Machine learning classification of patients with major depressive disorder and healthy controls using magnetic resonance imaging data

Dissertation

for the award of the degree

"Doctor rerum naturalium"

of the Georg-August-Universität Göttingen

within the doctoral program

GGNB Systems Neuroscience

of the Georg-August University School of Science (GAUSS)

submitted by

Vladimir Belov

from *Omsk, Russia*

Göttingen 2022

Thesis Committee

PD Dr. Roberto Goya-Maldonado

Dept. of Psychiatry and Psychotherapy, University Medical Center Göttingen

Prof. Dr. Annekathrin Schacht

Department of Affective Neuroscience and Psychophysiology, Georg-August-Universität Göttingen

Prof. Dr. Alexander Ecker

Institute of Computer Science, Georg-August-Universität Göttingen

Members of the Examination Board

1st Referee:

PD Dr. Roberto Goya-Maldonado

Dept. of Psychiatry and Psychotherapy, University Medical Center Göttingen

2nd Referee:

Prof. Dr. Annekathrin Schacht

Department of Affective Neuroscience and Psychophysiology, Georg-August-Universität Göttingen

Further members of the Examination Board

Prof. Dr. Alexander Ecker

Institute of Computer Science, Georg-August-Universität Göttingen

Prof. Dr. Andrea Antal

Dept. of Clinical Neurophysiology, University Medical Center Göttingen

PD Dr. Peter Dechent

Dept. of Cognitive Neurology, University Medical Center Göttingen

Prof. Dr. Fabian Sinz

Institute of Computer Science, Georg-August-Universität Göttingen

Date of oral examination: 15th February 2023

Declaration

Herewith I declare that I prepared the dissertation titled ***"Machine learning classification of patients with major depressive disorder and healthy controls using magnetic resonance imaging data"*** in my own capacities with no other sources and aids than quoted.

Vladimir Belov
Göttingen
December 2022

Acknowledgments

First, I would like to thank my supervisor PD Dr. Goya-Maldonado for his support and constant involvement and interest in my project. You have shown me how to crystalize scientific ideas and hypotheses, an experience that every student should be craving at the beginning of their scientific career. In addition, I would like to thank my TAC members, Prof. Dr. Ecker and Prof. Dr. Schacht, for their support throughout this project. Furthermore, I would like to thank all of the members of the ENIGMA Consortium. I truly enjoyed countless scientific discussions within this brilliant community. This work would not be possible without the supportive people within the SNIP lab: Grant, Vladislav, Tracy, Adytia, Anna, Tatiana, Valerie, Lara, Fabian, Asude, Yonka, Ali, Henrike, Caspar, and Joy. I could write a whole book about all the events we went through together and how they bonded us. Without a doubt, I had the best team of people to work with. Finally, my deepest appreciation is to my family and close friends, who supported me in the hardest times of this journey: my mom and dad, my grandma, Adi, Rashi, Elina, Rita, Lisa, Max, Valerie, Anna, Fabi, Asude, and Ayyash. I don’t know, where I would be mentally and physically without you.

Abstract

Major depressive disorder (MDD) is a prevalent and complex psychiatric disorder affecting more than 300 million people worldwide. It is characterized by a highly heterogeneous spectrum of symptoms, including low mood and self-esteem, loss of interest, sleep disturbances, and loss/gain of appetite. Psychotherapy and pharmacotherapy are established as the first line of treatment. Nevertheless, up to a third of patients do not respond to such interventions. The development of personalized, more successful treatment strategies requires a comprehensive understanding of the pathophysiology of depression and a set of corresponding biomarkers. There is a growing interest in investigating large-scale structural and functional brain alterations via neuroimaging techniques. Machine learning techniques have gained popularity in neuroimaging due to higher performances in classifying mental disorders and elucidating the brain's structural and functional connectivity patterns. The current findings' reproducibility is restricted by either small sample sizes or inadequate consideration of demographic factors and site-related differences – site effect - in the case of large multi-site samples. Careful heed of demographic factors' role in classification performance and meticulously considering site-related differences between subjects is expected to fill this gap. In this work, I focus on discriminating depressive subjects from healthy controls using shallow machine learning algorithms based on structural pre-segmented brain features. An unprecedented initiative in terms of the number of sites included, a large dataset from ENIGMA MDD Consortium allows for extensive analysis and generalizable results.

Furthermore, I investigate if higher classification performances can be achieved by analyzing high-resolution cortical vertex-wise maps and integrating volumetric characteristics, such as cortical thickness, with shape characteristics (sulcal depth and cortical curvature). Moreover, I test if deep non-linear classification algorithms, such as convolutional neural network (CNN), could potentially reveal complex patterns of brain organization, contributing to better detection of depression-related alterations compared to simple classification models.

The results show that the investigated machine and deep learning models yielded accuracy close to random chance, regardless of the data modality or resolution. Furthermore, the integration of volumetric and shape characteristics did not yield high results. I detected the presence of the site effect, which was addressed by a ComBat harmonization tool. However, ComBat failed to improve the classification performance of the models. More sophisticated classification models that can incorporate both demographic and clinical information could improve classification performance based on the brain's morphology in future studies.

Finally, I investigate the validity of functional subject-specific parcellation maps as a potential predictor of MDD. This proof-of-concept study uses subject-specific resting-state fMRI-based parcellations to reveal the effect of a single session of 10 Hz rTMS on healthy subjects. I applied RSFC-Snowballing and RSFC-Boundary mapping parcellation methods to obtain complementary node and boundary maps of functional brain organization, which were analyzed via Support Vector Machines (SVM) with a novel feature selection method. This approach revealed a slower and more complex response in boundaries compared to nodes located primarily in the posterior cingulate cortex and precuneus. These results highlight the potential benefits of subject-specific parcellations in future psychiatric analyses as they might capture distinct temporal and spatial differences.

The development of new personalized, more successful treatment strategies requires a comprehensive understanding of the pathophysiology of depression and a set of corresponding biomarkers. The heterogeneity of large worldwide samples in terms of socio-demographic, clinical, and genetic factors requires more sophisticated analytical approaches and solutions to achieve a more general overview of the pathophysiology of depression. Recently established scientific consortiums enable the investigations of unprecedentedly large datasets, requiring more powerful big-data analytical tools.

Table of Contents

Chapter 1 Introduction

1.1 Overview and impact of Major Depressive Disorder

Major depressive disorder (MDD, also referred to as unipolar depression) is a highly prevalent mental disorder. The World Health Organization (WHO) estimates the total number of people suffering from MDD to exceed over 300 million worldwide [1], with a total number of Years Lived with Disability (YLD) of up to 50 million, making it one of the leading causes of disability worldwide. During the depressive episode, the patient may suffer from low mood and self-esteem, suicidal thoughts, anhedonia, impaired cognition, insomnia or hypersomnia, and loss or gain of appetite [2], [3]. According to the Statistical Manual of Mental Disorders – 5 (DSM5) [4], MDD is characterized by at least one depressive episode minimum of 2 weeks duration, when at least 5 of these symptoms are present, among which low mood and anhedonia are considered primary symptoms. Moreover, all aspects of suicidality, such as suicidal ideation and plan, suicide attempts, and completed suicides, are commonly presented in depression [5]. There is an almost 20-fold higher risk of dying of suicide in MDD compared to the healthy general population [6]. In terms of sex, MDD affects twice as many women compared to men [7], with pervasiveness increasing with age, reaching a maximum in the range of 55-74 years [1], yet it can occur both in children and adolescents [8]. There is a strong association between MDD and elevated risk of cardiovascular disease [9], type 2 diabetes [10], and stroke [11]. Additionally, depressive symptoms were found to be linked to conditions with nonspecific

somatic complaints, such as chronic fatigue and chronic pain, negatively affecting the life of the patients with depression even further [12].

Pharmacotherapy and psychotherapy are common approaches to treat depressive episodes [3], [13], with up to 66% patient response rate [14]. Nevertheless, approximately one-third of all patients do not benefit from conventional therapies. Those presenting an inadequate response to at least two antidepressants of adequate doses and duration are diagnosed with treatment-resistant depression (TRD) [15]. The presence of other psychiatric disorders and general medical conditions, including HIV and cancer, contributes to a higher risk of developing TRD [16]. There are other approaches to treat both MDD and TRD in the field of brain stimulation, such as electroconvulsive therapy (ECT) [17] and repetitive transcranial magnetic stimulation (rTMS) [18]. Nevertheless, the remission rate is currently low despite the development of novel treatment strategies, such as Stanford neuromodulation therapy (SNT)[19], [20].

Perhaps one of the main obstacles to the successful treatment of depression is the heterogeneity of depression itself. According to the STAR*D study, only 1.8% of all MDD patients shared the most common symptom profile, with 14% having unprecedented among other symptom profiles [21]. A natural response to such heterogeneity is to apply system medicine approaches to reveal subgroups of depression based on their symptom profile allowing the prescription of treatment strategies according to patient subgroup. Assignment of the patient to the corresponding subgroup will potentially increase the remission rate. The early model of splitting MDD into two subtypes, endogenous/melancholic and neurotic/reactive, failed as there was no neurobiological evidence supporting that despite particular success in the choice of the medication based on this subdivision [22]. Another theory suggested splitting depression into four subtypes: anxious, melancholic, atypic, and unspecific depressions [23], with a significant overlap between these subtypes. The subtype's overlap contributes to the idea that depressed patients are found on the complex symptom's severity spectrum, and thus, a discrete assignation of the patient to any subtype should be performed carefully or entirely avoided.

Another way to address the low remission rate is through personalized treatment approaches [24]. The main aim of such a framework is to analyze biological information to access optimal clinical decisions on diagnostic and predictive treatment levels. This personalized, biomarker-based approach achieved much better outcomes in cancer therapy. It became a part of a current clinical practice, in which treatment is determined via molecular evaluation of tumor/ analysis of tumor-derived organoids or liquid biopsy [25], [26]. Symptom

heterogeneity of depression underlines the potential of personalized treatment. The first step in developing personalized therapies is investigating the pathophysiology of depression and searching for distinct biomarkers of depression.

Hitherto, there has been a lack of conclusive understanding of the pathophysiology of depression. One of the first neurobiological hypotheses of depression is the monoamine hypothesis. It suggests that the main symptoms of depression arise from the brain monoaminergic neurotransmitter deficiency in the synaptic clefts, which involve serotonin (5-HT), dopamine, and norepinephrine [27], [28]. Neurophysiology suggests a vital role of monoaminergic systems in regulating both cognitive and vegetative brain functions [29], [30]. Clinical and animal studies validated this hypothesis with experiments that evidenced increased availability of neurotransmitters driven by antidepressants [31]–[33]. According to the monoamine hypothesis, the antidepressants act fast in the synaptic cleft and should rapidly lead to the improvement in symptoms, yet usually, it takes 2-4 weeks until the effects of antidepressants become clinically visible [34]. Despite its success and clinical relevance, this hypothesis cannot explain all pathophysiological mechanisms of depression. Therefore, complementary theories are required.

Considering an association between chronic stress and the occurrence of MDD [35] and that a significant part of patients exhibits hormonal abnormalities [36] [37], endocrine processes were considered to be a significant part of the pathophysiology of depression. A substantial subgroup of patients has shown increased cortisol levels during depressive episodes [38] and recovery [39]. Moreover, a decreased level of thyroid hormones [40] and dysfunctions in the hypothalamus-pituitary-adrenal (HPA) axis were detected. In line with these findings, the use of triiodothyronine (T_3) was found to be an effective auxiliary treatment [41], [42]. Remarkably, patients without hypothalamus-pituitary-thyroid axis alterations were found to have lower 5-HT concentrations [43], connecting it to the monoamine hypothesis. An increased prevalence of depression among women could be linked to alterations in sex hormones level such as progesterone [44], and a higher risk of onset during pregnancy and directly after the postpartum period [45]. Nevertheless, pharmacological modulation of the HPA axis has low clinical efficiency, motivating the further search for other biomarkers.

Another hypothesis suggests that disturbances of synaptic plasticity or adaptation of neural systems are strongly involved in the pathophysiology of depression. Of interest here are neurotrophic factors, such as a brain-derived neurotrophic factor (BDNF) and neurotrophin-3 (NT-3), which were shown to be responsible for the growth and activity-dependent plasticity of 5-HT neurons [46]. The expression of BDNF in serum was found to be decreased in MDD

patients compared to healthy controls [47], which negatively affects the morphology of hippocampal neurons, making them unresponsive to stimuli and inhibiting the dendritic ramification. Hippocampus is directly involved in the brain's cognitive function, such as short-term and long-term memory consolidation and learning [48]–[50]. When treated with antidepressants inhibiting 5-HT uptake, BDNF level is elevated in the rat hippocampus. These changes are evident after up to 3 weeks of antidepressant intake [51], similar to treatment the response time of common antidepressants. Considering the reduced hippocampal volume in depressed patients and the neuroprotective effect of antidepressants [52], the neurotropic hypothesis has a good foundation to complement the monoamine hypothesis. Nevertheless, there are inconsistencies in this hypothesis. Several studies provided an opposite relationship between mood and BDNF level [53], [54]. It highlights the more complex role of BDNF in the pathophysiology of depression.

Considering the inconclusiveness of the above-mentioned studies, further investigation of brain alterations from the molecular level to large-scale brain network dynamics is required to provide new insights regarding depression. Given the replication crisis in psychology and clinical medicine [55], there is a growing interest in analyses of large quantities of neurobiological data using big data analytical methods. As more and more researchers collect data all around the globe, large mega-analyses are required to produce stable, reproducible results instead of local small sample studies with a low replication rate. Furthermore, the development of neuroimaging techniques opens a new direction in depression research. A study of large-scale neuroanatomical and functional brain alterations in depression via neuroimaging can potentially give new insights into the pathophysiology of depression, supplementing already existing hypotheses and allowing for the development of new successful therapies. In the next chapter, I will introduce the current development in neuroimaging approaches aiming to find biomarkers of depression via univariate approaches.

1.2 Neuroimaging Advances in MDD

Neuroimaging is a noninvasive technique to study brain neuroanatomy *in vivo*. In the last several decades, neuroimaging became a promising tool in outlining alteration in the brain correlating brain structure and function and psychiatric disorders, neurological disorders, and detecting lesions. According to analyzed brain features, neuroimaging can be divided into structural and functional neuroimaging. I will introduce them separately regarding their

potential to reveal quantifiable biomarkers of depression via univariate cross-sectional and longitudinal analyses.

Structural Neuroimaging

Structural neuroimaging is used to study brain morphology. The most widely used acquisition technique in clinical and research settings is magnetic resonance imaging (MRI). Compared to other neuroimaging, an advantage of the MRI technique is that it does not produce harmful ionizing radiation, as in computed tomography [56]. MRI relies on the nuclear magnetic resonance (NMR) signal, where spin-lattice relaxation (T1) or spin-spin relaxation (T2) of the tissue is measured with up to 2 mm spatial resolution [57]. This resolution allows for precise delineation of grey and white matter. The development of the very high field MRI scanners (7T and higher) allowed for even higher spatial resolution up to 0.5 mm, revealing microvenous anatomy of subcortical regions [58]. Another practical advantage of MRI is that the same scanner can be used for various sequence types to be collected in the same scanning session. Researchers can obtain brain structure data via T1, T2, diffusion-weighted magnetic resonance imaging (DWI), and diffusion tensor imaging (DTI). In the same scanning session, the brain's activity can be captured by functional MRI. T1 and T2 are best suited for investigating cortical and subcortical morphology. Among different morphological features of the gray matter that can be estimated via structural neuroimaging, the most commonly used are cortical and subcortical volumetric features, surface area, and thickness. DWI and DTI are used to analyze white matter tracks via structural connectivity analysis. In contrast, the microstructure of white matter can be characterized via fractional anisotropy, axial diffusivity, mean diffusivity, and radial diffusivity.

The first attempts at deciphering MDD pathophysiology via neuroimaging started three decades ago. The early neuroimaging studies analyzed the cortical and subcortical size and volume alterations in the MDD group compared to healthy controls obtained via CT and MRI, identifying reduced hippocampal volume [52], [59], smaller frontal areas [60] and atrophy in sulcal and ventricle measures [61] with inconsistent reports regarding other brain areas [62]. These findings implied no evident global brain atrophy but rather regional changes to be associated with MDD and its symptoms.

More recent studies validated the atrophy in the frontal lobe, more precisely in the anterior cingulate cortex (ACC), orbitofrontal cortex (OFC), and dorsolateral prefrontal cortex (DLPFC) [63], [64]. ACC is anatomically connected to areas associated with reward function,

such as the orbitofrontal cortex, ventral striatum, memory consolidation (hippocampus), and emotional regulation (amygdala), making it a significant hub of neuronal circuitry for affect regulation. Mayberg and colleagues considered metabolism in ACC a potential predictor of treatment response [65]. Supplementing the potential of ACC to be a biomarker of depression, the thickness of ACC was found to be negatively correlated with Montgomery-Asberg Depression Rating Scale (MADRS) [66], implying greater symptom improvement with thicker ACC. This association was further confirmed by Järnum and colleagues in their longitudinal study [67]. OFC plays an essential role in decision-making [68], [69], and emotional regulation [70]. Bilateral medial OFC exhibited the strongest cortical thinning compared to other cortical regions [71] in both adult and adolescent depressed patients, suggesting an essential role of OFC in the development of depression. Already after five weeks of standard TMS treatment in the left DLPFC, an increase in ACC and OFC volumes was observed by Lan and colleagues [72].

The parietal lobe is another cortical region, structural alterations of which have been reported in multiple studies [73]–[75]. It is involved in language, tactile signal processing, spatial awareness, and navigation [76]. Studies reported an increased cortical thickness in the left inferior parietal gyrus [73] and increased volume in the right postcentral gyrus [74] in first episode MDD patients. According to the same study, the postcentral gyrus in adolescent depressed patients exhibited a smaller surface area compared to healthy adolescents and was weakly negatively correlated with the Beck Depression Inventory (BDI-II). Furthermore, depressed patients with a high level of anxiety displayed a larger gray matter volume of the postcentral gyrus compared to non-anxious patients, highlighting a potential link between differences in symptom manifestation and structural alterations and the neurobiological heterogeneity of MDD [77].

The thalamus, which is often referred to as the “Grand Central Station” of the brain, is a major hub of sensory signals, tightly involved in emotional control, memory consolidation, and arousal [78]–[80]. The majority of brain regions are connected with specific divisions of the thalamus, damage of which can lead to an amnestic syndrome [81]. Recent studies probed the thalamus as a potential biomarker of depression reporting - however, contradictory results. According to two studies, a larger thalamic volume is observed among first-episode MDD patients [82], [83]. In another study, Lu and colleagues revealed thalamic atrophy in the first episode MDD group [84].

Another promising biomarker of depression is the amygdala. The amygdala is a part of the limbic system and plays a vital role in decision-making, memory encoding, and emotional

responses, including negative emotions, anxiety, and fear. According to a literature review by Belanni [85], the majority of small sample studies revealed reduced amygdala size in MDD compared to HC [86]–[88]. Nevertheless, a small portion of the studies reported unchanged [89] or even increased amygdala volume[90], [91].

The above-mentioned univariate studies are characterized by small sample sizes and even reporting contradictory results, limiting the credibility of reported findings. Large-scale data collection is not always possible due to the high financial cost of scanning sessions and limited access to the MDD population. Furthermore, local studies are typically focused on subjects from one geographical location, biasing the results to specific socio-economic, ethnic, and demographic groups. As more and more researchers conducted small sample neuroimaging studies worldwide, the meta-analytical approaches allowed for a more systematic and thus credible overview of MDD-related cortical and subcortical structural alterations. A cross-sectional meta-analysis (20 sites worldwide, 2148 MDD and 7957 Healthy Controls (HC)) performed by Schmaal and colleagues revealed thinner cortices in bilateral medial OFC and right caudal ACC, thus, confirming previous single-site findings, yet with smaller effect sizes [71]. Furthermore, no significant alteration in DLPFC and parietal lobe was reported, detecting only surface area reduction in the right inferior parietal cortex in adolescent patients and bilateral inferior parietal cortex in adolescents with recurrent depression.

Another recent subcortical meta-analysis from Ho and colleagues, which included 1781 MDD patients and 2953 HC, demonstrated smaller thickness and surface area of the amygdala in early-onset MDD vs. HC comparison, in line with the previous subcortical meta-analysis by Schmaal [92], and in multiple depressive episodes MDD vs. single episode MDD. Additionally, significant atrophy of the hippocampus in early (<21 years old) age of onset patients and patients suffering from multiple depressive episodes was detected [93], yet exhibiting more modest effect sizes compared to single-site studies [94], [95]. Furthermore, the results did not reveal significant hippocampal differences between general MDD and HC groups, in contradiction with a meta-analysis from Schmaal [92], despite a large sample size in the studies and partial overlap in the analyzed sample. Ho and colleagues revealed a reduction in both thickness and surface area in three subfields, cornu ammonis, dentate gyrus, and subiculum, in the early age of onset group. Inconsistent findings in hippocampal alterations highlight the necessity for future studies. Lastly, no significant shape differences were found in the thalamus in general MDD vs. HC comparison [93], suggesting more careful consideration of thalamic subregions in the pathophysiology of depression.

Overall, current findings are inconclusive regarding depression pathophysiology and how it is manifested in cortical and subcortical morphological properties. There is a trend of modest effect sizes in large sample studies and inconsistent findings in small sample studies with larger effect sizes. One possible explanation for such discrepancy is demographic and clinical heterogeneity in large sample studies versus more narrow demographic distribution with predominantly age- and sex-matched MDD and HC groups in small sample studies. Noticeably, sample size itself is a major factor in brain-wide association studies (BWAS). That was explicitly and conclusively shown in a study by Maker and colleagues [96]. They performed billions of univariate analyses to find an association between structural brain features and cognitive ability and psychopathology to measure the reproducibility, statistical error, and effect size as a function of sample size by using sample sizes from small (N=25) to large (N=35,572). Their results showed that effect size in BWAS had a high chance of being inflated in small sample studies. Moreover, there was a high chance of observing a sign error, i.e., detecting a correlation with an opposite sign to the whole sample result. A high false negative rate (>75%) and a low false positive rate (<25%) were observed even for large samples (N=1000), explicitly demonstrating a high chance for small sample studies to produce unreplicable results. Additionally, they revealed ***the underpowered BWAS paradox***: *"At smaller sample sizes, the largest, most inflated BWAS effects are most likely to be statistically significant and therefore, paradoxically, the most likely to be published."* [96] This paradox results from a high chance for small sample studies to generate strong random associations, which remain significant even after strict in-sample statistical thresholding, while more modest in terms of effect size associations are rejected. Thus, the remaining inflated results will most likely not be replicated in the independent sample.

A similar trend of modest effect sizes was observed in a mega-analysis by Winter [97]. In a large sample multi-site mega-analysis (861 MDD and 948 HC) by Winter and colleagues, a striking similarity between MDD and HC structural cortical and subcortical features in the univariate analysis was revealed. Right hemisphere volume was found to be most altered among other features with a small effect size and 91.6% distributional overlap between MDD and HC. When trying to classify subjects according to the diagnosis (MDD vs. HC), a balanced accuracy of 54.7% was achieved. This low accuracy explicitly shows that univariate structural biomarkers might not be informative enough to differentiate MDD from HC on a single-subject level. Moreover, subgroup analysis based on demographic (male MDD vs. male HC; female MDD vs. female HC) and clinical (acutely depressed MDD vs. HC, chronically depressed MDD vs. HC) factors yielded similar results.

Overall, structural alterations associated with MDD exhibit small effect sizes when computed via univariate analyses. Furthermore, there are concerns about the reproducibility of reported findings due to small sample sizes. In the following section, I will introduce the status of functional correlates of depression derived via univariate cross-sectional approaches.

Functional Neuroimaging

Functional neuroimaging allows the noninvasive studying of brain circuits involved in highly relevant functions, such as behavior, cognition, and emotion regulation. Compared to the structural counterpart, functional neuroimaging acquisition methods suffer from lower spatial resolution but provide information on neuronal activity patterns [98]. The most commonly used acquisition techniques are functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG), and positron emission tomography (PET). fMRI measures the blood oxygen level-dependent (BOLD) temporal changes. BOLD contrast is detected as a magnetic field change occurring due to increased oxygen influx caused by increased deoxygenated hemoglobin concentration after neural activation. This sequence of events is called a hemodynamic response to neural activation [98]. EEG measures the brain's neural activity by using electrodes placed on top of the scalp and sensitive to the activity occurring in both cortical sulci and gyri. MEG measures the magnetic field induced by neural activity using magnetometers and is the most sensitive to the activity happening in sulci. Like fMRI, PET measures the brain's blood flow by injecting subjects with radioactive isotopes such as fluorodeoxyglucose (18F). It is less sensitive to head movement than fMRI. Compared to other methods, fMRI has a higher spatial resolution. However, its losses in terms of temporal resolution due to 6 seconds delay between neural activation and BOLD changes [98]. Nevertheless, Ogawa and colleagues have found a new approach to scale down the fMRI signal time to milliseconds [99]. Another advantage of fMRI versus other functional modalities is that the same scanner is suited for structural MRI acquisition, thus allowing for collecting different neuroimaging datatypes in one session.

Generally, two major groups of analyses can be performed with functional neuroimaging. One option is to investigate activity patterns in a distinct brain region, for example, as a response to external stimuli or while performing cognitive or motor tasks. Furthermore, one can analyze a temporal concurrence of spatially distant neurophysiological events, referred to as functional connectivity (FC) [100]. A basic assumption of FC analysis is that if an event triggers two brain regions to be simultaneously (or with a certain delay) to be

activated that these regions should relate to each other. Thus, brain regions that exhibit consistently high FC are presumed to be coupled or belong to the same *network*. Large-scale brain network analysis shifted a previously considered paradigm that cognitive tasks are performed by distinct brain regions working independently but rather by global networks comprised of several brain regions functionally connected [101]. Historically, networks were first identified when subjects were performing cognitive tasks and thus referred to as task-positive networks [102], [103]. Large-scale networks can be captured by clustering algorithms or, for example, spatial independent component analysis (ICA) and further analyzed by graph theory methods. Crucially, it was revealed by Biswal and colleagues that there is a physiological fluctuation of the activity in the motor, auditory, and visual cortex even in the absence of any task – in a resting state [104]. These spatially distant regions exhibited a high intrinsic functional connectivity and thus form a distinct network, active in the resting state and deactivated in the presence of the motor or cognitive task - a default mode network (DMN) [105].

It has been demonstrated that alterations of functional networks occur in many brain disorders, including Alzheimer's disease [106], multiple sclerosis [107], and stroke [108]. Alteration of FC patterns has also been associated with MDD, suggesting that dysfunctions of large-scale brain networks may play a significant role in the pathophysiology and manifestation of MDD [109]. DMN is considered as one of the most critical networks in the pathophysiology of MDD. It is composed of the perigenual parts of ACC, posterior cingulate cortex (PCC), medial prefrontal cortex (mPFC), precuneus, and angular gyrus [105]. DMN is active during mind-wandering and processing self-related information. The aberrant activity profile in rostral ACC has been presumed to be related to negative introspective thinking [110]. Greicius and colleagues showed that sgACC was functionally hyperconnected with the whole DMN in depressed patients [111], in line with other studies showing increased connectivity within DMN [52], [112]. However, Connolly and colleagues found sgACC hypoconnected to left precuneus in depressed adolescents, which suggests more complicated interactions within DMN [113].

The central executive network (CEN), also known as the frontoparietal network (FPN), is involved in a variety of cognitive functions, such as decision-making, working memory, and concentration [114]. CEN consists of DLPFC and the posterior parietal cortex (PPC). Opposite to DMN, CEN is deactivated during the resting-state and anti-correlated with DMN both in the presence of the cognitive task and during the resting-state [115]. Taking into account the presence of cognitive impairment and inability to concentrate in the symptom profile of MDD, CEN was frequently studied in depressed patients. DLPFC exhibited decreased brain activity in the resting-state in MDD, which was normalized after remission [116]. Complementary, an

increased activity within DMN was observed during the cognitive task in remitted MDD [117], linking to the inability to suppress DMN, which subsequently leads to reduced activity in CEN. These findings suggest depression-related changes occur not only within each network but rather global miscommunication between DMN and CEN.

The salience network (SN) regulates communication between DMN and CEN – a network involved in detecting and processing salient stimuli [118]. It includes the dorsal anterior cingulate cortex (dACC), amygdala, ventral striatum, ventral tegmental area, and anterior insula (AI). Considering the close interaction between DMN and CEN via SN, a triple network model of pathophysiology was proposed, the malfunctioning of which is assumed to play an important role in most psychiatric disorders [119]. According to a study by Sridharan and colleagues, SN regulates the switch between DMN and CEN during external cognitive tasks and introspective cognitive processes [120], [121]. When a external salient stimulus occurs, SN activates CEN and deactivates DMN, respectively. Dysfunction of this mechanism leads to an aberrant activation pattern in the networks, leading to cognitive impairment and affective dysregulation.

In addition to the triple network model, functional alterations in the limbic system may provide additional information on the pathophysiology of depression as it is tightly involved in emotion regulation and reward-related decision-making [122]. It includes the amygdala, hippocampus, nucleus accumbens (NAcc), and OFC. There are inconsistent results regarding the functional alterations in the hippocampus. According to one fMRI study, the hippocampus is hyper-activated in MDD patients who experienced multiple depressive episodes [123], while another study demonstrated an opposite trend in first-episode unmedicated patients [124]. Moreover, Hao and colleagues identified an increased FC between cornu ammunis and the left premotor cortex, while the dentate gyrus exhibited increased FC with OFC and left ventrolateral prefrontal cortex (vlPFC) [125]. In conclusion, since the reported functional alterations were widespread in the brain, whole-brain analyses are required to decipher global functional changes. Moreover, due to the high dimensionality of the functional neuroimaging data, large-sample studies are required to obtain consistent results.

Considering on average small sample size in functional neuroimaging studies, there is a concern regarding the reproducibility of functional biomarkers of depression. In a recent mega-analysis, Yan and colleagues primarily examined DMN resting-state FC patterns [75]. This study included 794 MDD patients and 848 HCs from 17 sites in China. A significantly decreased FC within DMN was exhibited in depressed patients with multiple depressive episodes compared to HCs, but not in the single-episode group. That contradicts findings from

a meta-analysis by Kaiser and colleagues, reporting hyperconnectivity between different DMN regions [109]. Another mega-analysis from Javaheripour and colleagues, including 606 MDD and 476 HCs, revealed only a trend of decreased resting-state FC within DMN in the MDD group [126]. Moreover, there was no significant alteration of FC between DMN and CEN. Nevertheless, authors assumed that previously found statistical effects of DMN and CEN alterations were inflated due to small sample sizes and biases in selecting regions of interest (ROIs). Overall, even large-sample studies yield inconsistent results between each other, arguably due to the even higher dimensionality of the functional neuroimaging compared to the structural counterpart.

The above-mentioned inconsistencies in reported structural and functional alterations raise concerns regarding the usefulness and reproducibility of potential neuroimaging biomarkers of depression driven by univariate analyses. Usefulness of the existing univariate biomarkers is low, as reported effect sizes are small in large sample studies, and they do not allow for correct single-subject predictions [97]. Reproducibility is a major concern, as small sample studies tend to produce inflated, over-optimistic results [96]. Considering the complexity and the widespread nature of the reported alterations, much hope has been placed on the use of multivariate approaches, including shallow and deep machine learning models, to reveal neuroimaging biomarkers of depression. A straightforward approach is to perform MDD vs. HC classification task to estimate the differentiability between these groups.

1.3 Application of machine learning in MDD

There is growing interest in identifying phenotypes and biomarkers of depression by using machine learning [127]. Machine learning is an implementation of computational systems capable of learning and adapting their outputs according to received input information. By using adaptive algorithms and statistical models, it can find patterns in the input data to solve pre-defined tasks. Machine learning algorithms can be divided into three categories based on the data presented to the algorithm and the desired output:

- Unsupervised learning. The goal of unsupervised-learning algorithms is to find patterns in existing input, potentially grouping or clustering the data.

- Supervised learning. Algorithms belonging to this group require both the inputs and the desired output to make an inference. The most commonly used tasks of supervised-learning algorithms are classification and regression.
- Reinforcement learning. The algorithm that belongs to this class is trained to take actions in the environment according to the external data to maximize a reward function. Reinforcement learning is a rare guest in neuroimaging studies and has no direct implication in the scope of this project; thus, I will not discuss it in detail.

A standard pipeline of machine learning MDD vs. HC classification includes data splitting step, feature reduction, model training, and performance evaluation of the trained model.

Before the training or any computations, a common strategy is to split the data into training and test sets. The purpose of the training set is to train the model, while a test set is used to estimate the trained models' performance. The train-test split is appropriate when both train and test sets are "sufficiently large", i.e., both contain a full representation of the problem domain. Considering the sparsity of available neuroimaging data in most studies, well-established practice is to perform a k-fold cross-validation (CV) to estimate the actual performance of the trained model. In this case, data is split into k-folds. Each one of them is used as a test set, while k-1 is used as the training set. Thus the model is trained/tested k times in total. More specialized CV splits include leave-N-out CV, where N subjects are taken out from the training set and used as the test set. The most extreme version is the leave-one-out CV with one subject used to evaluate the model's performance, and the procedure is repeated for every subject.

Data splitting into training and test sets should occur before any step in the analysis pipeline, including feature extraction. Otherwise, the test set will influence either the training procedure or the feature extraction, causing a data leakage, often referred to as "double dipping" [128]. It inflates the results drastically, and it was detected in many published neuroimaging studies [129]. Another potential problem is a covariate shift – a mismatch between the distribution of the independent variables in the training and the test sets. This is a common phenomenon when a model is trained and tested on the data collected from different sources. One possible solution to that problem is to ensure the heterogeneity of the training set in terms of its independent variables [130]. Clinical multi-site datasets have the advantage of greater sample sizes. Furthermore, the data is more heterogeneous in terms of demographic factors

(age, sex, ethnicity, socio-economic status), clinical profile (severity of depression, comorbidities, number of depressive episodes), and imaging acquisition (scanner model, acquisition settings, preprocessing). In this scenario, the measures of performance should be additionally estimated for each site separately, thus performing leave-sites(s)-our CV [131], [132] to see how translatable the models are to data from unseen sites.

The performance of the classification model depends on the complexity of the task vs. the amount of available data. There is a high risk of model overfitting due to the high dimensionality of neuroimaging data and predominantly small sample sizes. Overfitting refers to the situation when the trained model learns all the available information from the training set, including noise, and learns it as a representation of the classes. It yields higher accuracies when measured on the training data, however, these representations are poorly translated to the unseen data, yielding much lower classification performance on the unseen data. More complex algorithms, i.e., containing more trainable parameters, tend to have a greater risk of overfitting [133]. Bashir and colleagues proposed one way to frame overfitting, suggesting that overfitting is the symptom of a mismatch between the model's informational capacity and the data's complexity [134]. One of the methods to tackle the overfitting problem is to reduce the data's dimensionality by extracting only patterns that are significant for the classification and removing noise or random fluctuations in the data. Feature reduction reduces the dimensionality of the data and subsequently reduces the risk of overfitting. Feature reduction methods include feature selection and feature extraction methods. In the case of feature selection, the most informative features are chosen for the analysis, while the rest is discarded. An example of a feature selection method is the two-sample t-test. Feature extraction methods project all original features into lower dimensional space. It can be either linear projection, such as principal component analysis (PCA), or non-linear (Isomap [135], t-distributed stochastic neighbor embedding [136]).

The discriminative model is trained via the optimization of trainable parameters to maximize the discrimination of one class from another. Examples of trainable parameters are characteristic of the hyperplane for a support-vector machine (SVM) or the weights in the artificial neural network (ANN). Additionally, one can optimize “hyper-parameters”- the model’s top-level features (such as C parameter and the choice of a kernel for SVM; learning rate, and choice of optimization function for ANN), by performing the nested CV. This procedure is computationally expensive and requires extensive data.

The performance of the classification algorithms can be measured by various categorical and rank-based metrics [137]. Categorical metrics can be derived from a confusion matrix – a

matrix with predicted labels in rows and observed labels in columns. The most commonly used metric is the accuracy, calculated as the number of correctly classified samples divided by the total number of samples. The major flaw of accuracy as the performance metric is that it does not consider the ratio between classes. In highly unbalanced datasets, the classification model can always learn to predict the majority class, yielding high accuracy but no actual discriminative power. The balanced accuracy accounts for the ratio between classes and is calculated as the mean of sensitivity (true positives divided by all positives) and specificity (true negatives divided by all negatives). The most common rank-based metric is the area under the receiver operating characteristic (ROC) curve (AUC), which estimates the separation between two class distributions separated by the model. It describes how sensitivity and specificity change as a function of the classification threshold (in categorical metrics, it is usually set to 0.5 as no class should be preferred by default).

Next, I will introduce the advances and pitfalls of supervised-learning algorithms applied to both structural and functional imaging data.

MDD vs. HC classification

Supervised-learning algorithms can be directly applied to differentiate healthy subjects from MDD patients. Numerous studies approached MDD vs. HC classification based on structural and functional brain images [127]. According to a review of 66 MDD studies by Gao and colleagues [127], the most commonly used classification algorithm is SVM, followed by the Gaussian process classifier (GPC) and linear discriminant analysis (LDA). SVM is a robust algorithm that finds a hyperplane that maximizes the width of the gap between two classes [138]. GPC is a non-parametric algorithm based on the Bayesian framework, which is a generalization of Gaussian probability distribution and requires a kernel to estimate the covariance function of the data [139]. LDA estimates the mean and variance for every class while calculating the probability of assigning new samples. There was significant variability of reported accuracies in considered studies, ranging from 52% to 97%, potentially because of small sample sizes, as only one study exceeded 700 subjects. Due to the small sample sizes, the most commonly used CV strategy was leave-one-out CV, although it is susceptible to overfitting and exhibits high variance [140]. According to Gao's previous work in classifying MDD vs. BD, 10-fold CV produces more stable results [141]; thus, it is recommended in future studies. On average, according to the review paper by Gao and colleagues, resting-stage fMRI

studies reported greater accuracy (≈86%) compared to both task-based fMRI (≈79%) and structural MRI (≈75%) studies [127]. Considering the high variability of results and small sample sizes, there is no conclusive evidence on the differentiability of HC and MDD with supervised-learning algorithms using neuroimaging data.

Counterintuitively, large-sample studies did not report higher classification performance. A predictive analysis competition (PAC) was conducted, in which researchers (49 teams worldwide) developed and applied machine learning models to perform MDD vs. HC classification based on structural MRI (https://www.photon-ai.com/pac), including 759 MDD patients and 1033 HC coming from 3 different sites. The best machine learning model yielded an accuracy of 65%, up to 10% lower than the average small-sample classification study. Using PAC data, Flint and colleagues investigated the effect of both training and test set sizes on classification performance [142]. They trained and tested machine learning models on the full data to acquire the benchmark of classification performance and then varied sample sizes from N=4 to N=150 to mimic the typical sample sizes in neuroimaging studies. The sample size was found to impact the variability of classification performances significantly. Samples with 20 subjects yielded accuracies in the range of 10% to 95%, while samples with 100 subjects produced a more shallow range of accuracies between 35 and 81%. Importantly, this effect was symmetrical, yet the lower range of performance is rarely presented in any studies, potentially due to publication bias. This trend was primarily driven by the small test sample size and not the training sample size. Therefore, large test sets are essential to obtain realistic classification estimates.

In a recent large sample study, Stolicyn and colleagues explored the classification ranges of different classification algorithms (SVM, decision tree, and logistic regression) applied to structural brain features [143]. The highest accuracy of 75% was achieved in the small age/sex-matched subsample (30 MDD, 30 HC) from the STRADL site (Stratifying Depression and Resilience Longitudinally, [144]) with formally diagnosed participants. However, when they used another set of HCs, the accuracy dropped to 62%. The highest accuracy was not replicated in the larger community-based cohort UK Biobank (N=8,959), yielding the highest accuracy of 60%. Several surface area features were found to be the most informative for MDD vs. HC classification: precentral cortex, ACC, superior frontal cortex, and lingual gyrus, in line with previously mentioned univariate mega-analyses. However, the direction of alterations in these regions was not evaluated. Stolicyn performed a similar to Flint's analysis using the UK Biobank dataset, where he varied the sample size to see its effect on classification performance. Higher accuracy of 75% was only achieved in the small (N<100) sample, exponentially

dropping to 53% with the large sample (N>4000), thus demonstrating a similar trend as in the Flint study. Another potential contributor to low accuracy in the UK Biobank site was the absence of the formal diagnosis of patients at scan time. The diagnosis was based on the current and past self-reported symptoms, arguably resulting in a less severe form of depression prevailing on this site. However, when Stolicyn stratified the UK Biobank sample based on past depression (remitted and lifetime-experienced), the top accuracies did not surpass 61% either. The lifetime-experienced group was assumed to represent a more severely depressed group. Thus, the severity of depression did not play a significant role as it was initially assumed.

Despite the consistency of the classification performances in previously mentioned studies [142], [143], several limitations must be highlighted. Firstly, none of the studies considered demographic covariates and their impact on the classification. As Flint mentioned, smaller samples tend to have a balanced and narrow demographic profile, thus making the results less translatable to the independent sample corresponding to the real-world scenario. In the Stolicyn study, age and sex were matched in the samples for all of the comparisons. Flint did not use this information to build the sample. Considering an existing effect of accelerated brain aging in depressed patients [145] and sex-related differences in unmedicated patients [146], age and sex should be accounted for in the analysis. As it was estimated by Snoek and colleagues, counterbalancing, as in the Stolicyn case, is less effective than regression models, in which the effect of covariates is regressed out from the brain features [147]. It is critical to incorporate this step within the CV routine to avoid data leakage, estimating the regression coefficients in the training set and applying them to the test set. Other options are to remove the effect of covariates directly from predictions, as was tested in the Dinga study [148], or to incorporate them directly in the deep learning model – an artificial neural network that uses multiple layers to extract and process low- to high-level features [149].

Secondly, none of the multi-site studies directly accounted for the site-related differences such as scanner type, acquisition protocols, inclusion/exclusion criteria, and demographic profiles. Flint detected a statistically significant effect of the scanner distribution on the accuracy of the hold-out test set, however, the site effect was not controlled in the main analysis. According to research by Solanes and colleagues [150], if not adequately addressed, it may introduce another bias to the classification performance. They created data that appeared to come from two different sites with and without differences in their distribution of the corresponding classes. Additionally, they introduced the site effect in case the real effect was strong and in the absence of the real effect. It revealed that the accuracy might be both inflated and shrunk in the presence of the site effect. The shrinkage was observed in the scenario when

both sites had a very strong real effect. Inflation of the accuracy occurred when the real effect was weak, while a strong site-effect was introduced. Solanes' solution to deal with the site-effect was to include site-covariate into the analysis by, for example, regressing it out. This method, however, does not account for a site-effect that changes the entire shape of the input distribution, and it may strongly correlate with other covariates. Both of these problems can be addressed via site-removal techniques such as ComBat [151], which estimates the additive and multiplicative effect of the site, while preserving the effect of other covariates. I will address both demographic and site-related factors directly by age and sex regression and application of ComBat and its modifications, respectively.

Overall, there are concerns about the unbiased classification performance of even the most commonly used shallow machine learning models and where the corresponding alterations occur. Moreover, the studies by Flint and Stolicyn considered only well-harmonized large-scale sites, thus lacking variability in acquisition, clinical factors, and demographics. In contrast, small-scale sites represent a substantial part of existing neuroimaging data (N<100) and, if combined, may provide a more general picture due to the variability in factors mentioned above. In the neuroimaging field and, more specifically, in MDD-related studies, the translation of machine learning models on unseen small-scale sites is underexplored. This work aims to analyze a large multi-site dataset and thus obtain a comprehensive overview of classification performances.

Furthermore, it is still an open question if unsatisfactory classification performance can be further boosted by analyzing more fine-grained features such as vertex-wise brain meshes or the voxel-wise structural images instead of a rather sparse set of atlas-based features. While surface area, thickness, and volumes of cortical and subcortical regions were analyzed extensively, the shape characteristics of the cortical landscape are underexplored. Integration of shape morphometric modalities, including gyral and sulcal shapes, the deformation of which were associated with MDD [152], [153], may lead to higher classification performances. In the recent study by Gao and colleagues, cortical vertex-wise brain features, including thickness, sulcal depth, and curvature, were used to predict the sex of the healthy subjects from the Human Connectome Project (HCP) S1200 release [154] and perform autism vs. HC classification [155]. Deep learning models exhibited significantly higher accuracy than a shallow model – SVM, highlighting non-linear models' potential in identifying neuroanatomical brain alterations. A similar trend was observed in the study by Abrol and colleagues [156]. They used structural MRI images directly to predict the age and sex of healthy subjects via machine and deep learning algorithms. Deep learning models surpassed more shallow classification models. I will

test if deep learning enables higher classification performance than shallow machine learning models when applied to vertex-wise cortical maps.

Machine learning models can also analyze patterns of altered brain activity. According to a review paper by Gao and colleagues [127], 23 resting-state fMRI studies yielded higher accuracy on average (≈85%) than structural. The most discriminative connectivity features were located within and across DMN, visual network, affective network, and cerebellum. Nevertheless, the risk of overfitting in resting-state fMRI studies is considerably higher due to the higher dimensionality of the functional data. The presence of a time component in fMRI compared to structural MRI increases the dimensionality manifold. That can be addressed directly by analyzing atlas-based connectivity features. However, the number of connectivity features from whole-brain atlases still dramatically surpasses the number of structural whole-brain segmentation features. Another source of a higher risk of overfitting is a low average sample size in resting-state fMRI studies – a general trend in clinical neuroimaging studies [157]. Considering both of these factors, there are reasonable concerns if the reported accuracies will hold in large sample studies.

To my knowledge, there are few large-sample multi-site resting-state fMRI studies to classify MDD and HC. Drysdale and colleagues performed the classification of MDD biotypes vs. HC [158], resulting in 89.2% accuracy in the training set and 86.2% accuracy in an independent dataset from a different site. Another resting-state fMRI multi-site study by Nakano and colleagues included 163 MDD and 195 HCs collected from four different sites [159]. They used 137 ROIs from Brainvisa Sulci Atlas [160] for FC analysis and performed leave-subject-out CV with all sites mixed in all folds. In line with recommendations from Woo to evaluate models on unseen sites [131], they also performed a leave-one-site-out CV. There was a striking difference between the results from these two validation schemes. The best performance in leave-one-out CV yielded 73.3% accuracy. In contrast, leave-site-out CV yielded 53.3% accuracy on average. The sensitivity and specificity exhibited extreme values in different sites, indicating the influence of site-related biases on the classification. To overcome that issue, they regressed out the site from both the training and test set. Critically, part of the independent set was used to estimate site effect, contradicting to leave-site-out strategy as the same site was presented in both training and test sets. Nevertheless, the accuracy did not improve significantly and yielded 54.7%. This low accuracy is in line with the previously mentioned Winter study, in which logistic regression was trained on the single FC feature with the largest effect size, yielding a balanced accuracy of 55.4% [97].

Yamashita and colleagues achieved higher a model performance of 66% accuracy validated on 449 subjects from five independent imagining sites [161]. They trained logistic regression with the least absolute shrinkage and selection operator (LASSO) [162] on connectivity matrices from 713 participants collected in 4 sites. Additionally, they employed their recently developed harmonization method to remove site-related differences [163]. The main idea of this method is to separate site-related differences into sampling bias and measurement bias. This method was compared with harmonization via ComBat and without controlling for the site-effect. There were inconsistent results on which harmonization tool yields higher model performance. Nevertheless, the results were similar across all training and test sites.

Qin and colleagues achieved a noteworthy classification performance in the largest up-to-date sample, including 1586 participants (821 MDD and 765 HC) from 16 sites of the Rest-meta-MDD consortium [164]. They deployed a graph convolutional network (GCN), resulting in an accuracy of 83.1% in leave-one-site-out CV. The core element of GCN is the spectral graph convolution filter [165], applied on the irregular graph instead of Euclidean data. In this study, the input layer of GCN received a graph calculated from the whole-brain FC matrix. ROIs from the pre-defined atlas represent the nodes of the graph. Edges of the graph are determined via clustering k-nearest neighbors (KNN) algorithm applied on the nodes. Age and sex were regressed out from the data via a non-linear Gaussian process, while the site-effect was controlled by ComBat. In line with Yamashita's study, simple models achieved up to 70% accuracy, while GCN drastically outperformed all other models.

A potential limitation in obtaining even higher classification accuracies and reliable functional signatures of depression may arise from standardized parcellation templates used in the group analyses, such as pre-defined brain atlases or pre-defined ROIs, which are usually built upon healthy population [166]–[168]. In considered studies, it is assumed that spatial topography does not depend on the diagnosis and remains unchanged across tasks [169]. Moreover, these studies do not consider the intra-subject variability of network organization, which may result in one ROI belonging to different functional networks in different subjects [170]. Averaging the results obtained from atlas-based ROI will inevitably obstruct reproducibility and lower statistical significance. Salehi and colleagues explicitly demonstrated how spatially variable functional networks were between subjects and across different tasks [171]. They defined three entropy classes of regions (referred to as nodes in the study), representing how stable is the assignment of a particular node is to the corresponding network. According to their results, primary and secondary visual networks, as well as the frontal part of

DMN, contained predominantly steady nodes, i.e., areas that maintained their network affiliation independent of the task or subject. Dorsal-medial part of DMN consisted mostly of transient nodes, which change their network affiliation in tasks and are highly variable across subjects, highlighting the potential weakness of atlas-based group-level analyses of depression. Moreover, in their next study, Salehi and colleagues estimated the robustness of fixed group-level atlases in which boundaries were defined anatomically or functionally [172]. Their findings suggested that there is no fixed single functional atlas of the brain, which implies the necessity of individual-level parcellation schemes. Moreover, the reconfiguration of the nodes according to the tasks can itself be informative and should be included in the interpretation of the changes in connectivity patterns. To my knowledge, there is no study to date that analyzed psychiatric disorders alterations via single-subject parcellations.

In summary, there are several limitations in above-mentioned structural and functional machine learning studies. First, multi-site structural studies analyzed homogeneous samples in terms of the demographic factors, thus limiting the generalizability to the unseen demographic collected elsewhere. Secondly, site-related differences were not systemically addressed in the structural studies. Therefore, the results can be biased toward scanner type, acquisition protocols, or inclusion/exclusion criteria. Furthermore, low classification accuracies obtained in large sample structural studies could result from the low-resolution atlas-based features. The investigation of high-resolution cortical maps and taking into account the shape of the cortical landscape could potentially lead to higher model performances. Lastly, inter-subject variability of functional network organization is lost when we apply population-based atlases in functional analyses. Functional network organization can be extracted via subject-specific parcellation schemes, the analysis of which may result in a boost in MDD vs. HC classification performances.

There are three primary goals of the work presented in the following chapters. The first goal is to evaluate the classification performance of the shallow machine learning models applied to structural atlas-based brain features to distinguish MDD from HC. I extensively analyzed a demographically heterogeneous large multi-site dataset provided by ENIGMA MDD Consortium. This analysis is unprecedented in terms of the number of included sites worldwide, allowing for an extensive investigation of the generalizability of the classification performance across sites. I extensively control for site effect and analyze its impact on the classification performance, which was remained lacking in past studies.

Furthermore, previous large-sample studies have analyzed only atlas-based cortical structural features. To further extend our understanding of cortical brain alterations in MDD, I

analyzed cortical vertex-wise meshes, including thickness sulcal depth and curvature, which provide a more detailed description of cortical morphology. Considering a substantial number of available subjects from sites, I hypothesize higher model performance using cortical meshes compared to atlas-based features. I apply a convolutional neural network (CNN) (pre-trained Dense Net[173]) able to reveal complex non-linear biomarkers as the provided ENIGMA MDD dataset is large enough to minimize the risk of overfitting. I expect higher accuracy of the deep learning model compared to the linear machine learning model, as it may capture non-linear MDD-related alterations.

The third goal of this work is to examine the validity of personalized parcellations as a potent data domain to be investigated as an MDD predictor in future psychiatric studies. In this proof of concept study, I applied subject-specific resting-state fMRI-based parcellations to unravel the effect of a single session of personalized 10 Hz rTMS on healthy subjects. I apply RSFC-Snowballing and RSFC-Boundary mapping parcellation methods to obtain complementary node and boundary maps of functional brain organization [174], which were subsequently analyzed via SVM with a novel feature selection method. I expect 10 Hz rTMS to perturb both nodal and boundary maps in the spatially distant areas that functionally interact with the left DLPFC (stimulation point).

Chapter 2 Scope of the dissertation

Numerous studies tackle the classification task of MDD vs. HC, analyzing anatomical brain features and functional brain patterns[143], [175]–[178]. Despite that, the current findings' reproducibility is restricted either by small sample sizes or inadequate consideration of the site effect in the case of large multi-site samples. Careful heed of demographic factors' role in classification performance and meticulously considering site-related differences between subjects will fill this gap. Moreover, non-linear classification algorithms, such as deep learning models, could potentially reveal complex patterns of brain organization, contributing to better detection of depression-related alterations compared to simple classification models. Lastly, the application of subject-specific functional parcellations may contribute to better MDD diagnosis, as commonly used group-based parcellations do not account for diagnosis-related differences.

MDD vs. HC classification on atlas-based cortical and subcortical morphometric measures

Even though work has been done to classify MDD and HC based on cortical and subcortical structural measures, limited sample sizes and narrow demographic and clinical profile distributions may hinder reproducibility in the current neuropsychiatric studies with machine learning as a predictive tool. Several large sample studies addressed some of these

issues [142], [143]. Yet, there are still concerns about the translatability of the results to unseen sites, as only 1) locally collected and demographically matched samples were analyzed 2) other site-related differences, such as differences in acquisition protocol, were not rigorously addressed. My goal is to present real-world global difference between MDD and HC groups by analyzing cortical and subcortical atlas-based features in the large multi-site sample from ENIGMA MDD consortium, including 30 sites worldwide, via shallow linear and non-linear machine learning algorithms (Chapter 3). I evaluated classification performance via two different CV splitting: 1) Splitting by Age/Sex, where we balance age and sex distributions across all CV folds, and 2) Splitting by Site, in which site can be found only in one particular CV fold. Splitting by Age/Sex reveals the unbiased classification performance regarding demographic factors, while in Splitting by Site, one estimates the classification performance on the unseen during the training sites, thus measuring the generalizability of the models. The differences in classification performances between these two CV splitting strategies would point to the presence of the site effect. Additionally, I address the site effect by applying ComBat and its variations, such as ComBat GAM and CovBat [179], [180], to improve the classification performance. Lastly, I stratify the data according to these factors to address if clinically and demographically more homogenous subgroups would yield higher classification performance.

MDD vs. HC classification on cortical vertex-wise maps: A deep learning approach

Despite numerous attempts to apply machine learning algorithms [142], [143], including the first part of my study (see Chapter 3), the classification of depression based on structural atlas-based ROIs did not yield high classification performance. The development of highly non-linear classification algorithms, such as deep neural networks, led to higher classification performance in neuroimaging compared to more commonly used classification algorithms [155], [164]. Therefore, the application of deep classification models could improve the differentiability between MDD and HC based on structural information. Furthermore, previous large sample studies analyzed atlas-based cortical features, such as cortical surface areas, thickness, and subcortical volumes. The investigation of the full cortical landscape via cortical vertex-wise meshed could improve low accuracies. In addition, other morphometric shape modalities, such as cortical curvature and sulcal depth, were not integrated into the analysis. Aggregation of cortical thickness with shape characteristics could lead to higher classification performances.

In this study, I will test if the deep learning model (pre-trained DenseNet), which was successfully implemented in similar data types before [155], can differentiate MDD from HC using vertex-wise maps and outperforms a robust shallow machine learning classification model - SVM. I perform multi-site MDD vs. HC classification based on structural cortical thickness and shape (sulcal depth and curvature) maps, expecting higher classification performance of both models when all morphometric characteristics are combined compared to separate analyses. The sample is provided by ENIGMA MDD Consortium, with a total of 7,012 participants (2,772 MDD and 4,240 HC) from 31 sites worldwide. In line with my previous study (Chapter 3), to address the site-related biases, I obtain the classification performance by splitting the dataset according to demographic factors (Splitting by Age/Sex) and site affiliation (Splitting by Site). Lastly, I will apply ComBat to mitigate site effect, expecting that it will remove site-related differences and, thus, lead to more generalizable results.

The effect of high frequency rTMS on subject-based functional connectivity nodes and boundaries in healthy subjects

rTMS is a well-established and successful treatment strategy for neuropsychiatric disorders, including MDD. Nevertheless, the exact mechanism of rTMS on the brain's functional level is still not well understood. Group-based parcellation schemes can map the major nodes of brain network organization, which are present across individuals, but these schemes are typically built on healthy subjects. Hence, this approach is blind to subject-specific brain organization, which may be especially relevant to reveal functional alterations due to psychiatric disorders. These subject-specific characteristics can be extracted via subject-based data-driven parcellations, such as the resting-state functional connectivity (RSFC) Snowballing algorithm [174]. It outputs a connectivity peak density map, representing brain network hubs. A complementary parcellation algorithm, RSFC-Boundary mapping, identifies the location of abrupt changes in connectivity patterns, thus revealing the boundaries between snowballing hubs [174]. In this proof-of-concept study, I apply both parcellations to reveal the effect of 10 Hz rTMS within the first hour of the stimulation in nodes and boundaries. I develop a novel feature selection method for both nodes and boundaries, which are analyzed via SVM. We hypothesize that 10 Hz rTMS will affect nodes and boundaries distant from the personalized stimulation location in the left DLPFC.

Chapter 3 Multi-site benchmark classification of major depressive disorder using machine learning on cortical and subcortical measures.

Multi-site benchmark classification of major depressive disorder using machine learning on cortical and subcortical measures

Vladimir Belov[1], Tracy Erwin-Grabner[1], Ali Saffet Gonul[2], Alyssa R. Amod[3], Amar Ojha[4], Andre Aleman[5] ,Annemiek Dols[6] , Anouk Schrantee[7], Aslihan Uyar-Demir[2], Ben J Harrison[8], Benson Mwangi[9,10] , Bianca Besteher[11], Bonnie Klimes-Dougan[12], Brenda W. J. H. Penninx[6], Bryon A. Mueller[13], Carlos Zarate[14], Christopher G. Davey[8], Christopher R. K. Ching[15], Colm G. Connolly[16], Cynthia H. Y. Fu[17,18], Dan J. Stein[19], Danai Dima[20,21], David E. J. Linden[22,23,24,25], David M. A. Mehler[22,23,26], Edith Pomarol-Clotet[27], Elena Pozzi[28,29], Elisa Melloni[30], Francesco Benedetti[30], Frank P. MacMaster[31], Hans J. Grabe[32], Henry Völzke[33], Ian H. Gotlib[34], Jair C. Soares[9,10], Jennifer W. Evans[35], Kang Sim[36,37,38], Katharina Wittfeld[32,39], Kathryn Cullen[13], Liesbeth Reneman[7], Mardien L. Oudega[6], Margaret J. Wright[40], Maria J. Portella[41], Matthew D. Sacchet[42], Meng Li[11], Moji Aghajani[6,43], Mon-Ju Wu[9,10], Natalia Jaworska[44], Neda Jahanshad[15], Nic J. A. van der Wee[45], Nynke Groenewold[3], Paul J. Hamilton[46,47], Philipp G. Sämann[48], Robin Bülow[49], Sara Poletti[29], Sarah Whittle[50], Sophia I. Thomopoulos[15], Steven J.A. van, der Werff[51], Sheri-Michelle Koopowitz[3], Thomas Lancaster[21,22], Tiffany C. Ho[52,53], Tony T. Yang[52], Zeynep Basgoze[13], Dick J. Veltman[6], Lianne

Schmaal[28], Paul M. Thompson[15], and Roberto Goya-Maldonado[1,*], for the ENIGMA Major Depressive Disorder working group[54]

Affiliations:

[1] Laboratory of Systems Neuroscience and Imaging in Psychiatry (SNIP-Lab), Department of Psychiatry and Psychotherapy, University Medical Center Goettingen (UMG), Georg-August University, Von-Siebold-Str. 5, 37075 Goettingen, Germany;

[2] SoCAT Lab, Department of Psychiatry, School of Medicine, Ege University, Izmir, Turkey;

[3] Department of Psychiatry & Mental Health, University of Cape Town, Cape Town, South Africa;

[4] Center for Neuroscience, University of Pittsburgh, Pittsburgh, PA, USA; Center for Neural Basis of Cognition, University of Pittsburgh, Pittsburgh, PA, USA;

[5] Department of Biomedical Sciences of Cells and Systems, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands;

[6] Department of Psychiatry, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Neuroscience, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands;

[7] Amsterdam University Medical Centers, location AMC, Department of Radiology and Nuclear Medicine, Amsterdam, the Netherlands;

[8] Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne, Parkville, Victoria, Australia;

[9] Louis A. Faillace, MD, Department of Psychiatry and Behavioral Sciences, The University of Texas Health Science Center at Houston, Houston, TX, USA;

[10] Center Of Excellence On Mood Disorders, Louis A. Faillace, MD, Department of Psychiatry and Behavioral Sciences at McGovern Medical School, The University of Texas Health Science Center at Houston, TX, USA;

[11] Department of Psychiatry and Psychotherapy, Jena University Hospital, Jena, Germany;

[12] Department of Psychology, University of Minnesota, Minneapolis, MN, USA;

[13] Department of Psychiatry and Behavioral Science, University of Minnesota Medical School, Minneapolis, MN, USA;

[14] Section on the Neurobiology and Treatment of Mood Disorders, National Institute of Mental Health, Bethesda, MD, USA;

[15] Imaging Genetics Center, Mark & Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Marina del Rey, CA 90274, USA;

[16] Department of Biomedical Sciences, Florida Stateu University, Tallahassee FL, USA;

[17] School of Psychology, University of East London, London, UK;

[18] Centre for Affective Disorders, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK;

[19] SA MRC Research Unit on Risk & Resilience in Mental Disorders, Department of Psychiatry & Neuroscience Institute, University of Cape Town, Cape Town, South Africa;

[20] Department of Psychology, School of Arts and Social Sciences, City, University of London, London, UK;

[21] Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

[22] Cardiff University Brain Research Imaging Center, Cardiff University, Cardiff, UK;

[23] MRC Center for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK;

[24] Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK;

[25] School of Mental Health and Neuroscience, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, 6229 ER, The Netherlands;

[26] Department of Psychiatry, Psychotherapy and Psychosomatics, Medical School, RWTH Aachen University, Germany;

[27] FIDMAG Germanes Hospitalàries Research Foundation, Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Barcelona, Catalonia, Spain

[28] Orygen, The National Centre of Excellence in Youth Mental Health, Parkville, VIC, Australia;

[29] Centre for Youth Mental Health, The University of Melbourne, Parkville, VIC, Australia;

[30] Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milano, Italy;

[31] Departments of Psychiatry and Pediatrics, University of Calgary, Calgary, AB, Canada;

[32] Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany;

[33] Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany;

[34] Department of Psychology, Stanford University, Stanford, CA, USA;

[35] Experimental Therapeutics and Pathophysiology Branch, National Institute for Mental Health, National Institutes of Health, Bethesda, MD, USA;

[36] West Region, Institute of Mental Health, Singapore;

[37] Yong Loo Lin School of Medicine, National University of Singapore, Singapore;

[38] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore;

[39] German Center for Neurodegenerative Diseases (DZNE), Site Rostock/ Greifswald, Greifswald, Germany;

[40] Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia;

[41] Sant Pau Mental Health Research Group, Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau, Barcelona, Catalonia, Spain. CIBERSAM, Madrid, Spain;

[42] Meditation Research Program, Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA;

[43] Institute of Education & Child Studies, Section Forensic Family & Youth Care, Leiden University, Leiden, The Netherlands;

[44] Department of Psychiatry, McGill University, Montreal, Quebec, Canada;

[45] Deparment of Psychiatry, Leiden Institute for Brain and Cognition and Theme NeuroscienceLeiden University Medical Center, Netherlands;

[46] Center for Social and Affective Neuroscience, Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden;

[47] Center for Medical Imaging and Visualization, Linköping University, Linköping, Sweden;

[48] Max Planck Institute of Psychiatry, Munich, Germany;

[49] Institute for Radiology and Neuroradiology, University Medicine Greifswald, Greifswald, Germany;

[50]Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne & Melbourne Health, Melbourne, VIC, Australia;

[51] Department of Psychiatry, Leiden University Medical Center, Leiden, Netherland;

[52] Department of Psychiatry and Behavioral Sciences, Division of Child and Adolescent Psychiatry, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA;

[53] Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA;

[54] https://enigma.ini.usc.edu/ongoing/enigma-mdd-working-group/

***Corresponding author:**
PD Dr. Roberto Goya-Maldonado
Laboratory of Systems Neuroscience and Imaging in Psychiatry (SNIP-Lab)
Department of Psychiatry and Psychotherapy
University Medical Center Göttingen (UMG)
Von-Siebold Str. 5, 37075 Göttingen
e-mail: roberto.goya@med.uni-goettingen.de

Contributions

RGM and VB conceptualized and developed the analysis pipeline, which was approved by ENIGMA MDD working chair LS, co-chair DJV, ENIGMA PI PMT. VB performed all the analyses mentioned in the manuscript and RGM closely supervised them. TEG and EP helped collecting and preparing the data from all participating cohorts. All authors participated in collecting and preprocessing data from their respective sites, reviewed and provided intellectual contribution to the manuscript.

Abstract

Machine learning (ML) techniques have gained popularity in the neuroimaging field due to their potential for classifying neuropsychiatric disorders. However, the diagnostic predictive power of the existing algorithms has been limited by small sample sizes, lack of representativeness, data leakage, and/or overfitting. Here, we overcome these limitations with the largest multi-site sample size to date (n=5,356) to provide a generalizable ML classification benchmark of major depressive disorder (MDD). Using brain measures from standardized ENIGMA analysis pipelines in FreeSurfer, we were able to classify MDD vs healthy controls (HC) with around 62% balanced accuracy, but when harmonizing the data using ComBat balanced accuracy dropped to approximately 52%. Similar results were observed in stratified groups according to age of onset, antidepressant use, number of episodes and sex. Future studies incorporating higher dimensional brain imaging/phenotype features, and/or using more advanced machine and deep learning methods may achieve more encouraging prospects.

Introduction

Major depressive disorder (MDD) is a psychiatric disorder with great impact on society, with a lifetime prevalence of 14% [181], often resulting in reduced quality of life [182] and increased risk of suicide for those affected [5]. Considering the possibility of treatment resistance [183] and accelerated brain aging [184], early recognition and implementation of effective treatments are critical. Unfortunately, there are no reliable biomarkers to date to diagnose MDD, to predict its highly variable natural progression or response to treatment [185]. Until now, the diagnosis of MDD relies exclusively on self-reported symptoms in clinical interviews, which - despite great efforts - present risk of misdiagnosis due to subjectivity and limited specificity of some symptoms, especially in the early stage of mental disorders. Furthermore, comorbid conditions such as substance use disorders, anxiety spectrum disorders [186], and other mental and somatic diseases [187] may contribute to the difficulty of correctly diagnosing and treating MDD.

With modern neuroimaging techniques such as magnetic resonance imaging (MRI), it has become possible to investigate cortical and subcortical brain alterations associated with MDD with high spatial resolution. Numerous studies reveal structural brain differences in MDD compared to healthy controls (HC) [71], [83], [92], [188], [189], with patients presenting, on average, smaller hippocampal volumes as well as lower cortical thickness in the insula, temporal lobes, and orbitofrontal areas. However, inference at the group level and small effect sizes preclude clinical application. Analytic tools such as machine learning (ML) that allow multivariate combinations of brain features and enable inference at the individual level may result in better discrimination between MDD patients and HC, thereby potentially providing clinically relevant biomarkers for MDD.

Current literature shows MRI-based MDD classification accuracies ranging from 53% to 91% [127], [190] with inconsistencies regarding which regions are the most informative for the classification. This lack of consensus in the literature raises concerns regarding the generalizability of the classification methods and their related findings. A major contributor to high variability in classification performances is sample size [142], [143]. Specifically smaller samples of the test data set tend to show more extreme results in both directions [142], whereas studies with larger sample sizes in the test set tend to converge to an accuracy of around 60% [143]. In the presence of publication bias, which favors the reporting of overestimations, published literature can quickly accumulate inflated results [191]. Further, overestimations in

the neuroimaging field [128], [192] may also be driven by data leakage, which refers to the use of test data in any part of the training process.

Another factor contributing to inconsistencies in results is the heterogeneity of samples in relation to demographic and clinical characteristics, which plays a significant role both in MDD and in the general population [184], [193], [194]. As large representative samples within a single cohort is difficult (e.g., due to financial cost, access to patient population, etc.), there is a growing interest in performing multi-site mega-analyses to address these issues.

ENIGMA MDD is a large-scale worldwide consortium, which curates and applies standardized analysis protocols to MRI and clinical/demographic data of MDD patients and HC from 52 independent sites from 17 countries across 6 continents (for review, [195]). Such large-scale approaches with global representation are necessary for identifying brain alterations associated with MDD that are realistic, reliable, and generalizable [196]. Therefore, we consider data from different international cohorts included in ENIGMA MDD a powerful and efficient resource to benchmark the robustness of representative examples of shallow linear and non-linear ML algorithms. Such algorithms include support vector machines (SVM), logistic regression with least absolute shrinkage and selection operator (LASSO) and ridge regularization, elastic net, and random forests. An additional advantage of ENIGMA MDD is that the inclusion of thousands of participants allows the stratification of several important factors related to cortical and subcortical brain alterations in MDD such as sex, age of MDD onset, number of depressive episodes, and antidepressant use. However, unifying multi-site data presents challenges. The global group differences between cohorts - referred to here as a site effect - may arise from different MR acquisition equipment and acquisition protocols [197], and/or demographic and clinical factors [198], [199]. Ignoring the site effect may lead to construction of suboptimal less-generalizable classification models [150], hindering the generalizability of the results. Along these lines, a commonly used strategy to mitigate site effect is to apply a harmonization technique such as ComBat [200]. Adopted from genomic studies, NeuroComBat estimates and statistically corrects for (harmonizes) differences in location (mean) and scale (variance) across different cohorts, while preserving or perhaps even enhancing the effect size of the variables of interest [201]–[203]. There are only a few studies attempting large sample multi-site MDD classification using structural brain metrics [142], [143]; however, site effects were not addressed in their analyses.

The main goal of this study was to establish a benchmark for classification of MDD vs HC based on structural cortical and subcortical brain measures in the largest sample to date. We profiled the classification performance of representative examples of linear and shallow non-linear models, including SVM with linear and rbf kernels with and without feature selection (PCA, t-test), logistic regression with LASSO/ridge regularization, elastic net and random forest. The model's performance is estimated via balanced accuracy, area under the receiver operating characteristic (AUC), sensitivity and specificity. We hypothesized that all models would be able to classify MDD vs HC with balanced accuracy higher than random chance, based on provided brain measures. We pooled preprocessed structural data from ENIGMA MDD participants, including 5,365 subjects (2,288 MDD and 3,077 HC) from 30 cohorts worldwide. As we were equally interested in general structural brain alterations in MDD as well as the generalizability of classification performance in sites unseen in the training phase, the data were split according to two strategies. First, age and sex (Splitting by Age/Sex) were evenly distributed across all cross-validation (CV) folds, where each fold is used as a test set once and the rest of folds is used as a training set iteratively. Second, the sites (Splitting by Site) were kept whole across CV folds, so the algorithms were trained and tested on different sets of cohorts, resulting in large between-sample heterogeneity of training and test sets, potentially resulting in lower classification performance [204], especially if large site effects are present. Because MDD is a highly heterogeneous diagnosis - and previous work from ENIGMA MDD [71], [92] has identified distinct alterations in different clinical subgroups - we also stratified MDD based on sex, age of onset, antidepressant use, and number of depressive episodes to investigate whether classification accuracy could be improved when considering more homogenous subgroups. Additionally, we investigated which brain areas were most relevant to classification performance.

In summary, we expected that (1) All models would correctly classify MDD above chance level, (2) Splitting by Site would yield lower performance versus Splitting by Age/Sex, (3) Application of ComBat would improve classification performance for all models, and (4) Stratified analyses according to demographic and clinical characteristics would yield higher classification performance. We also explored the impact of other approaches to remove site effects (ComBat-GAM [205] and CovBat [206]) from structural brain measures prior to feeding these measures into the classification models.

Results

Participants and Data Splitting

From 5,572 participants, 207 were excluded due to less than 75% of combined cortical and subcortical features being provided, resulting in 5,365 subjects (2,288 MDD and 3,077 HC) used in the analysis.

Substantial differences in age (87% of pairwise comparisons between cohorts were significant, t-test, $p<0.05$) and sex (54%, t-test, $p<0.05$) distribution exist in the investigated cohorts (Table 1, Supplementary Table 4). In the Splitting by Age/Sex strategy, all cohorts were evenly distributed across the folds, resulting in a similar number of subjects in each of fold (Table 2 left). In the Splitting by Site strategy, entire cohorts were kept into single folds, this time balancing the total number of subjects in each fold as close as possible (Table 2 right). This resulted in an irregular number of participants in each fold, with some folds containing only one of the larger cohorts (e.g., SHIP-T0, SHIP-S2, MPIP) and others containing multiple smaller cohorts.

Full Data Set Analysis

The classification performance of all models was similar and is presented in Table 3. When sites were evenly distributed across all CV folds (Splitting by Age/Sex), the highest balanced accuracy of 0.639 was achieved by SVM with rbf kernel, when trained using all cortical and subcortical features. The application of ComBat harmonization resulted in a performance drop of all models close to chance level. This pattern of lower classification performance, when ComBat was applied, was also observed across other classification metrics (see Supplementary Table 5-7). Yet specificity was found to be up to 10% higher than sensitivity, possibly related to potential imbalances in ratio MDD to HC and its effect on the classification. For the Splitting by Site strategy, classification performances did not change significantly based on whether the ComBat harmonization was performed or not. Balanced accuracy was close to random chance, indicating that the models were not able to differentiate MDD subjects from HC. The application of ComBat did not result in higher classification accuracies (Table 3). By exploring the classification performances measured on only a subset of cortical and subcortical features, we observed very similar results with classification around chance level. Similarly, there was no improvement when more sophisticated harmonization algorithms such as ComBat-GAM and CovBat were applied (see Supplementary Table 8).

When no harmonization step was applied, the choice of CV splitting strategy affected all measures of classification performance. Splitting by Age/Sex strategy yielded a balanced accuracy above 0.60 compared to roughly 0.51 accuracy for the Splitting by Site strategy. The ComBat harmonization step evened the classification performance of algorithms for the different splitting strategies, both being close to random chance. Information on the balanced accuracy changes via ComBat performing leave-one-site-out CV, can be found in Supplementary Table 9.

As the performance of the models were similar across all conditions, we assessed the weights of SVM with linear kernel to investigate, which regions contributed the most to the classification. The performance of SVM with and without application of ComBat was primarily driven by roughly the same set of cortical features, which could be observed by examining the feature weights. Feature weights of the SVM with linear kernel are presented in Figures 1 and 2. Even though the harmonization step affected the weights of the features, most of the informative features (with absolute weight >0.1) remained present. Cortical thickness features had greater weights compared to cortical surface areas, among which the left caudal middle frontal, left inferior parietal, left and right inferior temporal, left medial orbitofrontal, left postcentral, left precuneus, left superior frontal, right lingual, right paracentral, and right superior temporal regions were informative with and without the harmonization step. In the case of the regional surface areas, left and right cuneus, left inferior temporal, left medial orbitofrontal, left postcentral, and right precentral regions were found to be most informative for classification. Among subcortical volumes, no features remained informative after removing site effect via ComBat.

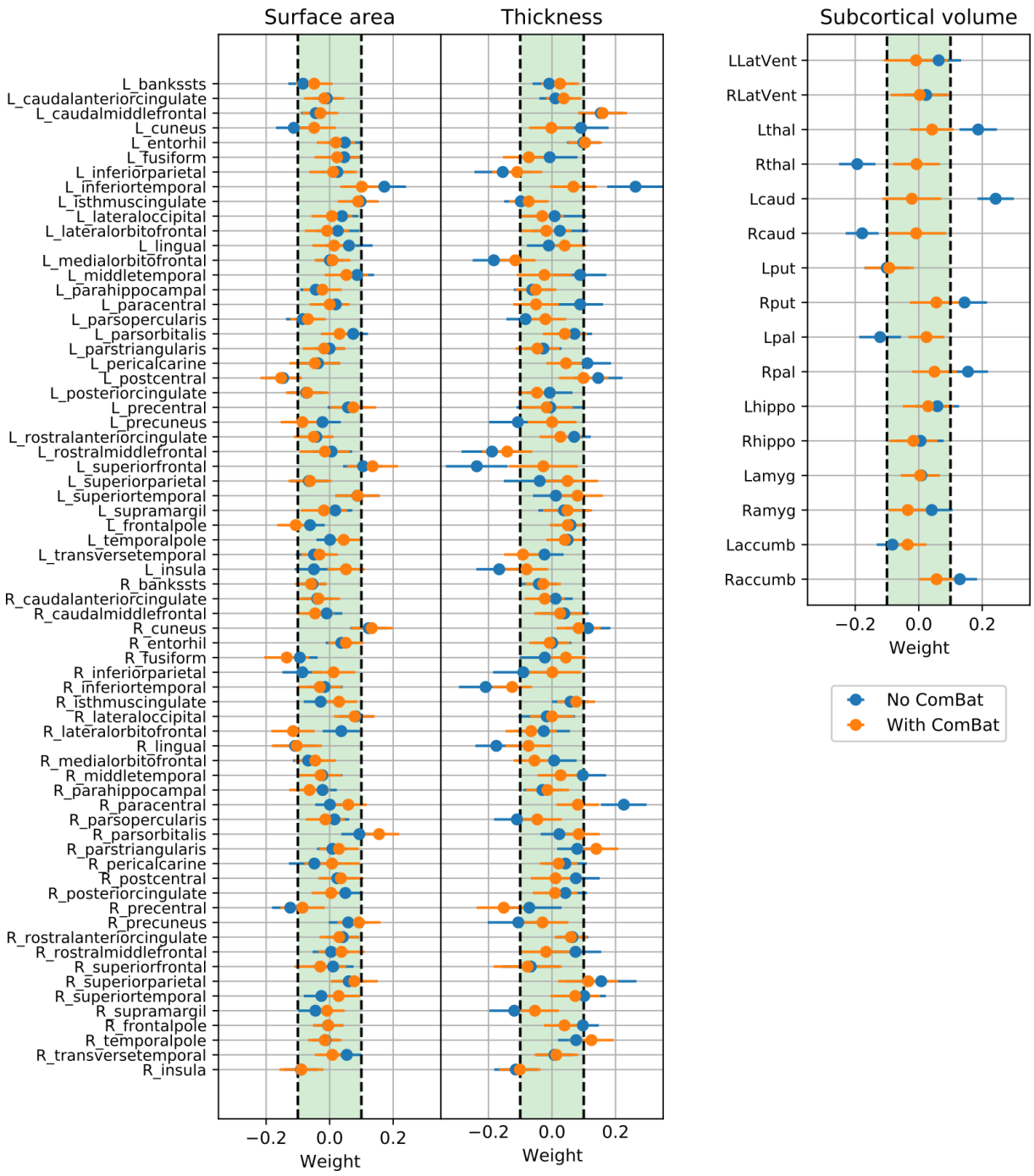


Figure 1: Feature weights of support vector machines (SVM) with the linear kernel. To assess the decision-making of SVM to differentiate subjects with major depressive disorder (MDD) from healthy controls (HC), we investigate the importance of the structural brain features by looking at the corresponding feature weights for the regional cortical surface areas, cortical thicknesses and subcortical volumes. The horizontal bars indicate the 95% confidence interval calculated using percentile method via bootstrapping.

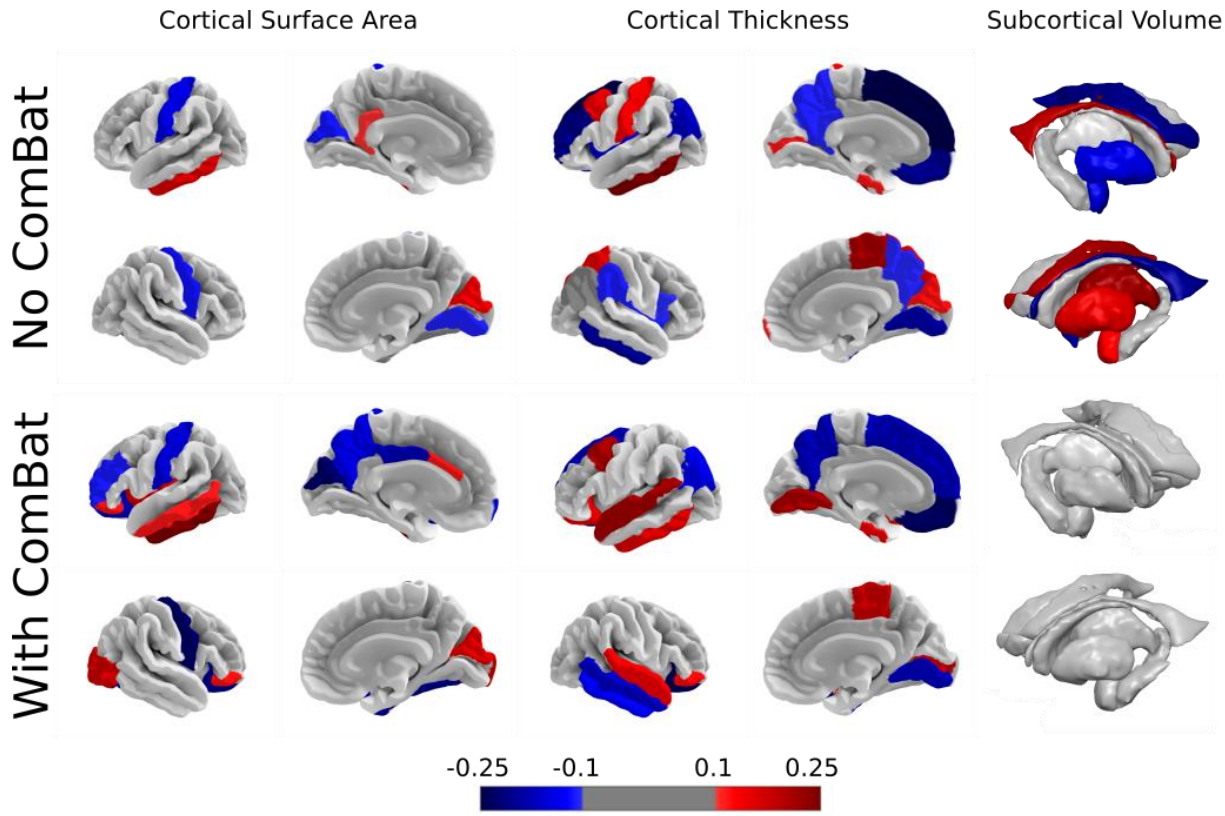


Figure 2: The most informative features for classification including regional cortical surface areas, thicknesses and subcortical volumes, trained on the whole data set without and with ComBat harmonization. Increased and decreased feature weight values in the major depressive disorder (MDD) group are represented by red and blue colormap, respectively.

Data Stratification

Next, we investigated the classification performance of models trained and tested on stratified data by demographic and clinical characteristics. The general pattern of the highest accuracy achieved by Splitting by Age/Sex strategy without ComBat and the significant drop in the accuracy when ComBat is applied was observed in all stratified analyses (below). In the Splitting by Site strategy, the classification performance did not change significantly when ComBat was applied. Information on the feature weights may be found in Supplementary Figures 1-4.

Males vs females

The number of male subjects is 2,131 and female subjects is 3,227 (7 male participants from the Episca cohort were not considered as we could not split them into 10 CV folds). In the Splitting by Age/Sex strategy without the harmonization step, the highest balanced accuracy of

0.632 was achieved when trained and tested on males - compared to maximum of 0.585 for females. When ComBat was applied, the accuracy dropped to 0.530 for males and to 0.529 for females, showing that there were minimal differences in classification results for males and females. For Splitting by Site, the accuracy did not change depending on the use of ComBat for both males (0.513 to 0.506) and females (0.519 to 0.517). Nevertheless, different brain regions were found important for classification in subgroups. In general, more features were found to be important for classification for males compared to females; this is especially noticeable for the regional surface areas (Supplementary Figure 1).

Age of onset

For Splitting by Age/Sex, when only patients first diagnosed in adolescence were included in the analysis, yielding 3,794 subjects in total, an accuracy of 0.626 was achieved, compared to 0.623 when patients who were first diagnosed in adulthood were analyzed. These accuracies dropped to 0.548 and 0.521 respectively, when ComBat was applied. In the Splitting by Site strategy, the balanced accuracy metrics did not change substantially for both subgroups: 0.541 to 0.544 for the adolescent-onset group and 0.546 to 0.518 for the adult-onset group, highlighting the absence of significant differences between these groups (Supplementary Figure 2).

Antidepressant use vs antidepressant free (at the time of MR scan)

Both subgroups showed a drop in balanced accuracy when ComBat was applied. In case of Splitting by Age/Sex, it reduced from 0.564 to 0.529 for the antidepressant-free subgroup (4,408 subjects) and from 0.716 to 0.534 for the antidepressant subgroup (3,988 subjects). When Splitting by Site, the balanced accuracy metrics did not change significantly for any of the subgroups when ComBat was used. For the antidepressant-free subgroup, it decreased from 0.564 to 0.528, while for the antidepressant group, it dropped from 0.560 to 0.483 (Supplementary Figure 3).

First episode vs recurrent episodes

Similarly, a drop in accuracy was observed when the data set was stratified based on the number of depressive episodes with vs without ComBat. In Splitting by Age/Sex, the balanced accuracy for the first episode subgroup dropped from 0.559 to 0.518 when ComBat was applied. For individuals with more than one episode, the balanced accuracy decreased from 0.644 to 0.520 with ComBat. In the Splitting by Site strategy, the algorithm's performance was not majorly

affected by ComBat in the single episode subgroup, yielding 0.482 to 0.512 in balanced accuracy and an insignificant drop from 0.521 to 0.505 for the recurrent episodes subgroup (Supplementary Figure 4).

Discussion

In this work, we benchmarked ML performance on the largest multi-site data set to date, using regional cortical and subcortical structural information for the task of discriminating patients with MDD vs HC. We applied shallow linear and non-linear models to 152 atlas-based features of 5,365 subjects from the ENIGMA MDD working group. To investigate brain characteristics common to MDD, as well as realistic classification metrics for unseen sites, we used two different data splitting approaches. Balanced accuracy was up to 63%, when data was split into folds according to *Splitting by Age/Sex*, and up to 51%, when data was split into folds according to *Splitting by Site* strategy. The harmonization of the data via ComBat evened the classification performance for both data splitting strategies, yielding up to 52% of balanced accuracy. This classification level implies that initial differences in performances were due to the site effects, most likely stemming from differences in MRI acquisition protocols across sites. Lastly, the data set was stratified based on demographic and clinical factors, but we found only minor differences in terms of classification performances between subgroups.

Data Splitting and Site Effect

Splitting of the data plays an important role in formulating and testing the hypotheses as well as validating them. As shown in [207], different data splitting techniques in combination with machine and deep learning algorithms in medical mega-analytical studies may introduce unwanted biases influencing classification or regression performances. Here we aimed to consider two data splitting paradigms: Splitting by Age/Sex and Splitting by Site. With Splitting by Age/Sex, we investigated general MDD alterations in contrast to HC using ML methods to obtain unbiased results regarding these important demographic factors. When we look at the weights of the SVM with linear kernel estimated on the entire data set, they correspond to the performance from Splitting by Age/Sex, as every CV fold contains all sites and demographically corresponds closely to the whole data set. With Splitting by Site, we wanted to see if the knowledge learned in one subset of cohorts could be translated to unseen cohorts - this can only be realistically measured when data is split according to the site it belongs to. To

the best of our knowledge, this is the first study to systematically emphasize differences in MDD vs HC classification performance in the context of data splitting strategies and the impact of ComBat in these strategies. The balanced accuracy of algorithms trained on data from Splitting by Age/Sex was up to 10% higher compared to Splitting by Site, confirming our expectations. This is a common trend in multi-site neuroimaging analyses [208], which indicates site effect and emphasizes how the nuances in data splitting strategies can strongly influence the classification performance. The presence of the site effect was additionally confirmed by training the SVM model to classify subjects according to their respective site, yielding substantially higher balanced accuracy compared to the main task of MDD vs HC classification (see Supplementary section "Harmonization methods"). The possibility that the site effect still reflected the demographic differences across cohorts, as cortical and subcortical features undergo substantial changes throughout lifespan [205] and differ between males and females [193], [194], was not supported. Regressing out these sources of demographic information did not significantly change the level of classification when predicting site belonging . According to our results, a major source of the site effect comes from the different scanner models and acquisition protocols, since we achieved the highest accuracy when attempting to classify scanner type (see Suppl. "Harmonization methods").

In addition to scanning differences, demographic and diagnostic characteristics distribution were different across the sites. Therefore, we explored if balancing the sample in terms of age and sex distributions would lead to higher classification performance. However, balancing of age/sex distributions across sites did not improve classification performance in Splitting by Site (balanced accuracy 52.6%/50.7% without/with ComBat). Thus, balancing age and sex did not contributed to better performance. As the MDD/HC ratio also varied across sites, an influence of site affiliation to the main MDD vs HC task could exist. Therefore, we additionally explored if the classification performance would drop to random level by equalizing the MDD/HC ratio in every site before splitting the data according to Splitting by Age/Sex. Sites without HC were discarded from this analysis. Indeed, we observed a substantial drop of the balanced accuracy from 61% to 53% with 1:1 MDD to HC ratio, confirming our assumption of likely incorporation of the site affiliation in the diagnosis classification.

Building on this, ComBat was able to remove the site effect, as all classification models could not differentiate between sites after its application. Subsequently, there were no differences between classification results across splitting approaches, with around 0.52 in balanced accuracy. Such a low accuracy – close to random chance – is consistent with another large

sample study based on two cohorts [143]. In their study, self-reported current depression was speculated as a reason for low accuracy, but this possibility is unlikely explaining our classification results. Moreover, similar classification levels in our and their study support the notion that a more balanced ratio between classes is not the main aspect behind the low power of discrimination.

Similar to ComBat, other more sophisticated harmonization methods such as ComBat-GAM and CovBat were able to remove site effect, but did not improve the balanced accuracy in Splitting by Site strategy. We cannot exclude the possibility that ComBat-like harmonization tools may overcorrect the data and remove weaker group differences of interest [209]. Hence, encouraging such evaluations in large data sets as well as implementing new methods to be tested [210], [211] on both the group and the single subject prediction level could be of great benefit for the imaging community.

Machine Learning Performance

In our study, the selection of shallow linear and non-linear classification algorithms was guided its low computational complexity and robustness. According to previous studies [127], [143], SVM is the most commonly and successfully used algorithm in previous analyses. We have tested other commonly used linear ML algorithms, such as logistic regression with LASSO, logistic ridge regression and elastic net logistic [127], [212], [213]. Given that logistic regression models already have an in-built feature selection procedure, we also included feature selection algorithms such as the two-sample t-test and PCA [175], [214], [215], for a fair comparison with SVM. Lastly, we included kernel SVM and random forest as representative shallow non-linear models. There was no single winner with a significantly higher classification performance across all algorithms, with a balanced accuracy up to 64%, when applied in data split by age/sex, and up to 53%, when split according to subsets of site. A similar trend was observed with AUC. In general, specificity was up to 5 % higher than sensitivity, possibly because of the imbalanced MDD/HC data sets, even when the impact of both classes was weighted by its ratio during the training.

Considering such a low balanced accuracy, future studies could apply more sophisticated classification methods such as Convolutional Neural Networks [129], which are able to detect nonlinear interactions between all the features as well as to consider spatial information of the given features. As it was demonstrated previously on both real and simulated data [130],

regressing out covariates can lead to lower classification performance, therefore one could use an importance weighting instead. Another option would be to include other data modalities such as vertex-wise cortical and subcortical maps [216], [217] or even voxel-wise T1 images to capture even more fine-grained changes [97], which are also present in shapes of subcortical structures[93] or diffusion MRI[218]. A recent resting-state fMRI multi-site study by Qin [164] reported an accuracy of 81.5%. Thus, integration of structural and functional data modalities may result in even higher classification performances.

Predictive Brain Regions

Our results do not support the hypothesis that MDD can be discriminated from healthy controls by regional structural features; classification performance, when site effects were removed, was close to chance level. Nevertheless, during investigation of the most discriminative regions, even after ComBat, we found an overlap with previously reported MDD-related regions. Multiple cortical and subcortical regions were found as the most discriminative between MDD and HC. Most of the cortical regions were identified in previous ENIGMA MDD work [71], which overlaps with our study set of cohorts. Shape differences in left temporal gyrus were reported previously in a younger population with MDD [219]. Left postcentral gyrus and right cuneus surface area were associated with severity of depressive symptoms, while left superior frontal gyrus, bilateral lingual gyrus and left entorhinal cortical thickness were decreased in MDD group [71], [220]. In a previous study, MDD subjects exhibited reduced cortical volume compared to HC [221]. Differences in orbitofrontal cortex between MDD and HC were also previously identified [71]. Overall, the effect sizes for case-control differences in these studies were small, which is in line with our current results showing low classification accuracies of these structural brain measures. Additionally, we also found increased thickness of left caudal middle frontal gyrus, right pars triangularis, right superior parietal and right temporal pole in MDD group, which was not previously reported. All subcortical volumes identified as informative for classification became uninformative after ComBat was applied, suggesting that either previously identified alterations in subcortical regions [92] cannot be directly used as MDD predictors at an individual level or ComBat removed differences significant for classification. One possible reason is that subcortical volumes tend to exhibit complex association with the age. Therefore, linear age regression might be an overly simplistic

representation of aging trajectories both in ComBat and residualization step. While some of the regions were found also to be predictive in the previous large sample MDD vs HC study from Stolicyn and colleagues [143], it is dificult to draw a consistent conclusion as they highlight the regions based on the selection frequency by the decision tree model, without reporting the direction of the modulation.

When models were trained and tested only on the subset of features in Splitting by Age/Sex, cortical thicknesses and subcortical volumes yielded higher balanced accuracy compared to cortical surface areas, which is consistent with the previous Enigma MDD meta-analysis, due to an overlap of study cohorts. When data was harmonized, there was no distinct subgroup of features providing more discriminative information. Together, we observed more changes in weights for cortical thicknesses and subcortical volumes after applying ComBat. One possibility is that differences are more pronounced in thickness than surface area, which is in line with previous findings from univariate approaches [222]. Another possibility is that differences in scanners and acquisition protocols may affect thickness features more strongly than surface areas, in line with previous works [223]. This is a very pertinent topic to be further investigated using multi-cohort mega-analyses on volumetric measures, particularly when the site effect is systematically considered.

Importantly, identified features correspond to the Splitting by Age/Sex strategy as the SVM model was trained on the whole data set with entirely mixed cohorts. While these regions were found to be informative according to the SVM with linear kernel , this model (and every other considered model) failed to differentiate MDD from HC on an individual level, thus one has to be cautious when interpreting these results. Structural alterations in myelination, gray matter, and curvature were found to be associated with MDD-associated genes (Li et al., 2021). Furthermore, a small sample study revealed MDD-related alterations in sulcal depth [224], while white matter topologically-based MDD classification led to up to 76% in accuracy [225]. Thus, the performance could be potentially elevated by integrating morphological shape features with white matter characterestics, such as sulcal depth and curvature, and myelination density as it led to improved performance when classifying sex and autism [155].

Data stratification

When the data set was stratified, we found substantial differences in balanced accuracies between the groups only for Splitting by Age/Sex strategy without harmonization step, yet these results were strongly influenced by the site effect. Harmonization step equalizes the accuracies

within all pairs of comparisons to a roughly chance probability. Same balanced accuracy was observed when the Splitting by Site strategy was used. This suggests that the demographic and clinical subgroups that we considered do not contain information to predict MDD on an individual level and do not differ in terms of the resultant accuracy, at least according to simplest ML models, despite the group level differences reported prior [71], [146]. Large sample meta-analysis of white mater characteristics that investigated similar subgroups also did not reveal significant differences [226], suggesting that the inclusion of these features into ML analysis might not be beneficial for classification improvement. Similarly, a large sample MDD classification study including structural and functional neuroimaging data did not reveal any significant differences between males and females [97]. However, we speculate that other clinically relevant stratifications such as the number of depressive episodes [164], [227] and course of disease [164], [228] using functional data in further large studies may improve classifications.

Conclusion

We benchmarked the classification of MDD vs HC using shallow linear and non-linear ML models applied to regional surface area features, cortical thickness features and subcortical volumes in the largest multi-site global data set to date. We systematically addressed the questions of general MDD characteristics and generalizability of models to perform on unseen sites by splitting the data according to demographic information (Splitting by Age/Sex) and site affiliation (Splitting by Site), which were complemented by ComBat harmonization. A classification accuracy up to 63% was achieved when all cohorts were present in the test set, which decreased down to 52% after ComBat harmonization. Here we have shown that most commonly used ML algorithms may not be able to differentiate MDD from HC on the single subject level using only structural morphometric brain data, even when trained on data from thousands of participants. Furthermore, the performance was not higher in stratified, clinically and demographically more homogeneous groups. Additional work is required to examine if more sophisticated algorithms also known as deep learning can achieve higher predictive power or if other MRI modalities such as task-based or resting state fMRI can provide more discriminative information for successful MDD classification.

Material and methods

Participant Sample

A total of 5,365 participants, 2,288 patients with MDD and 3,077 healthy controls, from 30 cohorts participating in the ENIGMA MDD working group, were included in the analyses. Information on sample characteristics, inclusion/exclusion criteria for each cohort can be found in Supplementary Table 1. Subjects with less than 75% of combined cortical and subcortical features and/or missing demographic/clinical information required for a particular analysis were excluded from the analysis. We implemented 75% as a reasonable cut-off value, which allowed us to accommodate a large amount of the available data without incurring biased model estimations. Furthermore, after exclusion of the subjects with less than 75% of existing data, total number of missing values was less than 10% from the remaining participants. According to the third guideline by Newman [229], i.e., "for construct-level missingness that exceeds 10% of the sample, ML and multiple imputation (MI) techniques should be used under a strategy that includes auxiliary variables and any hypothesized interaction terms as part of the imputation/estimation model", we performed data imputation by considering age and sex factors as „auxiliary variables". Missing cortical and subcortical features for the remaining subjects (2% of all data) were imputed by using multiple linear regression with age and sex of all subjects (regardless of diagnosis) as predictors, estimated for each cohort separately. All participating sites reported approval from their institutional review boards and local ethics committees and also obtained written informed consent for all participants.

Brain Imaging Processing

Structural T1-weighted 3D brain MRI scans of participating subjects were acquired from each site and preprocessed according to the rigorously validated ENIGMA Consortium protocols (http://enigma.ini.usc.edu/protocols/imaging-protocols/). Information on the MRI scanners and acquisition protocols used for each cohort can be found in Supplementary Table 2. To facilitate the ability to pool the data from different cohorts, cortical and subcortical parcellation was performed on every subject via the freely available FreeSurfer (Version 5.1,5.3, 6 and 7.2) software [230], [231]. Every cortical and subcortical brain parcellation was visually inspected as part of a careful quality check (QC) and statistically evaluated for outliers, according to the ENIGMA Consortium protocol (https://enigma.ini.usc.edu/protocols/imaging-protocols/). Cortical gray matter segmentation was based on the Desikan–Killiany atlas [167], yielding cortical surface area and cortical thickness measures for 68 brain regions (34 for each

hemisphere), resulting in 136 cortical features. Subcortical segmentation was based on the *Aseg* atlas [167], providing volumes of 40 regions (20 for each hemisphere), of which we included 16: lateral ventricle, thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and nucleus accumbens, bilaterally.

Data Splitting into Cross-Validation Folds

We applied two different strategies to split the data into training and test sets: *Splitting by Age/Sex* and *Splitting by Site*. For both strategies, data was split into 10 folds, 9 of which were used for the training and the remaining fold was used as a test set. This was repeated iteratively until each fold was used once as a test set, thus performing the 10-fold CV. We investigated the general differences in brain volumes that can characterize MDD by using the Splitting by Age/Sex strategy. In this way, the age and sex distribution as well as number of subjects between the folds were balanced to mitigate the effect of these factors on the classification performance. However, it should be noted that with each site represented in every CV fold the potential site effects in this strategy, if any, would be diluted between the folds, which would not represent a realistic clinical scenario, where a classification model likely has to generalize to unseen sites. Therefore, we used a second strategy, Splitting by Site, which would yield more realistic metrics of classification performance for unseen sites. Using this strategy, every site was present only in one fold, thus the model is always trained and tested on different sets of sites and sites were distributed across folds to balance the number of subjects in every fold as close as possible. In this scenario, potential site-specific confounders (e.g., different MR scanners/acquisition protocols, demographic and clinical differences, etc.) were not equally distributed between the training and test sets. In this way, we can fairly evaluate the generalizability from one cohort to another. Finally, to assess the performance estimates for each site, we explored leave-site-out CVs. Further details on both splitting strategies can be found in Supplementary Section “CV splitting strategies”.

Classification models

We have chosen representative examples of shallow linear and non-linear classification models to establish a benchmark of MDD vs HC classification. For the linear models, we selected SVM with linear kernel [138], and logistic regression with different types of regularization: L1 (LASSO), L2 (Ridge), and L1+L2 (Elastic Net)[232]. Both models are commonly used classification models used in neuroimaging [127] due to their low computational complexity. As regularization serves as an in-built feature selection algorithm, we evaluated SVM with

additional feature selection via PCA and t-test. As many classification tasks are not linearly separable, potentially including MDD vs HC, we additionally evaluated robust shallow non-linear models, including SVM with RBF kernel [233], and ensemble classification algorithm - random forest [234], [235]. While, other shallow linear/non-linear models were evaluated for MDD vs HC classification task previously [127], including linear discriminant analysis (LDA) [236], SVM with other non-linear kernels, a large sample benchmark analysis revealed no significant advantage of their application in the general neuroimaging setting[237].

Analysis Pipeline

After distributing the data into CV folds corresponding to the splitting strategies, 9 folds were used for the training, while the remaining fold was held out as a test set (Figure 3). CV folds were residualized normatively, partialling out the linear effect of age, sex and ICV from all cortical and subcortical features. In this step, age, sex and ICV regressors were estimated on the HC from training CV folds and applied to remove the effect of age, sex, and ICV fro, brain measures in the MDD training data and to all test data. After normalizing all features to have mean of zero and standard deviation of one based on the mean and standard deviation estimaes from the training set's initial features' distributions, training and test folds were used for training and performance estimation, respectively. Additionally, class weighting was performed to mitigate an unbalanced training set across classes. Models' hyperparameters were estimated in the training data via nested 10-folds cross-validation using grid search (random splits, for both Splitting by Site and Splitting by Age/Sex), before the performance was measured on the test data to avoid data leakage through the choice of hyperparameters. The list of hyperparameters that were adjusted can be found in Supplementary Table 3. We evaluated the performance of SVM with linear kernel, SVM with rbf kernel, logistic regression with LASSO regularization, logistic regression with ridge regularization, elastic net, and random forest by using balanced accuracy, sensitivity, specificity and AUC as performance metrics. For the model-level assessment [238], all models were also trained on the subset of features, i.e. only cortical surface areas, only cortical thicknesses and only subcortical volumes. Lastly, we investigated which features contributed most to the classification performance by looking at the decision-making of the most successful model, in line with established guidelines [238]. In case no performance differences across models were found, we reported the weights of the SVM with linear kernel as the representative classifier. These weights correspond to the classification performance of Splitting by Age/Sex strategy as all sites are used for weight's estimation. To

assess confidence intervals of the feature weights, we performed 599-bootstrap [239], [240] on the whole data set.

Further analyses were performed by stratifying the data according to demographic and clinical categories, including sex, age of onset (<21 years old vs >21 years old), antidepressant use (yes/no at time of scan), and number of depressive episodes (first episode vs recurrent episodes). The subjects with missing information on these factors were not included in these analyses, while they were still considered for the main analysis.

All the steps from CV folds to classification were repeated with feature specific harmonization of site effects via ComBat. Variance explained by age, sex and ICV was preserved in the cortical and subcortical features during harmonization step. The harmonized folds were then residualized normatively with all subsequent steps from the analysis without harmonization step. Furthermore, we compared ComBat with two modifications: ComBat-GAM and CovBat. More detailed description of ComBat, ComBat-GAM and CovBat as well as their implementation for both splitting strategies can be found in Supplementary section "Harmonization methods".

We used Python (version 3.8.8) to perform all calculations. All classification models and feature selection methods were imported from sklearn library (version 1.1.2). We modified ComBat script (https://github.com/Jfortin1/ComBatHarmonization) to incorporate ComBat-GAM (https://github.com/rpomponio/neuroHarmonize) and CovBat (https://github.com/andy1764/CovBat_Harmonization) in one function for both splitting strategies.

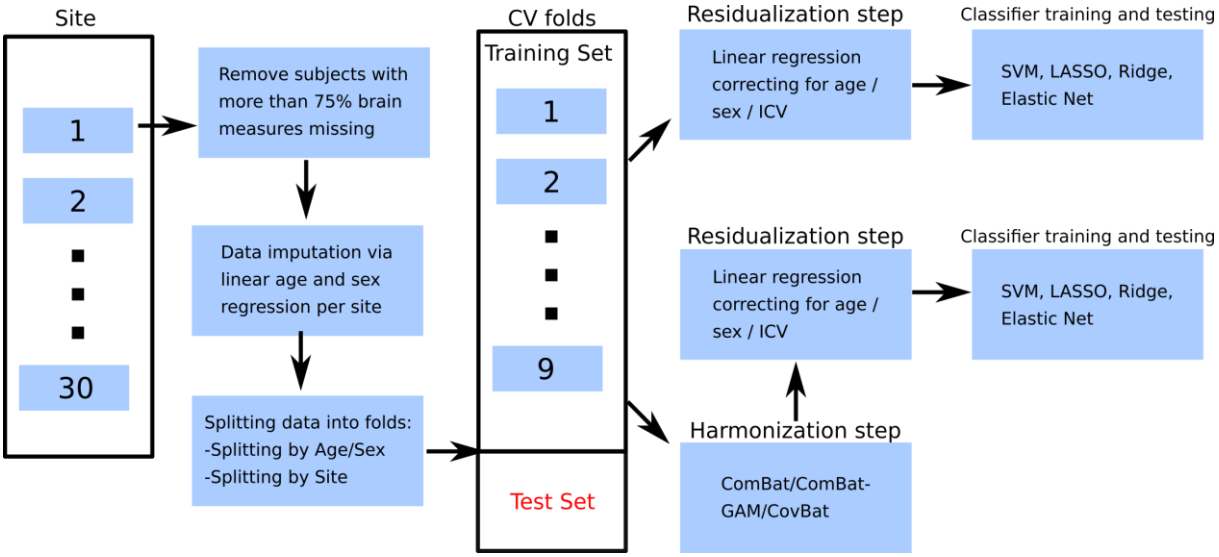


Figure 3: Detailed analysis pipeline. Initial data from all cohorts is split into training and test sets according to splitting strategies (Splitting by Age/Sex and Splitting by Site) after removing subjects with more than 75% missing data and data imputation steps. The corresponding training folds are then residualized directly to remove ICV, age and sex related effects and fed to the classification algorithms. In case of harmonization by ComBat, the residualization step takes place after the harmonization step is conducted. If training folds were harmonized by ComBat, the test fold was harmonized as well by using ComBat estimates from the training folds. Next, the test fold was residualized by using estimates obtained from the training folds. We estimated classification performance on the residualized test fold. This routine was performed iteratively for each combination of training and test folds.

Tables

Table 1: ENIGMA MDD participating cohorts in the study. Each cohort is presented with number of total subjects, number of patients with major depressive disorder (MDD) and healthy controls (HC), as well as their mean age (in years) and sex (number and % of females).

Cohort		Number of subjects	Age Mean (SD)	Number of Females (%)
AFFDIS	**Total**	79	39.75 (14.67)	36(46)
	HC	46	39.87 (14.29)	22(48)
	MDD	33	39.58 (15.18)	14(42)
Pharmo (AMC)	**Total**	51	29.37 (4.64)	51(100)
	HC	0	nan	nan
	MDD	51	29.37 (4.64)	51(100)
Barcelona-StPau	**Total**	94	46.66 (7.81)	72(77)
	HC	32	46.03 (8.00)	23(72)
	MDD	62	46.98 (7.68)	49(79)
CARDIFF	**Total**	40	46.55 (11.74)	27(68)
	HC	0	nan	nan
	MDD	40	46.55 (11.74)	27(68)
CSAN	**Total**	109	34.70 (12.88)	74(68)
	HC	49	33.20 (12.07)	34(69)

	MDD	60	35.92 (13.38)	40(67)
Calgary	**Total**	107	17.03 (4.12)	60(56)
	HC	52	15.81 (5.03)	29(56)
	MDD	55	18.19 (2.51)	31(56)
DCHS	**Total**	79	30.91 (6.71)	79(100)
	HC	61	31.49 (6.82)	61(100)
	MDD	18	28.94 (5.89)	18(100)
ETPB	**Total**	60	35.03 (9.86)	36(60)
	HC	26	33.88 (10.22)	16 (62)
	MDD	34	35.91 (9.48)	20(59)
Episca (Leiden)	**Total**	49	15.00 (1.54)	42(86)
	HC	30	14.73 (1.53)	26(87)
	MDD	19	15.42(1.46)	16(84)
FIDMAG	**Total**	69	47.22 (12.29)	44(64)
	HC	34	45.94 (11.49)	22(65)
	MDD	35	48.46 (12.90)	22(63)
Groningen	**Total**	41	44.27 (13.67)	30(73)
	HC	21	44.05 (13.96)	16(76)
	MDD	20	44.50 (13.34)	14(70)
Houston	**Total**	290	28.72 (16.30)	169(58)
	HC	186	26.76 (15.91)	105(56)
	MDD	104	32.23 (16.39)	64(62)
Jena	**Total**	107	46.76 (15.00)	52(49)
	HC	77	47.75 (15.93)	36(47)
	MDD	30	44.20 (11.92)	16(53)
LOND	**Total**	130	49.67 (8.62)	79(61)
	HC	61	51.72(7.87)	32(53)
	MDD	69	47.86(8.85)	47(68)
MODECT	**Total**	42	72.71 (9.25)	28(67)
	HC	0	nan	nan
	MDD	42	72.71 (9.25)	28(67)
MPIP	**Total**	548	48.66 (13.59)	315(57)
	HC	211	49.53 (13.02)	124 (59)
	MDD	337	48.12 (13.90)	191(57)
Melbourne	**Total**	245	19.42 (2.88)	130(53)
	HC	102	19.58 (2.97)	54(53)
	MDD	143	13.31 (2.80)	76(53)
Minnesota	**Total**	110	15.47 (1.89)	79(72)
	HC	40	15.68 (1.98)	26(65)
	MDD	70	15.36 (1.83)	53(76)
Moraldilemma	**Total**	70	18.81 (1.94)	70(100)
	HC	46	18.50 (1.75)	46(100)
	MDD	24	19.42 (2.14)	24(100)
NESDA	**Total**	219	38.11 (10.32)	145(66)
	HC	65	40.29 (9.67)	42(65)
	MDD	154	37.19 (10.45)	103(67)
QTIM	**Total**	386	22.08 (3.25)	267(69)
	HC	284	22.11 (3.30)	190(67)
	MDD	102	22.01 (3.11)	77(75)
UCSF	**Total**	163	15.46 (1.31)	91(56)
	HC	88	15.32 (1.28)	42(48)
	MDD	75	15.63 (1.33)	49(65)
SHIP_S2	**Total**	579	55.01 (12.57)	294(51)
	HC	443	55.44 (12.80)	198(45)
	MDD	136	53.59 (11.68)	96(71)
SHIP_T0	**Total**	1229	50.15 (13.69)	607(49)
	HC	919	50.50 (14.18)	405(44)
	MDD	310	49.12 (12.04)	202 (65)
SanRaffaele	**Total**	45	49.07 (13.51)	32(71)
	HC	0	nan	nan
	MDD	45	49.07 (13.51)	32(71)
Singapore	**Total**	38	39.50 (6.43)	18(47)
	HC	16	38.69 (4.59)	8(50)
	MDD	22	40.09 (7.43)	10(45)
Socat_dep	**Total**	179	37.85 (13.34)	161(90)
	HC	100	36.42 (13.57)	90 (90)
	MDD	79	39.66 (12.81)	71 (90)
StanfFAA	**Total**	32	32.71 (9.56)	32(100)
	HC	18	30.44 (9.96)	18(100)
	MDD	14	35.63 (8.14)	14(100)
StanfT1wAggr	**Total**	115	37.18 (10.27)	69(60)
	HC	59	37.24 (10.43)	36(61)
	MDD	56	37.11 (10.09)	33(59)

TIGER	**Total** **HC** **MDD**	60 11 49	15.63 (1.34) 15.18 (1.03) 15.73 (1.38)	38(63) 5(45) 33(67)
All sites	**Total** **HC** **MDD**	5365 3077 2288	39.84 (17.69) 40.82(18.09) 38.52 (17.05)	3227(60) 1706(55) 1521(66)

Table 2: Data splitting strategies. The differences in strategies are manifested in the distribution of age, sex, and diagnosis between cross-validation folds.

Splitting By Age/Sex				**Splitting By Site**			
Fold	**Age mean (SD)**	**Number of Females (%)**	**Number of subjects (%MDD)**	**Fold**	**Age mean (SD)**	**Number of Females (%)**	**Number of subjects (%MDD)**
1	39.98 (17.40)	322 (60)	536 (42)	1	50.15 (13.69)	607 (49)	1229 (25)
2	39.63 (17.81)	324 (60)	538 (42)	2	55.01 (12.57)	294 (51)	579 (23)
3	39.85 (17.57)	325 (60)	538 (43)	3	48.66 (13.59)	315 (57)	548 (61)
4	39.66 (17.94)	322 (60)	535 (39)	4	22.90 (4.97)	299 (72)	418 (28)
5	39.99 (17.56)	323 (60)	538 (44)	5	36.72 (19.69)	272 (60)	451 (51)
6	39.75 (17.25)	317 (60)	531 (43)	6	22.53 (10.92)	293 (65)	450 (68)
7	40.15 (17.89)	327 (60)	541 (42)	7	35.94 (12.96)	295 (71)	418 (59)
8	39.81 (17.93)	322 (60)	535 (44)	8	38.85 (12.66)	348 (81)	431 (45)
9	39.86 (17.73)	320 (60)	535 (44)	9	24.79 (16.16)	203 (54)	377 (42)
10	39.74 (17.80)	325 (60)	538 (43)	10	34.95 (15.45)	301 (65)	464 (55)

Table 3: Balanced accuracy measured on the entire data set, after being divided into cross-validation folds using the Splitting by Age/Sex and Splitting by Site strategies. We evaluated classification performances when models are trained on combined cortical and subcortical features , cortical thickness, cortical surface area, and subcortical volume. Furthermore, all models were trained/tested without and with ComBat harmonization.

Splitting by Age/Sex								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
SVM linear	0.616	0.524	0.577	0.504	0.572	0.518	0.602	0.524
SVM rbf	0.639	0.525	0.600	0.515	0.578	0.510	0.619	0.513
SVM + PCA	0.638	0.529	0.601	0.513	0.575	0.518	0.622	0.513
SVM + ttest	0.627	0.515	0.581	0.515	0.567	0.526	0.619	0.521
LASSO	0.612	0.524	0.583	0.499	0.578	0.516	0.596	0.518

Ridge	0.610	0.523	0.585	0.498	0.573	0.515	0.594	0.520
Elastic Net	0.609	0.523	0.584	0.500	0.569	0.517	0.593	0.520
Random Forest	0.613	0.507	0.593	0.514	0.574	0.509	0.611	0.511
Splitting by Site								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
SVM linear	0.512	0.519	0.498	0.495	0.499	0.506	0.506	0.521
SVM rbf	0.513	0.515	0.493	0.499	0.493	0.513	0.503	0.519
SVM + PCA	0.527	0.520	0.502	0.512	0.504	0.524	0.520	0.520
SVM + ttest	0.502	0.512	0.487	0.499	0.507	0.508	0.510	0.527
LASSO	0.513	0.517	0.491	0.489	0.508	0.513	0.507	0.512
Ridge	0.514	0.514	0.497	0.490	0.505	0.509	0.507	0.514
Elastic Net	0.513	0.514	0.498	0.489	0.503	0.514	0.507	0.514
Random Forest	0.518	0.506	0.495	0.501	0.491	0.503	0.519	0.501

Competing Interest

PMT and NJ received a research grant from Biogen, Inc., for research unrelated to this manuscript. HJG has received travel grants and speakers honoraria from Fresenius Medical Care, Neuraxpharm, Servier and Janssen Cilag as well as research funding from Fresenius Medical Care unrelated to this manuscript. JCS has served as a consultant for Pfizer,Sunovion, Sanofi, Johnson & Johnson, Livanova, and Boehringer Ingelheim. The remaining authors declare no conflict of interest.

Data availability

Authors are not allowed to share the data of participating sites to third parties inside or outside the ENIGMA MDD consortium. Some sites may provide data upon request.

Acknowledgements

ENIGMA MDD: This work was supported by NIH grants U54 EB020403 (PMT) and R01MH116147 (PMT) and R01 MH117601 (NJ & LS). AMC: supported by ERA-NET PRIOMEDCHILD FP 6 (EU) grant 11.32050.26. AFFDIS: this study was funded by the University Medical Center Goettingen (UMG Startfoerderung) and VB and RGM are supported by German Federal Ministry of Education and Research (Bundesministerium fuer Bildung und Forschung, BMBF: 01 ZX 1507, “PreNeSt - e:Med”). Barcelona-SantPau: MJP is funded by

the Ministerio de Ciencia e Innovación of the Spanish Government and by the Instituto de Salud Carlos III through a 'Miguel Servet' research contract (CP16–0020); National Research Plan (Plan Estatal de I + D + I 2016–2019); and co-financed by the European Regional Development Fund (ERDF). CARDIFF supported by the Medical Research Council (grant G 1100629) and the National Center for Mental Health (NCMH), funded by Health Research Wales (HS/14/20). CSAN: This work was supported by grants from Johnson & Johnson Innovation (S.E.), the Swedish Medical Research Council (S.E.: 2017–00875, M.H.: 2013–07434, 2019–01138), the ALF Grants, Region Östergötland M.H., J.P.H.), National Institutes of Health (R.D.: R01 CA193522 and R01 NS073939), MD Anderson Cancer Support Grant (R.D.: P30CA016672) Calgary: supported by Canadian Institutes for Health Research, Branch Out Neurological Foundation. FPM is supported by Alberta Children's Hospital Foundation and Canadian Institutes for Health Research. DCHS: supported by the Medical Research Council of South Africa. ETPB: Funding for this work was provided by the Intramural Research Program at the National Institute of Mental Health, National Institutes of Health (IRP-NIMH-NIH; ZIA-MH002857). Episca (Leiden): EPISCA was supported by GGZ Rivierduinen and the LUMC. FIDMAG: This work was supported by the Generalitat de Catalunya (2014 SGR 1573) and Instituto de Salud Carlos III (CPII16/00018) and (PI14/01151 and PI14/01148). Gron: This study was supported by the Gratama Foundation, the Netherlands (2012/35 to NG). Houst: supported in part by NIMH grant R01 085667 and the Dunn Research Foundation. LOND This paper represents independent research (BRCDECC, London) part-funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. MODECT: This study was supported by the Department of Psychiatry of GGZ inGeest and Amsterdam UMC, location VUmc. MPIP: The MPIP Sample comprises patients included in the Recurrent Unipolar Depression (RUD) Case-Control study at the clinic of the Max Planck Institute of Psychiatry, Munich, Germany. We wish to acknowledge Rosa Schirmer, Elke Schreiter, Reinhold Borschke, and Ines Eidner for MR image acquisition and data preparation, and Benno Pütz, and Bertram Müller-Myhsok for distributed computing support and the MARS and RUD Study teams for clinical phenotyping. We thank Dorothee P. Auer for initiation of the RUD study. Melbourne: funded by National Health and Medical Research Council of Australia (NHMRC) Project Grants 1064643 (Principal Investigator BJH) and 1024570 (Principal Investigator CGD). Minnesota the study was funded by the National Institute of Mental Health (K23MH090421; Dr. Cullen) and Biotechnology Research Center (P41 RR008079; Center for

Magnetic Resonance Research), the National Alliance for Research on Schizophrenia and Depression, the University of Minnesota Graduate School, and the Minnesota Medical Foundation. This work was carried out in part using computing resources at the University of Minnesota Supercomputing Institute. Moral dilemma: study was supported by the Brain and Behavior Research Foundation and by the National Health and Medical Research Council ID 1125504 to SLW. NESDA: The infrastructure for the NESDA study (www.nesda.nl) is funded through the Geestkracht program of the Netherlands Organisation for Health Research and Development (Zon-Mw, grant number 10–000–1002) and is supported by participating universities (VU University Medical Center, GGZ inGeest, Arkin, Leiden University Medical Center, GGZ Rivierduinen, University Medical Center Groningen) and mental health care organizations, see www.nesda.nl. QTIM: The QTIM data set was supported by the Australian National Health and Medical Research Council (Project Grants No. 496682 and 1009064) and US National Institute of Child Health and Human Development (RO1HD050735). UCSF: This work was supported by the Brain and Behavior Research Foundation (formerly NARSAD) to TTY; the National Institute of Mental Health (R01MH085734 to TTY; K01MH117442 to TCH) and by the American Foundation for Suicide Prevention (PDF-1-064-13) to TCH. SHIP: The Study of Health in Pomerania (SHIP) is part of the Community Medicine Research net (CMR) (http://www.medizin.uni-greifswald.de/icm) of the University Medicine Greifswald, which is supported by the German Federal State of Mecklenburg—West Pomerania. MRI scans in SHIP and SHIP-TREND have been supported by a joint grant from Siemens Healthineers, Erlangen, Germany and the Federal State of Mecklenburg-West Pomerania. This study was further supported by the EU-JPND Funding for BRIDGET (FKZ:01ED1615). SanRaffaele (Milano): Italian Ministry of Health, Grant/Award Number: RF-2011-02349921 and RF-2018-12367489 Italian Ministry of Education, University and Research (Miur). Number: PRIN - 201779W93T. Singapore: The study was supported by grant NHG SIG/15012. KS was supported by National Healthcare Group, Singapore (SIG/15012) for the project.SoCAT: Socat studies supported by Ege University Research Fund (17-TIP-039; 15-TIP-002; 13-TIP-054) and the Scientific and Technological Research Council of Turkey (109S134, 217S228). StanfFAA and StanfT1wAggr: This work was supported by NIH grant R37 MH101495. TIGER: Support for the TIGER study includes the Klingenstein Third Generation Foundation the National Institute of Mental Health K01MH117442 the Stanford Maternal Child Health Research Institute and the Stanford Center for Cognitive and Neurobiological Imaging TCH receives partial support from the Ray and Dagmar Dolby Family Fund.

Supplementary Materials

Supplementary Table 1: ENIGMA MDD Image acquisition and processing by cohort

Cohort	Scanner type	Sequence T1	FreeSurfer version	Slice orientation	Operating system
AFFDIS	3T Siemens Magnetom TrioTim	3D T1 (176 slices; TR = 2250 ms; TE = 3.26 ms; FOV 256; voxel size 1X1X1mm)	5,3	Sagittal	Linux CentOS
Pharmo (AMC)	3T Philips	T1 sequence details: 3D-TFE sequence TR= 9.7 ms, TE=4.6ms, matrix 192x192, voxel size = 0.875 x 0.875 x 1.2 mm; 120 slices. Axial plane. Philips 3T Ingenia 16 channel coil	5,3	Transverse	freesurfer-Linux-centos6_x86_64-stable-pub-v5.3.0
Barcelona	3T Philips Achieva	3D MPRAGE images (Whole-brain T1-weighted); TR=6.7ms, TE=3.2ms; 170 slices, voxel size 0.89X0.89X1.2 mm. Image dimensions 288X288X170; field of view: 256X256X204; slice thickness: 1.2 mm; with a sagittal slice orientation, T1 contrast enhancement, flip angle: 8º, grey matter as a reference tissue, ACQ matrix MXP = 256X240 and turbo-field echo shots (TFE) = 218.	6	Sagittal	Scientific Linux 5
Cardiff	A 3 Tesla whole body MRI system (General Electric, Milwaukee, USA) with an 8-channel head coil was used at the Cardiff University Brain Research Imaging Centre (CUBRIC).	High-resolution anatomical scan (Fast Spoiled Gradient-Recalled-Echo [FSPGR] sequence): 178 slices, TE=3 ms, TR=7.9 ms, voxel size=1.0×1.0×1.0 mm3, FA=15°, FOV=256×256	5,3		freesurfer-Linux-centos6_x86_64-stable-pub-v5.3.0
CSAN (Adf)	3T Siemens MAGNETOM PRISMA	Whole-head t1-weighted MPRAGE (TR = 2300 ms, TE = 2.34 ms, FOV 250 × 250 mm, voxel size = 0.9 × 0.868 × 0.868 mm, flip angle = 8°)	7.2	Sagittal	Ubuntu
Calgary	1.5T Siemens Magnetom Vision. 3T GE Discovery MR750	1.5T: A sagittal scout series was acquired to test image quality. 3D fast low angle shot (FLASH) sequence was used to acquire data from 124 1.5 mm-thick contiguous coronal slices through the entire brain (echo time = 5ms, repetition time = 25ms, acquisition matrix = 256 x 256 pixels, field of view = 24 cm and flip angle = 40°). 3T: Anatomical imaging acquisition parameters: axial acquisition, repetition time (TR), 2200 milliseconds (ms); echo time (TE), 3.04 ms; TI, 766, 780; flip angle, 13 degrees; 208 partitions; 256 × 256 matrix; and field of view, 256.	5,3	Dalhousie sample, coronal; Calgary sample, axial	MacOs Sierra
DCHS	3T Siemens Skyra	3D multi-echo MPRAGE, voxel size 1 mm x 1mm x 1.5mm, TR = 2530 ms, TE = 1.69 x 3.55 x 5.41 x 7.27ms, FOV: 256x256mm, flip angle = 7°	5.3	Sagittal	Linux-centos6_x86_64
ETPB	3T, GE HDx	Fast spoiled gradient recalled echo (FSPGR). Slice Thickness: 1.	5,3	Sagittal	Linux

		Repetition Time: 8.836. Echo Time: 3.496. Inversion Time: 450. Magnetic Field Strength: 3. Spacing Between Slices: 1. Echo Train Length: 1. Percent Sampling: 100. Percent Phase Field of View: 100. Pixel Bandwidth: 195.312. Reconstruction Diameter: 256. Acquisition Matrix: 0,256,256,0. In-plane Phase Encoding Direction: ROW. Flip Angle: 13			
EPISCA (Leiden)	3T Philips Achieva	a sagittal 3-dimensional gradient-echo T1-weighted image was acquired (repetition time = 9.8 ms; echo time = 4.6 ms; flip angle = 8°; 140 sagittal slices; no slice gap; field of view =256 × 256 mm; 1.17 × 1.17 × 1.2 mm voxels; duration = 4:56 min)	5,3	Sagittal	Ubuntu 14.04.5 LTS (Linux 3.13.0-153-generic x86_64)
FIDMAG	1.5T, GE Signa	3D T1: matrix size = 512 × 512, 180 contiguous axial slices, voxel resolution = 0.47 × 0.47 × 1mm, no slice gap, TE = 3.93ms, TR = 2000ms and inversion time (TI) = 710ms, flip angle = 15 degrees	6	Axial	Linux-centos6_x86_64
Groningen sample (DIP)	3T Philips	3D T1-weighted scan (170 slices; TR = 9ms; TE = 3.6ms; 256x231 matrix of 1×1×1 mm voxels)	5,3	Sagittal	SUSE Linux X86_64
Houston	subjects in 20000s: 1.5 T Philips Medical Systems Gyroscan Intera; subjects in 30000s: 3T Siemens Allegra	Subjects in the 20000s: Fast field echo sequence- repetition time (TR) = 24 ms, echo time (TE) = 4.99 ms, flip angle = 40°, slice thickness = 1 mm, matrix size = 256 × 256 and 150 slices. Subjects in 30000s: MPRAGE- repetition time (TR) = 1750 ms, echo time (TE) = 4.39 ms, flip angle = 8°, slice thickness = 1 mm, matrix size = 208 × 256 and 160 slices.	5,3	Subjects in 20000s: Sagittal; Subjects in 30000s: Transverse	Fedora 19
TiPs (Jena, Germany)	3T Siemens MAGNETOM Prisma_fit	MPRAGE sequence: TR 2300 ms, TE 3.03 ms, α 9°, 192 contiguous sagittal slices, in-plane field of view 256 mm, voxel resolution 1Å~1Å~1 mm; acquisition time 5:21 min	5,3	Sagittal	Linux
BRCDECC London	1.5T GE Signa HDx	ADNI-1 MPRAGE pulse sequence (details at http://adni.loni.ucla.edu/research/protocols/mri-protocols/)	5,3	Sagittal	Linux-centos4_x86_64
MODECT	3T (General Electric Signa HDxt, Milwaukee, WI, USA)	T1-weigthed dataset was acquired (flip angle=12°, repetition time=7.84 milliseconds, echo time=3.02 milliseconds; matrix 256x256, voxel size 0.94x0.94x1 mm; 180 slices).	5,3	Coronal	Linux
MPIP	1.5T GE and Siemens (the latter: only few cases)	#1: T1-weighted SPGR sagittal 3D volume. TR=1030 msec; TE=3.4 msec; 124 slices; matrix=256x256; FOV=23.0x23.0 cm2; voxel size=0.8975 mm x0.8975 mm x 1.2- 1.4 mm; flip angle=90°; birdcage resonator. #2: same scanner as #1, platform update Signa Excite, sagittal T1-weighted (spin echo sequence, TR=9.7 msec, TE=2.1 msec; FOV=25.0x25.0 cm2, voxel size=0.875 mm x0.875 mm x1.2 mm, 124- 132 slices, flip angle=90°. #3: Siemens 1.5 Tesla, Vario, 3D MPRAGE, TR=11.6 msec; TE=4.9 msec; FOV 23x23 cm2; matrix 512x512; 126 axial slices; voxel site 0.45 mm x 0.45 mm x 1.5 mm. (only N=2 subjects)	5,3	1.5 GE: sagittal. 1.5 Siemens: axial	Linux 2.6.37.1-1.2- desktop x86_64
Melbourne	3T GE Signa Excite	3D BRAVO sequence 140; TR=7900 ms; TE=3000 ms; flip angle=13º; FOV=256 mm; matrix=256 x 256	5,3	Axial	Linux Debian x86 64
Minnesota	3.0 Tesla Tim Trio scanner; Siemens Corp	A 5-minute structural scan was acquired using a T1-weighted, high-resolution, magnetization-prepared gradient-echo sequence: repetition	5,3	Coronal	Linux

		time, 2530 milliseconds; echo time, 3.65 milliseconds; inversion time, 1100 milliseconds; flip angle, 7°; field of view, 256 × 176 mm; voxel size, 1-mm isotropic; 224 slices; and generalized, autocalibrating, partially parallel acquisition acceleration factor, 2.			
Moral Dilemma	3T GE Signa Excite	3D BRAVO sequence: 140 contiguous slices; repetition time, 7900 ms; echo time, 3000 ms; flip angle, 13°; in a 25.6-cm field of view, with a 256 × 256 pixel matrix and a slice thickness of 1 mm (1 mm gap).	5,3	Axial	Linux Debian x86 64
NESDA	3T Phillips Achieva/Intera	3D gradient-echo T1-weighted sequence. TR=9 msec; TE=3.5 msec; flip angle 8º, FOV = 256 mm; matrix: 25x62x56; in plane voxel size = 1 mm × 1 mm x 1 mm; 170 slices.	5	Sagittal	SHARK HPC, Linux environment
QTIM	Bruker 4T Wholebody MRI	3D T1 weighted sequence. TR=1500 msec; TE=3.35 msec; flip angle=8°, 256 or 240 (coronal or sagittal) slices, FOV=240 mm, matrix 256x256x256 (or 256x256x240)	5,1	Coronal, then sagittal following software upgrade.	Linux-centos4_x86_64-stable-pub-v5.1.0
San Francisco UCSF	3T GE Discovery MR750	SPGR T1-weighted: TR=8.1 ms; TE=3.17 ms; TI=450 ms; flip angle=12°; 256x256 matrix; FOV=250x250 mm; 168 sagittal slices; slice thickness=1 mm; in-plane resolution=0.98x 0.98 mm	5,3	Sagittal	Linux-centos6_x86_64-stable-pub-v5.3.0.
SHIP	1.5T Siemens Avanto	3D T1-weighted (MP-RAGE/ axial plane); TR=1900 msec; TE=3.4 msec; Flip angle=15°; voxel size 1 mm x 1 mm x 1 mm	5.3 (cortical), 5.1 (subcortical)	Axial	Centos6_x86_64
SHIP/TREND	1.5T Siemens Avanto	3D T1-weighted (MP-RAGE/ axial plane); TR=1900 msec; TE=3.4 msec; Flip angle=15°; voxel size 1 mm x 1 mm x 1 mm	5.3 (cortical), 5.1 (subcortical)	Axial	Centos6_x86_64
San Raffaele Milano OSR	3T Philips Ingenia and 3T Philips Intera scanner	3D-MPRAGE sequence: TR 2500 ms, TE 4.6 ms, field of view FOV = 230 mm, matrix = 256 × 256, in-plane resolution 0.9 × 0.9 mm, yielding 220 transversal slices with a thickness of 0.8 mm.	5,3	axial	Linux Ubuntu 16.04
Singapore	Achieva 3T, Philips Medical Systems, Netherlands	Whole brain high resolution 3D MP-RAGE (magnetisation-prepared rapid acquisition with a gradient echo) volumetric scans (TR/TE/TI/flip angle 8.4/3.8/3000/8; matrix 256x204; FOV 240mm2) with axial orientation (reformatted to coronal)	5,3	Axial	Linux_Ubuntu12.04_6 4
SoCAT	3.0 T, Siemens Verio,Numaris/4,Syngo MR B17,Erlangen,Germany	3D T1 weighted MP-Rage/axial plane; TR=1900 msec; TE=3.4 msec; Flip angle=15°; Voxel size 1 mm x 1 mm x 1 mm	5,3	Axial	Ubuntu 18.04 LTS
Stanford FAA	3.0T GE Discovery MR750	Whole-brain T1-weighted images were collected using a spoiled gradient echo (SPGR) pulse sequence (186 sagittal slices; resolution = 0.9 mm isotropic; flip angle = 12°; repetition time [TR] = 6,240 ms; echo time [TE] = 2.34 ms)	5,3	Sagittal	Linux-centos6_x86_64
Stanford T1w Aggregate	1.5T GE Signa Excite	Whole-brain T1-weighted images were collected using a spoiled gradient echo (SPGR) pulse sequence (116 sagittal slices; through-plane resolution = 1.5 mm; in-plane resolution = 0.86 x 0.86 mm; flip angle = 15 degrees; repetition time [TR] = 8.3-10.1 ms; echo time [TE] = 1.7-3.0; inversion time [TI] = 300 ms; matrix = 256 x 192).	5,3	Sagittal	Centos6_x86_64, Linux-based HPC
TIGER	3T GE MR750	TR/TE/TI=8.2/3.2/600 ms; flip angle=12°; 156 axial slices; FOV=25.6 cm; matrix=256 mm x 256 mm, isotropic voxel=1 mm, total scan time: 3:40	6	Axial	Linux

Supplementary Table 2: ENIGMA MDD Instrument for diagnosing major depressive disorder and exclusion criteria by site

Cohort	Diagnosis measurment	Sample characteristics/Inclusion criteria	Exclusion criteria
AFFDIS	ICD-10/DSM-IV criteria	MDD subjects currently depressed and in day program or inpatient	All subjects exclusion criteria: current or history of neurological disorder or brain injury, current substance abuse or dependence (not including nicotine), pregnancy, MRI contraindications, inability to give consent. MDD specific: comorbid psychiatric diagnosis. Healthy control specific: current or history of psychiatric diagnosis.
Pharmo (AMC)	MINI Plus	48 subjects with lifetime diagnosis of either MDD and/or AD and 14 healthy controls. Patients were startified depending on exposure to SSRIs: early (before age 23) or late (after age 23) exposure to SSRI's, or no exposure at all (UN). 15 subjects were diagnosed with only MDD, 3 with only AD and 22 with both MDD and AD (8 subjects did not receive a diagnosis due to incomplete M.I.N.I. Plus assessment). According to the M.I.N.I. Plus, none of the HC subjects were ever diagnosed with MDD or AD	Less than three week medication-free interval before scanning, current psychotropic medication use, a history of chronic or neurological disorder, family history of sudden heart failure or epileptic attacks, pregnancy (tested via urine sampling prior to the assessment), breast feeding, alcohol dependence and contra-indications for an MRI scan (e.g., ferromagnetic fragments). Participants agreed to abstain from smoking, caffeine and alcohol use for 24 hours prior to the assessments.
Barcelona	DSM-IV-TR acc. to CIDI-interview and HAMD	Outpatients with MDD diagnosis (DSM-IV-TR), with a first episode, recurrent MDD or chronic MDD (TRD) age 18-65	The exclusion criteria for healthy participants were: lifetime psychiatric diagnoses, first-degree relatives with psychiatric diagnoses and clinically significant physical or neurological illnesses. Axis I comorbidity according to DSM-IV-TR criteria was an exclusion criteria for all participants.
Cardiff	Hamilton Depression Rating Scale (HDRS-17)	N= 40, MDD patients with a current moderate to severe depressive episode despite minimum three months of stable antidepressant treatment	Psychotic symptoms, current substance dependence, eating disorders, claustrophobia and other MRI contraindications, and ongoing non-pharmacological treatment.
CSAN (Adf)	MINI	Current MDD: Meets MINI criteria for depression; comorbid anxiety disorders are allowed; mood-congruent psychotic symptoms allowed.	Current MDD: a current DSM-5 diagnosis of substance use disorder, except nicotine; a psychotic disorder, except depression with mood-congruent psychotic features; new antidepressant medication during the month before study participation (two months for fluoxetine); change of the dose of psychotropic medications over the last month (antidepressant and antipsychotic medication) or the last two months (mood stabilizers and anticonvulsants).
Calgary	KSADS	First episode MDD and healthy controls (Dalhousie sample). Recurrent MDD and healthy controls, recruited via referral from clinicians in Calgary, Alberta and through advertisements in local clinics and at the University of Calgary (Calgary sample).	Dalhousie Sample: A history of neurological illness, medical illness, claustrophobia, >21 year of age, or the presence of a ferrous implant or pacemaker. University of Calgary: Left handed; history of seizures, epilepsy or other neurological or psychiatric diagnoses (specifically bipolar disorder, psychosis, pervasive developmental disorder, eating disorders, PTSD); pregnancy
DCHS	MINI	Women over the age of 18 years, who were between 20 and 28 weeks pregnant, who presented at either of the two recruitment clinics, and who had no intention of moving out of the area within the following year, and were able to give written consent	1) loss of consciousness longer than 30 minutes, 2) inability to speak English, 3) current/lifetime alcohol and/or substance dependence or abuse, 4) psychopathology other than PTSD and/or MDD, 5) traumatic brain injury, 6) standard MRI exclusion criteria
ETPB	HAMD,BDI, SHAPS,MADRS	Treatment resistant depression, at least one failed trial MADRS >20	Current or past diagnosis of Schizophrenia or any other psychotic disorder as defined in the DSM-IV. Subjects with a history of DSM-IV drug or alcohol dependency or abuse (except for nicotine or caffeine) within the preceding 3 months. Female subjects who are either pregnant or nursing. Serious, unstable illnesses including hepatic, renal, gastroenterologic, respiratory, cardiovascular (including ischemic heart disease), endocrinologic, neurologic,

			immunologic, or hematologic disease. Subjects with uncorrected hypothyroidism or hyperthyroidism. Subjects with one or more seizures without a clear and resolved etiology. Treatment with a reversible MAOI within 4 weeks prior to study phase I. Treatment with fluoxetine within 5 weeks prior to study phase I. Treatment with any other concomitant medication not allowed (Appendix A for Substudy 2; Appendix G for Substudy 4) 14 days prior to study phase I. No structured psychotherapy will be permitted during the study. Current NIMH employee/staff or their immediate family member. Additional Exclusion Criteria for substudy 2 (patients with MDD) Previous treatment with ketamine or hypersensitivity to amantadine. Additional Exclusion Criteria for Substudy 4 (patients with MDD or BD). Subjects who currently are using drugs (except for caffeine or nicotine), must not have used illicit substances in the 2 weeks prior to screen and must have a negative alcohol and drug urine test (except for prescribed benzodiazepines) urine test at screening. Presence of any medical illness likely to alter brain morphology and/or physiology (e.g., hypertension, diabetes) even if controlled by medications. Clinically significant abnormal laboratory tests. Presence of metallic (ferromagnetic) implants (e.g, heart pacemaker, aneurysm clip). Subjects who, in the investigator s judgment, pose a current serious suicidal or homicidal risk, or who have a MADRS item 10 score of >4.
EPISCA (Leiden)	ADIS	Inclusion criteria for the patient group were: having clinical depression as assessed by categorical and dimensional measures of DSM-IV depressive and anxiety disorders, no current and prior use of antidepressants, and being referred for CBT at an outpatient care unit. Inclusion criteria for the control group were: no current or past DSM-IV classifications, no clinical scores on validated mood and behavioral questionnaires, no history of traumatic experiences, and no current psychotherapeutic and/or psychopharmacological intervention of any kind.	Primary DSM-IV clinical diagnosis of ADHD, ODD, CD, pervasive developmental disorders, post-traumatic stress disorder, Tourette's syndrome, obsessive–compulsive disorder, bipolar disorder, and psychotic disorders; current substance abuse; history of neurological disorders or severe head injury; age < 12 or > 21 years; pregnancy; left-handedness; IQ score < 80 as measured by the Wechsler Intelligence Scale for Children (WISC) (Wechsler, 1991) or Adults (Wechsler, 1997); and general MRI contra- indications.
FIDMAG	DSM-IV-TR criteria	MDD patients within a current depressive episode (HDRS >= 17, only 1 patient was in remission), right-handed, age 18-65	Patients were excluded (i) if they were left-handed; (ii) if they were younger than 18 or older than 65 years; (iii) if they had a history of brain trauma or neurological disease; (iv) if they had shown alcohol/ substance abuse within 12 months prior to participation; and (v) if they had undergone electroconvulsive therapy in the previous 12 months.
Groningen sample (DIP)	MINI-SCAN	Outpatients with MDD diagnosis. Inclusion MDD: Outpatients treated in mental health care for depression, BDI-II>13 at screening, adults.	Exclusion MDD: Comorbid axis-I disorders other than anxiety disorders or past substance abuse, other psychotropic medication than stable use of SSRI/SNRI/TCA, established cardiovascular disease, active and concrete suicidal plans, inadequate language proficiency, cognitive impairments or neurological disease that interferes with task performance. Exclusion CTL: Same as MDD, lifetime history of MDD, BDI>8.
Houston	SCID interview	Outpatients	MDD subjects: age below 18; lifetime or current diagnosis of psychotic disorder, or bipolar I or II disorder; substance abuse/dependence in 6 months prior to study inclusion; current major medical problems. Control subjects: age below 18; current major medical problems; current psychiatric or neurologic disorder; history of psychiatric disorders in a first-degree relative;

			current major medical problems. Both groups: MRI contra-indications
TiPs (Jena, Germany)	SCID interview	Psychiatric inpatients and tinnitus patients with MDD or a disorder of the depressive spectrum (also adjustment disorders as pointed out in the data table); psychiatrically healthy controls were derived from community and tinnitus patients	MDD subjects: presence of axis-I disorders other than MDD or adjustment disorders. Control subjects: no Axis-I diagnosis, no medication use. Exclusion criteria for all subjects included history of neurological disease (e.g. tumour, head trauma, epilepsy) or untreated internal medical condtitions, intellectual and/or developmental disability. Only German native speakers were allowed to participate.
BRCDECC London	SCAN interview	Community based or outpatients, none were inpatients. MDD subjects: Less than two depressive episodes of at least moderate severity. Did not meet DSM-IV diagnostic criteria for recurrent major depressive disorder. Control group participants were clinically interviewed to ensure they had never experienced depressive symptoms. Exclusion criteria for all participants were for contraindications to MRI; other exclusion criteria were a diagnosis of neurological disorder, head injury leading to loss of consciousness or conditions known to affect brain structure or function (including alcohol or substance misuse), ascertained during clinical interview. Potential participants were also excluded if they or a first-degree relative had ever fulfilled criteria for mania, hypomania, schizophrenia or mood-incongruent psychosis.	Contraindications to MRI, diagnosis of neurological disorder, head injury leading to loss of consciousness or conditions known to affect brain structure or function (including alcohol or substance misuse), if they or a first-degree relative had ever fulfilled criteria for mania, hypomania, schizophrenia or mood-incongruent psychosis.
MODECT	MINI	Older adults, aged above 55, with severe depression admitted to be treated with ECT	Exclusion criteria were another major DSM-IV-TR diagnosis, such as schizophrenia, bipolar or schizoaffective disorder and a history of major neurological illness (including Parkinson's disease, stroke and dementia).
MPIP	M-CIDI/SCAN interview	M. A. R. S. sample: both first and recurrent episodes; RUD sample: only recurrent episodes with some patients scanned in remission	1. Munich Antidepressant Response Signature (MARS) study MDD subjects (clinical consensus diagnosis or M-CIDI (since 2008)): depressive syndromes secondary to any medical or neurological condition (e. g., intoxication, drug abuse, stroke), the presence of manic, hypomanic or mixed affective symptoms, lifetime diagnosis of alcohol dependence, illicit drug abuse or the presence of severe medical conditions (e.g., ischemic heart disease). Patients with bipolar depression were excluded for the current MR study. Control subjects: age > 65, MMSE<27, presence of severe somatic diseases or lifetime history of the following axis I disorders as assessed by the M-CIDI interview: alcohol dependence, drug abuse or dependence, possible psychotic disorder, mood disorder, anxiety disorder including OCD and PTSD, somatoform disorder, dissociative disorder NOS, and eating disorder 2. Recurrent unipolar depression (RUD) study: MDD subjects (SCAN interview): presence of manic episodes, mood incongruent psychotic symptoms, the presence of a lifetime diagnosis of intravenous drug abuse and depressive symptoms only secondary to alcohol or substance abuse or to medical illness or medication.Control subjects: presence of severe somatic diseases or life-time history of anxiety and affective disorders according to the Composite International Diagnostic-Screener (CIDI-S). All subjects: gross incidental MR findings such as territorial infarction, tumor, hydrocephalus, malformations and anatomical deviations (e.g. enlarged ventricles) that prevent appropriate image processing were additional exclusion criteria. 3. MR images of 9 additional controls acquired at the LMU,

			Munich, meeting equivalent criteria as the RUD control sample were included.
Melbourne	SCID interview	Youth depression sample: 15-25 years of age. Recruited as part of 2 large RCTs (incl. YoDA-C - Davey et al., 2014; Trials) and scanned prior to treatment randomisation. 60 patients unmedicated (YoDA-C).	MDD subjects: lifetime or current SCID-I diagnosis of psychotic disorder, or bipolar I or II disorder. Control subjects: any SCID-I diagnosis or medication use. Both groups: Acute or unstable medical disorder; general MRI contraindications
Minnesota	Schedule for Affective Disorders and Schizophrenia for School-Age Children–Present and Lifetime Version and the Children's Depression Rating Scale–Revised (CDRS-R).	Adolescents with MDD and HCs aged 12 to 19 years were recruited to participate through community postings and referrals from local mental health services. Adolescents with MDD were eligible if they had a primary diagnosis of MDD and had not received any psychotropic medication treatment for the past 2 months. Healthy adolescents were eligible if they had no current or past psychiatric diagnoses and were frequency matched to the MDD group on age and sex	Exclusion criteria for both groups included the presence of a neurologic or other chronic medical condition, mental retardation, pervasive developmental disorder, substance use disorder, bipolar disorder, or schizophrenia
Moral Dilemma	SCID interview	Youth depression sample: 15-25 years of age; recruited from outpatient service. Controls recruited from general community.	MDD subjects: lifetime or current SCID-I diagnosis of psychotic disorder, or bipolar I or II disorder; current antidepressant medication use. Control subjects: any SCID-I diagnosis or medication use. Both groups: Acute or unstable medical disorder; general MRI contraindications
NESDA	CIDI interview	DSM-4 based diagnosis of MDD (6 month recency), using CIDI interview. 93 (60%) MDD patients have a comorbid ANX diagnosis. Age range 18-65	
QTIM	CIDI interview	Retrospective questionnaire about depression episodes combined with an MRI study. The best described MDD episode is defined as the worst one (according to individuals). We have up to 5 supplementary episodes (briefly) described. Sample composed of twins and relatives. Population-based sample	MDD subjects: presence of axis-I disorders other than MDD and anxiety disorders Control subjects: antidepressant use, psychiatric disorders All subjects: relatedness between subjects, left handedness, history of neurological or other severe medical illness, head injury or current or past diagnosis of substance abuse, use of cognition affecting medication and general MRI contraindications
San Francisco UCSF	KSADS (semi-structured interview based on DSM) for MDD, DISC/DPS for HCL	Outpatient/community-based sample with DSM diagnosis, mostly antidepressant-naive and approximately 60% of MDD have comorbid anxiety disorders	Exclusion criteria for all participants included: 1) use of pharmacotherapeutics for treating psychiatric conditions within the past 6 months, 2) misuse of drugs within two months prior to MRI scanning; 3) two or more alcoholic drinks per week within the previous month (as assessed by the Customary Drinking and Drug Use Record; CDDR) (Brown et al, 1998); 4) a full scale IQ score of less than 75 (as assessed by the Wechsler Abbreviated Scale of Intelligence; WASI) (Wechsler, 1999); 5) contraindications for MRI including ferromagnetic implants and claustrophobia; 6) pregnancy or the possibility of pregnancy; 7) left-handedness; 8) prepubertal status (as assessed as Tanner stages of 1 or 2) (Tanner, 1962); 9) inability to understand and comply with procedures; 10) neurological disorder (including meningitis, migraine, or HIV); 11) head trauma; 12) learning disability; 13) serious health problems; and 14) complicated or premature birth (i.e., birth before 33 weeks of gestation). The MDD group was subject to the additional exclusion criterion of a primary psychiatric diagnosis other than MDD. The HCL group was subject to the additional exclusion criteria of: 1) history of mood or psychotic disorders in a first- or second-degree relative (as assessed by the Family Interview for Genetics; FIGS) (Maxwell, 1992); and 2) current or lifetime DSM-IV-TR Axis I psychiatric disorder.
SHIP	M-CIDI interview	Population based longitudinal cohort study	MDD subjects: presence of axis-I disorders other than MDD, anxiety disorders, conversion, somatization and eating disorder. Control

			subjects: no lifetime diagnosis of depression, no antidepressiva, and severity index=0 All subjects: We removed subjects with medical conditions (e.g. a history of cerebral tumor, stroke, Parkinson's diseases, multiple sclerosis, epilepsy, hydrocephalus, enlarged ventricles, pathological lesions) or due to technical reasons (e.g. severe movement artifacts or inhomogeneity of the magnetic field).
SHIP/TREND	M-CIDI interview	Population based longitudinal cohort study	MDD subjects: no special exclusion criteria Control subjects: no lifetime diagnosis of depression, no antidepressiva, and severity index=0 All subjects: We removed subjects with due to medical conditions (e.g. a history of cerebral tumor, stroke, Parkinson's diseases, multiple sclerosis, epilepsy, hydrocephalus, enlarged ventricles, pathological lesions) or due to technical reasons (e.g. severe movement artifacts or inhomogeneity of the magnetic field).
San Raffaele Milano OSR	SCID interview	adult MDD depressed inpatients	Other diagnoses on Axis I; pregnancy, history of epilepsy, major medical and neurological disorders; absence of a history of drug or alcohol dependency or abuse within the last six months. inflammation-related symptoms, including fever and infectious or inflammatory disease; uncontrolled systemic disease; uncontrolled metabolic disease or other significant uncontrolled somatic disorder known to affect mood; somatic medications known to affect mood or the immune system, such as corticosteroids, non-steroid anti-inflammatory drugs and statins.
Singapore	SCID interview	Inclusion: 1) DSM IV dx of MDD (Patients) 2) Age: 21-65 3) English speaking 4) Provision of informed written consent	Exclusion criteria 1) History of significant head injury 2)Neurological diseases such as epilepsy, cerebrovascular accident 3) Impaired thyroid function 4) Steroid use 5) DSM IV alcohol or substance use or dependence 6) Contraindications to MRI (e.g. pacemaker, orbital foreign body, recent surgery/procedure with metallic devices/implants deployed) using standard MRI Request Form from NNI 7)Pregnant women 8) Claustrophobia
SoCAT	SCID interview	Inclusion criteria: DSM IV dx for mdd patients Age: 18-65 right-handed currently depressed or remitted; Control subjects: any history of psychiatric disorder	Exclusion criteria 1) History of significant head injury 2)Neurological diseases such as epilepsy, cerebrovascular accident 3)Other diagnoses on Axis I disorders4)
Stanford FAA	SCID interview	Community-based DSM-diagnosed sample	MDD subjects: presence of axis-I disorders other than MDD, anxiety and eating disorders . Control subjects: control individuals did not meet diagnostic criteria for any current psychiatric. Both groups: alcohol / substance abuse or dependence within six months prior to MRI scanning, history of head trauma with loss of consciousness > 5 min, aneurysm, or any neurological or metabolic disorders that require ongoing medication or that may affect the central nervous system (including thyroid disease, diabetes, epilepsy or other seizures, or multiple sclerosis), MRI contraindications, or bad MRI data (e.g., extreme movement).
Stanford T1w Aggregate	SCID interview	Community-based DSM-diagnosed sample	MDD subjects: presence of axis-I disorders other than MDD, anxiety and eating disorders . Control subjects: control individuals did not meet diagnostic criteria for any current psychiatric. Both groups: alcohol / substance abuse or dependence within six months prior to MRI scanning, history of head trauma with loss of consciousness > 5 min, aneurysm, or any neurological or metabolic disorders that require ongoing medication or that may affect the central nervous system (including thyroid disease, diabetes, epilepsy or other seizures, or multiple sclerosis), MRI contraindications, or bad MRI data (e.g., extreme movement).
TIGER	KSADS	Community-based DSM-diagnosed sample	All subjects: Exclusion criteria were premenarchal status (for females), history of concussion within the past 6 weeks or history of

			any lifetime concussion with loss of consciousness, contraindications to MRI scanning (e.g. braces, metal implants, or claustrophobia), serious neurological or intellectual disorders that could interfere with the participant's ability to complete study components. MDD subjects: meeting lifetime or current DSM-IV criteria for any Bipolar Disorder, Psychosis, or Alcohol Dependence, or DSM-5 criteria for Moderate Substance Use Disorder with substance-specific threshold for withdrawal. CTL subjects: any current or past DSM-IV Axis I Disorder and first-degree relative with confirmed or suspected history of depression, mania, psychosis, or substance dependence.

Supplementary Table 3: List of hyperparameters of trained algorithms. Optimal hyperparameters were found by grid search.

Classification algorithm	Feature Selection	Hyperparameters	Inner CV
SVM Linear	None	C = $[10^{-4}, 10^{-3}, \dots, 10^{4}]$	10 fold
SVM Linear	PCA	C = $[10^{-4}, 10^{-3}, \dots, 10^{4}]$	10 fold
		% components = [10%,20%,100%]	
SVM Linear	Ttest (pvalue<0.5)	C = $[10^{-4}, 10^{-3}, \dots, 10^{4}]$	10 fold
SVM rbf	None	C = $[10^{-4}, 10^{-3}, \dots, 10^{4}]$	10 fold
LASSO	None	C = $[10^{-4}, 10^{-3}, \dots, 10^{4}]$	10 fold
Ridge	None	C = $[10^{-4}, 10^{-3}, \dots, 10^{4}]$	10 fold
Elastic Net	None	C = $[10^{-4}, 10^{-3}, \dots, 10^{4}]$	10 fold
		L1_ratio = $[0.1, 0.2, \dots, 1]$	

Supplementary Table 4: Balanced accuracy measured on the entire dataset

Splitting by Age/Sex								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
Elastic Net	0.609	0.523	0.584	0.500	0.569	0.517	0.593	0.520
LASSO	0.612	0.524	0.583	0.499	0.578	0.516	0.596	0.518
Ridge	0.610	0.523	0.585	0.498	0.573	0.515	0.594	0.520
SVM + PCA	0.638	0.529	0.601	0.513	0.575	0.518	0.622	0.513
SVM + ttest	0.627	0.515	0.581	0.515	0.567	0.526	0.619	0.521
SVM linear	0.616	0.524	0.577	0.504	0.572	0.518	0.602	0.524
SVM rbf	0.639	0.525	0.600	0.515	0.578	0.510	0.619	0.513
Splitting by Site								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
Elastic Net	0.513	0.514	0.498	0.489	0.503	0.514	0.507	0.514
LASSO	0.513	0.517	0.491	0.489	0.508	0.513	0.507	0.512
Ridge	0.514	0.514	0.497	0.490	0.505	0.509	0.507	0.514
SVM + PCA	0.527	0.520	0.502	0.512	0.504	0.524	0.520	0.520
SVM + ttest	0.502	0.512	0.487	0.499	0.507	0.508	0.510	0.527
SVM linear	0.512	0.519	0.498	0.495	0.499	0.506	0.506	0.521
SVM rbf	0.513	0.515	0.493	0.499	0.493	0.513	0.503	0.519

Supplementary Table 5: Area Under the Curve (AUC) measured on the entire dataset

Splitting by Age/Sex								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
Elastic Net	0.648539	0.523596	0.615843	0.499245	0.595426	0.517599	0.633831	0.523637
LASSO	0.650035	0.524172	0.617033	0.500462	0.59732	0.520961	0.634274	0.523566
Ridge	0.648338	0.523602	0.615255	0.500242	0.596384	0.517314	0.633636	0.523546
SVM	0.680602	0.54138	0.635524	0.526516	0.60116	0.516012	0.663276	0.51936
PCA + SVM ttest	0.666525	0.526525	0.61945	0.529764	0.588676	0.528114	0.655524	0.52274
SVM linear	0.653819	0.484991	0.613508	0.499734	0.597671	0.508618	0.63484	0.512361
SVM rbf	0.676804	0.536093	0.635927	0.529764	0.610408	0.512816	0.663968	0.527769
Splitting by Site								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
Elastic Net	0.523872	0.522363	0.501988	0.48749	0.504883	0.519347	0.523983	0.522332
LASSO	0.524027	0.523467	0.492272	0.486919	0.507039	0.520232	0.523876	0.52224
Ridge	0.523911	0.522412	0.496877	0.486571	0.504747	0.517977	0.524009	0.522273
SVM	0.524491	0.525502	0.493816	0.510443	0.498252	0.527865	0.525457	0.525901
PCA + SVM ttest	0.509798	0.520948	0.483935	0.50375	0.508691	0.513711	0.509827	0.528292
SVM linear	0.522964	0.508571	0.498851	0.506523	0.505694	0.506809	0.517339	0.524529
SVM rbf	0.522126	0.521114	0.492071	0.50375	0.496003	0.512841	0.51567	0.520161

Supplementary Table 6: Sensitivity measured on the entire dataset

Splitting by Age/Sex								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
Elastic Net	0.58099	0.514061	0.544332	0.488109	0.533295	0.508275	0.571971	0.509539
LASSO	0.587113	0.514919	0.545587	0.479235	0.532011	0.505767	0.575458	0.505136
Ridge	0.581884	0.512739	0.547864	0.488856	0.534633	0.510037	0.57285	0.509124
SVM	0.561453	0.449954	0.531587	0.392439	0.473415	0.489011	0.554732	0.480713
PCA + SVM ttest	0.57149	0.50289	0,470528	0,451512	0,369522	0,469517	0.557393	0.493869
SVM linear	0.563102	0.508836	0.475565	0.477826	0.489956	0.516014	0.574156	0.525332
SVM rbf	0.575977	0.413184	0.515986	0.451512	0.499034	0.487073	0.555247	0.475907
Splitting by Site								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
Elastic Net	0.508823	0.505946	0.49201	0.483858	0.442774	0.509582	0.523821	0.509649
LASSO	0.507419	0.508388	0.497816	0.483757	0.436127	0.512557	0.525108	0.507443
Ridge	0.508823	0.50562	0.495052	0.485361	0.440439	0.507268	0.524251	0.509898
SVM	0.441462	0.490569	0.458459	0.453159	0.359674	0.529208	0.445305	0.515571
PCA + SVM ttest	0.463461	0.516686	0.380577	0.425697	0.278024	0.483774	0.456027	0.501919
SVM linear	0.477417	0.50553	0.43298	0.478957	0.395238	0.508418	0.49214	0.52187
SVM rbf	0.454889	0.466977	0.402547	0.425697	0.378808	0.512755	0.451128	0.508139

Supplementary Table 7: Specificity measured on the entire dataset

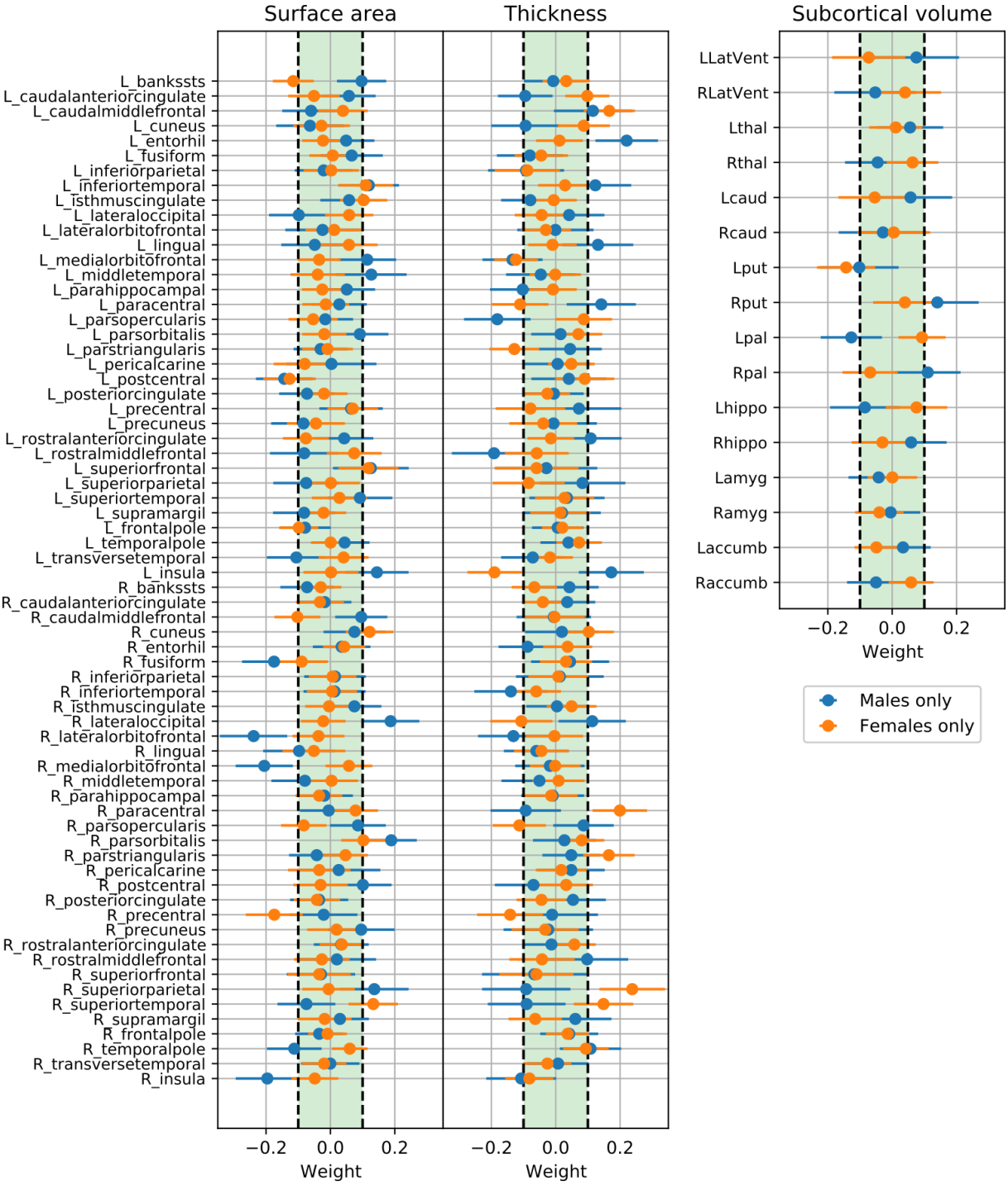
Splitting by Age/Sex								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
Elastic Net	0,637355	0,532388	0,622725	0,512761	0,60403	0,525794	0,614494	0,53087
LASSO	0,636702	0,533024	0,619805	0,518401	0,62387	0,527035	0,615774	0,529925
Ridge	0,637367	0,532727	0,62172	0,506681	0,61189	0,519568	0,614151	0,530534
SVM	0,713958	0,60902	0,670171	0,634246	0,67668	0,547546	0,689005	0,545735
PCA + SVM ttest	0,682816	0,52688	0,691055	0,577748	0,76375	0,582134	0,680718	0,547983
SVM linear	0,669546	0,538587	0,679278	0,530123	0,65501	0,520541	0,629424	0,521685
SVM rbf	0,70293	0,636271	0,683979	0,577748	0,65626	0,532894	0,682409	0,550219
Splitting by Site								
	Cortical + Subcortical		Cortical Thickness		Cortical Surface area		Subcortical Volume	
	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat	No ComBat	With ComBat
Elastic Net	0,517111	0,522542	0,504184	0,494112	0,564	0,517669	0,489776	0,517682
LASSO	0,517831	0,525393	0,483815	0,493526	0,58012	0,514426	0,488071	0,517208
Ridge	0,518686	0,522408	0,498261	0,493809	0,5705	0,511332	0,489386	0,518182
SVM	0,611771	0,549877	0,546142	0,571302	0,64857	0,519776	0,594934	0,525237
PCA + SVM ttest	0,540771	0,50779	0,592581	0,571477	0,7355	0,533098	0,564258	0,551745
SVM linear	0,546405	0,532711	0,563156	0,510389	0,60297	0,504562	0,519196	0,519815
SVM rbf	0,57041	0,562125	0,583274	0,571477	0,60703	0,51332	0,555772	0,529274

Supplementary Table 8: Performance of different harmonization options for every classification model

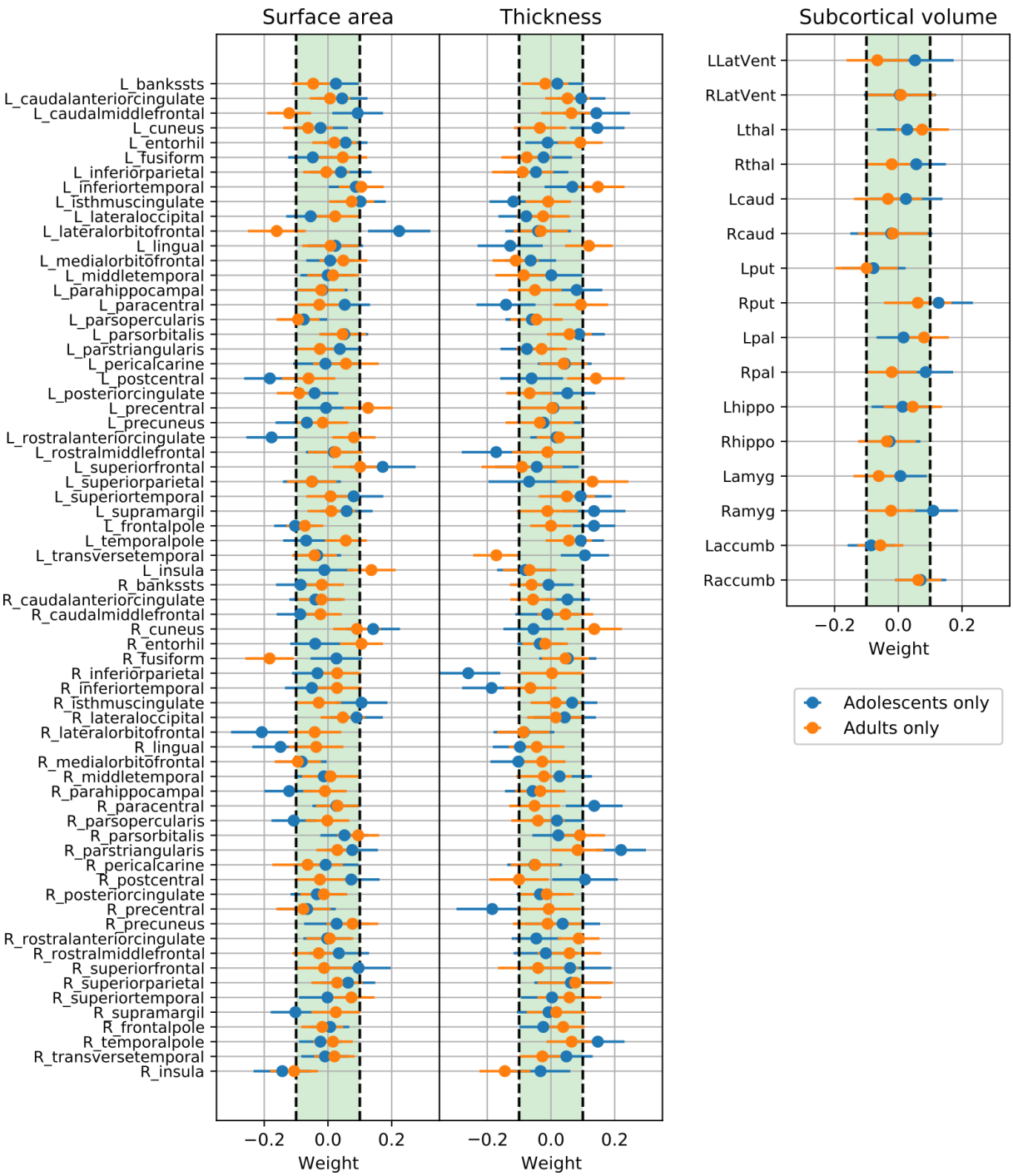
	Splitting by Age/Sex			Splitting by Site		
	ComBat	ComBat-GAM	CovBat	ComBat	ComBat-GAM	CovBat
Elastic Net	0.523	0.522	0.517	0.514	0.515	0.514
LASSO	0.524	0.523	0.517	0.517	0.513	0.514
Ridge	0.523	0.519	0.518	0.514	0.516	0.514
SVM PCA	0.529	0.521	0.523	0.520	0.520	0.528
SVM + ttest	0.515	0.503	0.513	0.512	0.505	0.512
SVM linear	0.524	0.526	0.521	0.519	0.520	0.518
SVM rbf	0.525	0.522	0.522	0.515	0.519	0.512

Supplementary Table 9: Balanced accuracy of SVM model trained and validated with Splitting by Site strategy. More extreme values of balanced accuracy are obtained for cohorts containing no healthy subjects. Note that ComBat brings these values closer to average across all cohorts.

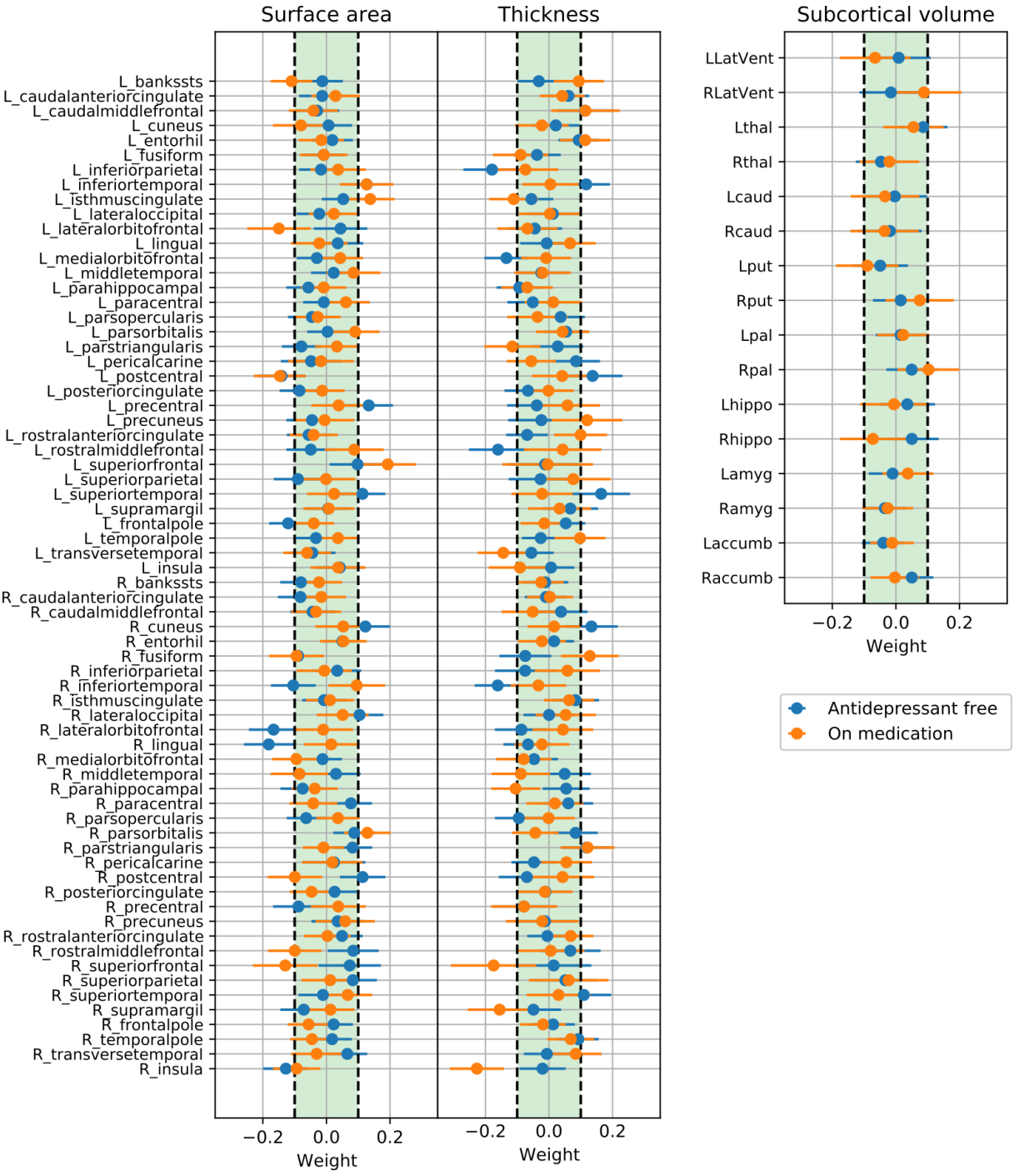
Name of site	No ComBat (Bacc)	With ComBat (Bacc)	Ratio MDD/HC
SHIP T0	0.503017	0.50513	0.3373232
SHIP S2	0.474821	0.516075	0.3069977
StanfT1wAggr	0.535866	0.476998	0.9491525
Minnesota	0.532143	0.521429	1.75
CSAN	0.489966	0.531973	1.2244898
Jena	0.498268	0.549567	0.3896104
Calgary	0.522727	0.553147	1.0576923
Barc	0.489415	0.467742	1.9375
DCHS	0.514117	0.607013	0.295082
AFFDIS	0.496377	0.455534	0.7173913
Moraldilemma	0.67663	0.586051	0.5217391
FIDMAG	0.487815	0.535714	1.0294118
MPIP	0.532092	0.533309	1.5971564
ETPB	0.561086	0.529412	1.3076923
TIGER	0.569573	0.517625	4.4545455
AMC	0.313726	0.431373	0
Episca(Leiden)	0.563158	0.513158	0.6333333
SanRaffaele	0.911111	0.644444	0
MODECT	0.785714	0.309524	0
Gron	0.519048	0.585714	0.952381
CARDIFF	0.85	0.5	0
Singapore	0.434659	0.400568	1.375
StanfFAA	0.468254	0.373016	0.7777778
QTIM	0.506663	0.509183	0.3591549
Houst	0.491057	0.477512	0.5591398
Melb	0.568216	0.546551	1.4019608
NESDA	0.52952	0.533417	2.3692308
Socat_dep	0.527722	0.547468	0.79
UCSF	0.522197	0.506061	0.8522727
LOND	0.531005	0.609527	1.1311475
ALL SITES	0.513	0.515	0.74



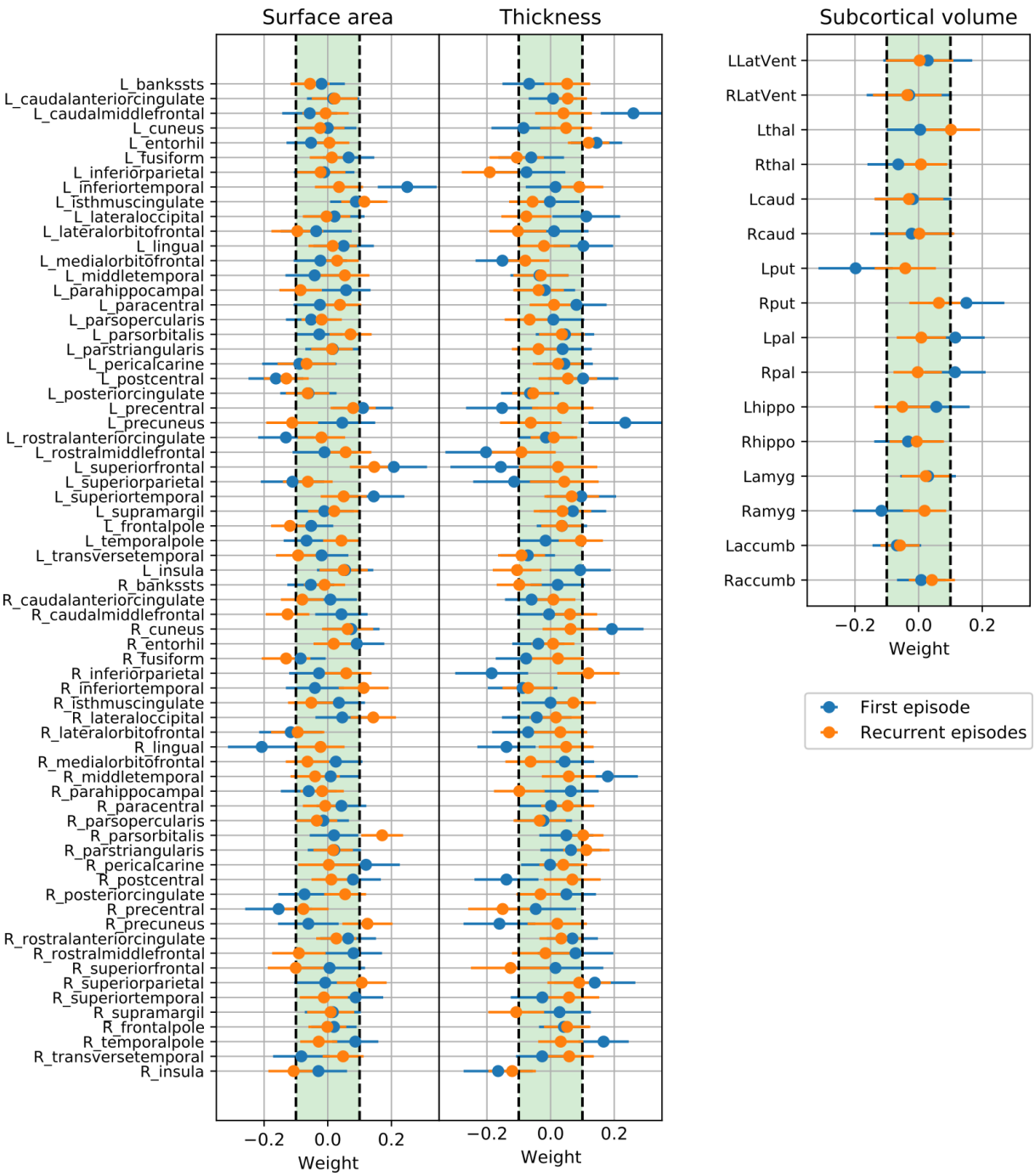
Supplementary Figure 1: Weights of SVM with linear kernel applied on stratified data by sex (no feature selection, with ComBat). The horizontal bars indicate the 95% confidence interval calculated using percentile method via bootstrapping.



Supplementary Figure 2: Weights of SVM with linear kernel applied on stratified data by age of onset (no feature selection ,with ComBat). The horizontal bars indicate the 95% confidence interval calculated using percentile method via bootstrapping.



Supplementary Figure 3: Weights of SVM with linear kernel applied on stratified data by use of antidepressant medication (no feature selection, with ComBat). The horizontal bars indicate the 95% confidence interval calculated using percentile method via bootstrapping.



Supplementary Figure 4: Weights of SVM with linear kernel applied on stratified data by number of episodes (no feature selection, with ComBat). The horizontal bars indicate the 95% confidence interval calculated using percentile method via bootstrapping.

CV splitting strategies

We wrote our own scripts in python to perform both splitting strategies, which were different in Splitting by Site and Splitting by Age/Sex. In Splitting by Site, we balanced the number of subjects across folds as much as possible .The code is publicly available (https://github.com/vlbl/Splitting-by-Site). Conversely, In Splitting by Age/Sex, we balanced the number of subjects across folds. It was achieved by assigning every subject to the fold, which leads to more even age/sex distribution (by comparing the mean of all folds to the mean of the fold when subject is added) across the folds. This process is repeated for every site separately. Thus, each fold contains almost equal number of subjects per site.

Harmonization methods

We harmonized individual cortical and subcortical features by implementing the well-established statistical harmonization algorithm, ComBat [200]. Its purpose was to adjust Location (mean) and Scale (variation) (L/S) of all features of the data collected from different cohorts by preserving the influence of biologically-significant factors of interest in the features. Additionally, it is assumed that the site effect is not independent across cortical and subcortical features and it uses empirical Bayes for site effect estimation. Subsequently, the cortical and subcortical features would be standardized, while the site effect would be removed. ComBat assumes that the data $Y_{i,j,k}$ for ROI *k*, site *i* and subject *j* can be represented by the following model:

$$Y_{ijk} = \alpha_k + X_{ij}\beta_k + \gamma_{ik} + \delta_{ik}\varepsilon_{ijk} \qquad (1)$$

Where α_k is an overall ROI value, X is a design matrix where X_{ij}j is a vector containing site affiliation and controlled covariates of participant j in site i.. In our case these are age, sex and ICV. β_k is the vector of regression coefficients corresponding to X_{ij}, γ_{ik} and δ_{ik} correspond to additive and multiplicative site effect and ε_{ijk} is an error term assumed to follow normal distribution with mean zero and variance σ_k^2 .After parameter estimation in the model above, the standardized data $Y^*{}_{ijk}$ can be calculated as follows:

$$Y^*{}_{ijk} = \frac{Y_{ijk} - \hat{\alpha}_k - X_{ij}\hat{\beta}_k - \hat{\gamma}_{ik}}{\hat{\delta}_{ik}} + \hat{\alpha}_k + X_{ij}\hat{\beta}_k \qquad (2)$$

where $\hat{\alpha}_k$, $\hat{\beta}_k$, $\hat{\gamma}_{ik}$ and $\hat{\delta}_k$ are estimated ComBat parameters. Additionally, it is assumed that the site effect is not independent across cortical and subcortical features.

All parameter estimations, which includes estimates of $\hat{\alpha}_k$, $\hat{\beta}_k$, $\hat{\gamma}_{ik}$ and $\hat{\delta}_k$, should be computed only on the training set, i.e. 9 CV folds, to avoid non-independence of the training and test data, also known as data leakage. After parameter estimations and training of the ML algorithm were complete, the calculated parameters were used to adjust the test data and the performance of the trained classification algorithm measured on the test set represented by the remaining CV fold. These parameters were directly used for adjusting data from unseen subjects from the test set only if these subjects belong to the same cohorts as in the training set. This scenario corresponds to Splitting by Age/Sex strategy as every CV fold contains subjects from all cohorts.

In Splitting by Site strategy, subjects from one cohort are included only in one CV fold, thus the direct usage of estimated ComBat parameters on the test set is imprudent. Here we adapted the approach of a reference batch adjustment [241], which constitutes fixing a reference sites, while other sites are adjusted to the mean and variance of the reference site according to the following equation:

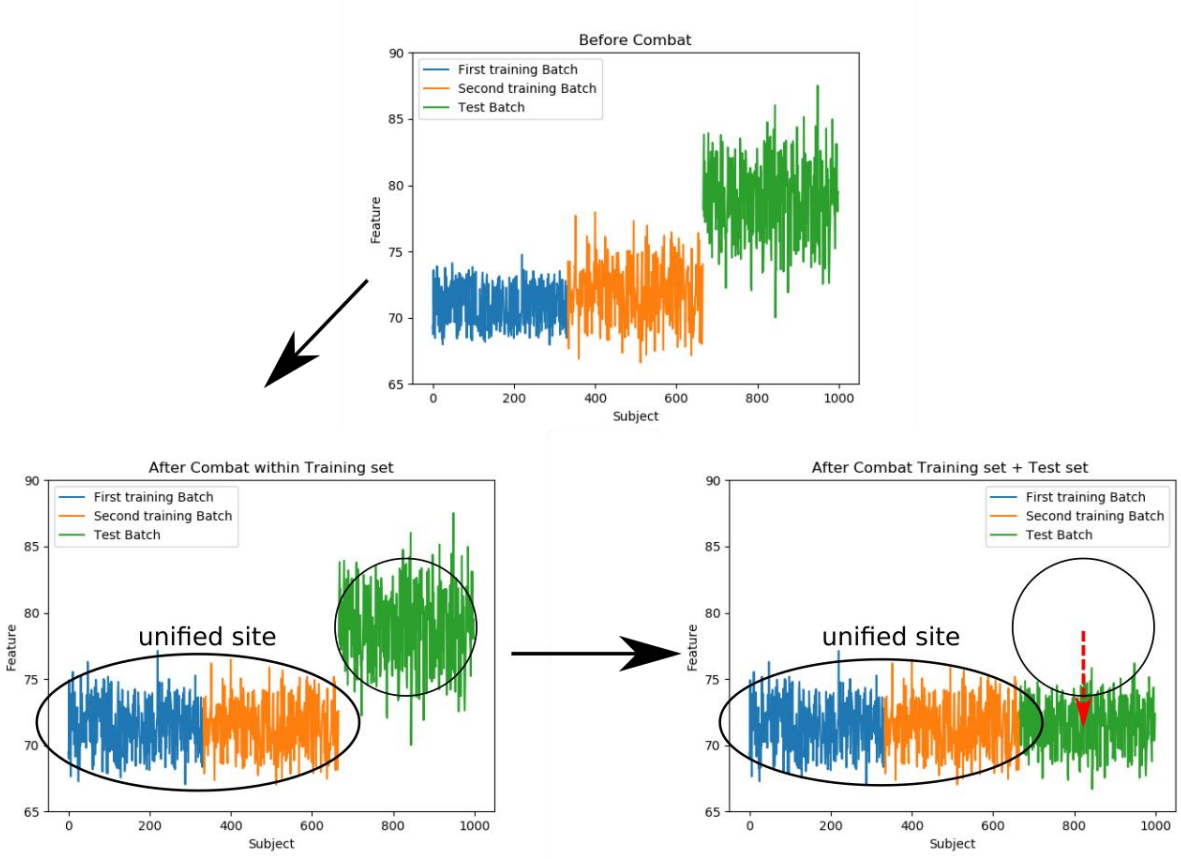
$$Y^*_{ijkr} = \frac{Y_{ijk} - \hat{\alpha}_{rk} - X_{ij}\hat{\beta}_{kr} - \hat{\gamma}_{ikr}}{\hat{\delta}_{ikr}} + \hat{\alpha}_{kr} + X_{ij}\hat{\beta}_{kr} \quad (3)$$

where α_{kr}, β_{kr} correspond to coefficients estimated on the reference site *r*. Additionally, γ_{ikr}, δ_{ikr} represent additive and multiplicative differences between site i and r. In our case, the test set was adjusted to a unified batch made by integrating all cohorts from the training set and adjusting to common mean and variance by the ComBat, which allowed to harmonize unseen cohorts without data leakage from the training set to the test set (Supplementary Figure 5).

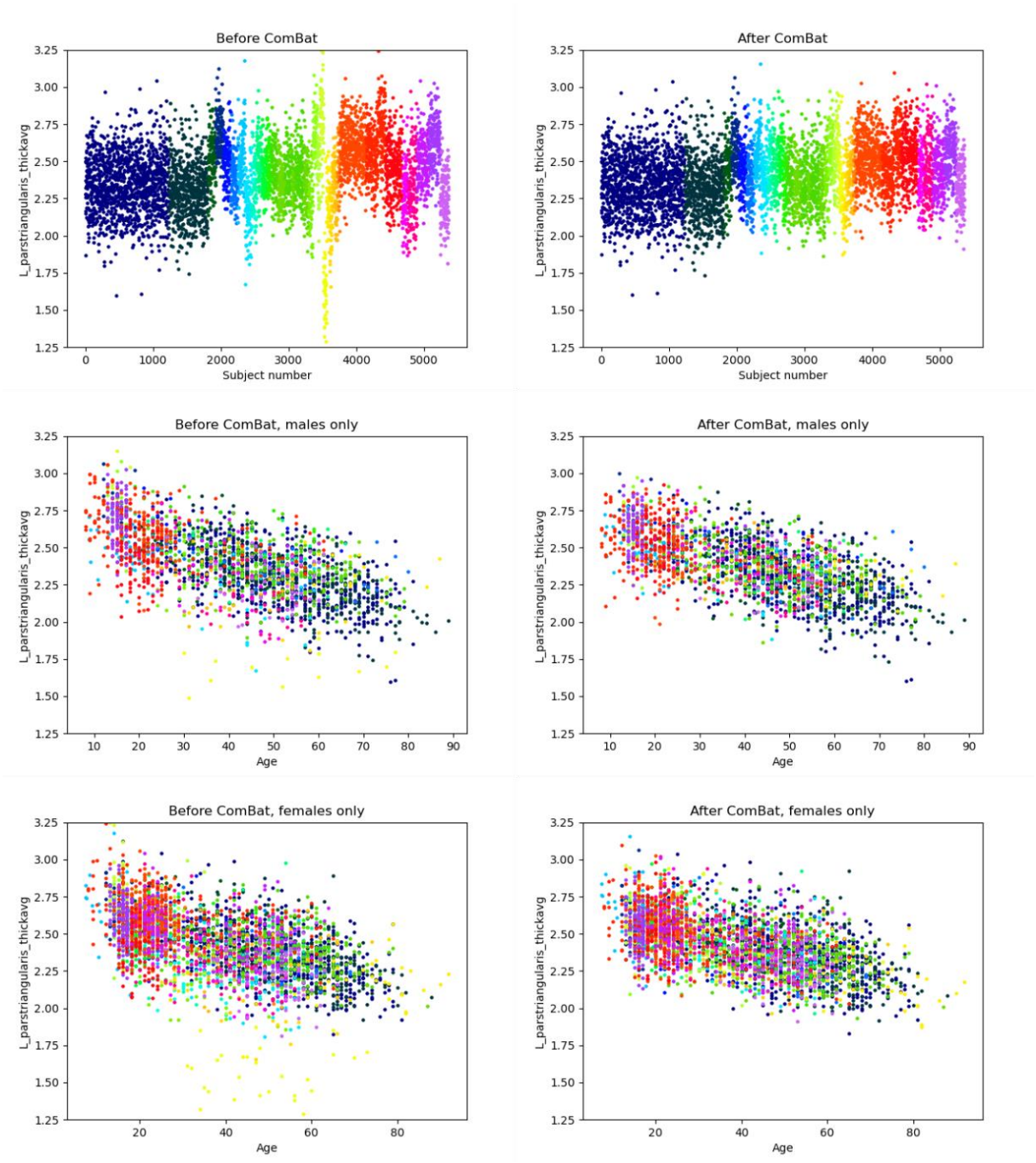
This framework was additionally extended to include non-linear preservation of the covariates by substituting $X_{ij}\beta_k$ with a Generalized Additive Model (ComBat-GAM) [205], allowing nonlinear age trends to be preserved during the harmonization step. Furthermore, we considered a CovBat model, which assumes an additional covariance site effect alongside with mean and variance corrections [206].We tested ComBat's harmonization ability to remove the site effect from the full data by training cortical and subcortical features via SVM with a linear kernel to predict the site. This was mostly tested in cases where the number of sites was below 10 [203],

[205], [242], so it is relevant in the context of our current analysis. We used the Splitting by Age/Sex strategy, since to predict site information from the test folds, this information should be presented in the training folds - and this would not be possible in the Splitting by Site strategy. The balanced accuracy was 0.854 before applying ComBat without correction for age, sex and ICV. Such a high performance indicated a strong site effect presented in the data, which may interfere with the main MDD vs HC classification task. The classification balanced accuracy dropped substantially to 0.031 after applying ComBat, indicating the significant removal of site-related information in cortical and subcortical features. Such a low accuracy comes from the fact that with ComBat site-related information seemed to be removed from the data, resulting in SVM always predicting SHIP_T0 – the biggest cohort. By assessing the confusion matrices we could evidence that - without the harmonization step - the classification algorithm was able to predict site affiliation of the subject successfully, except of sites coming from the same research group e.g., MPIP (two cohorts) and in case of SHIP_T0-SHIP_S2 and Melbourne-MoralDilemma pairs. As an illustration, an example of a feature being harmonized via ComBat may be seen in Supplementary Figure 6.

To further investigate whether differences across sites were mainly driven by irregular age and sex distributions, we repeated the classification task by also regressing out age and sex from the features. This resulted in 0.816 and 0.031 balanced accuracies for predicting site without and with ComBat harmonization respectively. By comparing the results with and without this residualization step, we could infer that the classification performance in differentiating sites only minorly came from age and sex distribution as it remained very similar to the previous classification results. To see if the site effect was due to the differences in scanners and scan acquisition protocols between cohorts, we trained SVM to predict scanner type from cortical and subcortical features. The resulting accuracy was 0.875, even higher than only site prediction. This hints to the site effect being primarily caused by differences in acquisition equipment across sites.



Supplementary Figure 5: Test set adjustment to the unified site. After ComBat is applied on the training set, all training sites are adjusted so that their residuals (after fitting covariates) have the same mean and variance, which we unify to build a unified site used for the classification training. After the training is complete, test set is harmonized to the fixed unified site allowing trained model to be evaluated on the test set.



Supplementary Figure 6**:** An example of site effect removal by ComBat for left pars opercularis thickness. Color corresponds to the site affiliation. While the differences between sites are reduced, remaining differences correspond to age- and sex-related differences between cohorts (middle and bottom).

Chapter 4 Discriminating major depressive disorder on cortical surface-based features: A deep learning approach

Machine learning classification of major depressive disorder and healthy individuals using vertex-wise cortical features

Vladimir Belov[1], Tracy Erwin-Grabner[1], Ling-Li Zeng[2], Alec J. Jamieson[3], Aleix Solanes[4], Aleks Stolicyn[5], Ali Saffet Gonul[6], Alyssa R. Amod[7], Amar Ojha[8], Andre Aleman[9], Annemiek Dols[10], Anouk Schrantee[11], Aslihan Uyar-Demir[6], Baptiste Couvy-Duchesne[12], Ben J Harrison[3], Benson Mwangi[13,14], Bianca Besteher[15], Bonnie Klimes-Dougan[16], Brenda W. J. H. Penninx[11], Bryon A. Mueller[17], Carlos Zarate[18], Christopher G. Davey[3], Colm G. Connolly[19], Dan J. Stein[20], David E. J. Linden[21,22,23,24], David M. A. Mehler[21,22,25], Dominik Grotegerd[26], Edith Pomarol-Clotet[27], Eduard Vieta[28], Elena Pozzi[29,30], Elena Rodríguez-Cano[27], Elisa Melloni[31], Emmanuelle Corruble[32,33], Francesco Benedetti[31], Frank P. MacMaster[34], Frederike Stein[35], Hans J. Grabe[36], Heather C Whalley[37], Henry Völzke[38], Ian H. Gotlib[39], Igor Nenadić[35], Jair C. Soares[14], Jan Ernsting[25], Jennifer W. Evans[40], Joaquim Radua[41], Kang Sim[42,43,44], Katharina Brosch[35], Katharina Wittfeld[36,45], Kathryn Cullen[17], Laura K.M. Han[30,46], Lukas Fisch[25], Mardien L. Oudega[47,48], Margaret J. Wright[49,50], Maria J Portella[51], Martin Walter[15,52], Matthew D. Sacchet[53], Meng Li[15], Mon-Ju Wu[14], Neda Jahanshad[54], Nils Winter[25], Nynke A. Groenewold[55], Paola Fuentes-Claramonte[27], Paul Hamilton[56], Ramona Leenings[25], Raymond Salvador[27], Robin Bülow[57], Romain Colle[32,58], Sara Poletti[31], Sarah Whittle[59], Sheri-Michelle Koopowitz[7], Sophia I. Thomopoulos[60], Susanne Meinert[25,61], Thomas Lancaster[62,63], Tiffany C. Ho[64,65], Tilo Kircher[35], Tim Hahn[25], Tony T. Yang[64], Udo Dannlowski[25], Yara J. Toenders[30], Yasumasa Okamoto[66], Yolanda Vives-Gilabert[67], Zeynep Basgoze[17], Dick J. Veltman[11],

Christopher R. K. Ching[68], Lianne Schmaal[30,46], Paul M. Thompson[54,60], Roberto Goya-Maldonado[1,*], for the ENIGMA Major Depressive Disorder working group[69]

Affiliations:

[1] Laboratory of Systems Neuroscience and Imaging in Psychiatry (SNIP-Lab), Department of Psychiatry and Psychotherapy, University Medical Center Göttingen (UMG), Georg-August University, Göttingen, Germany;

[2] College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China;

[3] Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne, Parkville, Victoria, Australia;

[4] FIDMAG Germanes Hospitalàries Research Foundation, Hospital Clinic, University of Barcelona, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Barcelona, Catalonia, Spain;

[5] Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Scotland, UK;

[6] SoCAT Lab, Department of Psychiatry, School of Medicine, Ege University, Izmir, Turkey;

[7] Department of Psychiatry & Mental Health, Neuroscience Institute, University of Cape Town, Cape Town, South Africa;

[8] Center for Neuroscience, University of Pittsburgh, Pittsburgh, PA, USA; Center for Neural Basis of Cognition, University of Pittsburgh, Pittsburgh, PA, USA;

[9] Department of Biomedical Sciences of Cells and Systems, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands;

[10] Department of Psychiatry, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Neuroscience, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands;

[11] Amsterdam University Medical Centers, location AMC, Department of Radiology and Nuclear Medicine, Amsterdam, the Netherlands;

[12] Sorbonne University, Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France;

[13] Department of Psychiatry & Behavioral Sciences, The University of Texas Health Science Center at Houston, Houston, TX, USA;

[14] Center Of Excellence On Mood Disorders, Louis A. Faillace, MD, Department of Psychiatry and Behavioral Sciences at McGovern Medical School, The University of Texas Health Science Center at Houston;

[15] Department of Psychiatry and Psychotherapy, Jena University Hospital, Jena, Germany;

[16] Department of Psychology, University of Minnesota, Minneapolis, MN, USA;

[17] Department of Psychiatry and Behavioral Science, University of Minnesota Medical School, Minneapolis, MN, USA;

[18] Section on the Neurobiology and Treatment of Mood Disorders, National Institute of Mental Health, Bethesda, MD, USA;

[19]Department of Biomedical Sciences, Florida Stateu University, Tallahassee FL;

[20]SA MRC Research Unit on Risk & Resilience in Mental Disorders, Department of Psychiatry & Neuroscience Institute, University of Cape Town, Cape Town, South Africa;

[21] Cardiff University Brain Research Imaging Center, Cardiff University, Cardiff, UK;

[22] MRC Center for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK;

[23] Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK;

[24] School of Mental Health and Neuroscience, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, 6229 ER, the Netherlands;

[25] Department of Psychiatry, Psychotherapy and Psychosomatics, Medical School, RWTH Aachen University, Germany;

[26] Institute for Translational Psychiatry, University of Münster, Germany;

[27] FIDMAG Germanes Hospitalàries Research Foundation, Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Barcelona, Catalonia, Spain;

[28] Hospital Clinic, University of Barcelona, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Barcelona, Catalonia, Spain

[29] Orygen, The National Centre of Excellence in Youth Mental Health, Parkville, VIC, Australia;

[30] Centre for Youth Mental Health, The University of Melbourne, Parkville, VIC, Australia;

[31] Division of Neuroscience, IRCCS Scientific Institute Ospedale San Raffaele, Milano, Italy;

[32] MOODS Team, CESP, INSERM U1018, Faculté de Médecine, Univ Paris-Saclay, Le Kremlin Bicêtre 94275, France;

[33] Service Hospitalo-Universitaire de Psychiatrie de Bicêtre, Hôpitaux Universitaires Paris-Saclay, Assistance Publique-Hôpitaux deParis, Hôpital de Bicêtre, Le Kremlin Bicêtre F-94275, France;

[34] Departments of Psychiatry and Pediatrics, University of Calgary, Calgary, AB, Canada;

[35] Department of Psychiatry and Psychotherapy, University of Marburg, Rudolf Bultmann Str. 8, 35039 Marburg, Germany;

[36] Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany;

[37] Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Scotland, UK;

[38] Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany;

[39] Department of Psychology, Stanford University, Stanford, CA, USA;

[40] Experimental Therapeutics and Pathophysiology Branch, National Institute for Mental Health, National Institutes of Health, Bethesda, MD;

[41] FIDMAG Germanes Hospitalàries Research Foundation, Hospital Clinic, University of Barcelona, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Barcelona, Catalonia, Spain;

[42] West Region, Institute of Mental Health, Singapore;

[43] Yong Loo Lin School of Medicine, National University of Singapore, Singapore;

[44] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore;

[45] German Center for Neurodegenerative Diseases (DZNE), Site Rostock/ Greifswald, Greifswald, Germany;

[46] Orygen, Parkville, VIC, Australia;
[47]Amsterdam UMC location Vrije Universiteit Amsterdam, Department of Psychiatry, Boelelaan 1117, Amsterdam, The Netherlands;

[48] GGZ inGeest Mental Health Care, Amsterdam, The Netherlands;

[49] Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia;

[50] Centre for Advanced Imaging, The University of Queensland, Brisbane, QLD, Australia;

[51] Sant Pau Mental Health Research Group, Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau, Barcelona, Catalonia, Spain. CIBERSAM, Madrid, Spain;

[52] Clinical Affective Neuroimaging Laboratory, Leibniz Institute for Neurobiology, Magdeburg, Germany;

[53] Meditation Research Program, Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA;

[54] Imaging Genetics Center, Mark & Mary Stevens Neuroimaging and Informatics InstituteInstitute for Neuroimaging and Informatics, Keck School of Medicine, University of Southern California, Marina del Rey, CA 90274, USA;

[55] Department of Psychiatry & Mental Health, Neuroscience Institute, University of Cape Town, Cape Town, South Africa;

[56] Center for Social and Affective Neuroscience, Department of Biomedical and Clinical Sciences, Linköping University, Linköping, Sweden;

[57] Institute for Radiology and Neuroradiology, University Medicine Greifswald, Greifswald, Germany;

[58] Service Hospitalo-Universitaire de Psychiatrie de Bicêtre, Hôpitaux Universitaires Paris-Saclay, Assistance Publique-Hôpitaux deParis, Hôpital de Bicêtre, Le Kremlin Bicêtre F-94275, France;

[59] Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne & Melbourne Health, Melbourne, VIC, Australia;

[60] Imaging Genetics Center, Mark & Mary Stevens Neuroimaging and Informatics Institute, Institute for Neuroimaging and Informatics, Keck School of Medicine, University of Southern California, Marina del Rey, CA 90274, USA;

[61] Institute for Translational Neuroscience, University of Münster;

[62] Cardiff University Brain Research Imaging Centre, Cardiff University, Cardiff, UK;

[63] MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK;

[64] Department of Psychiatry and Behavioral Sciences, Division of Child and Adolescent Psychiatry, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA;

[65] Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA;

[66] Department of Psychiatry and Neurosciences, Hiroshima University, Hiroshima, Japan;

[67] Intelligent Data Analysis Laboratory (IDAL), Department of Electronic Engineering, Universitat de València, Valencia, Spain;
[68] Imaging Genetics Center, Mark & Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Marina del Rey, CA 90274, USA;
[69] https://enigma.ini.usc.edu/ongoing/enigma-mdd-working-group/

***Corresponding author:**
PD Dr. Roberto Goya-Maldonado
Laboratory of Systems Neuroscience and Imaging in Psychiatry (SNIP-Lab)
Department of Psychiatry and Psychotherapy
University Medical Center Göttingen (UMG)
Von-Siebold Str. 5, 37075 Göttingen
e-mail: roberto.goya@med.uni-goettingen.de

Contributions

RGM and VB conceptualized and developed the analysis pipeline, which was approved by ENIGMA MDD working chair LS, co-chair DJV, ENIGMA PI PMT and all ENIGMA MDD members. VB performed all the analyses mentioned in the manuscript and RGM closely supervised them. TEG and EP helped collecting and preparing the data from all participating cohorts. All authors participated in collecting and preprocessing data from their respective sites, reviewed and provided intellectual contribution to the manuscript.

Abstract

Major depressive disorder (MDD) is a complex psychiatric disorder that affects the lives of hundreds of millions of individuals around the globe. Even today, researchers debate if morphological alterations in the brain are linked to MDD, likely due to the heterogeneity of this disorder. The application of deep learning tools to neuroimaging data, capable of capturing complex non-linear patterns, has the potential to provide diagnostic and predictive biomarkers for MDD. . However, previous attempts to demarcate MDD patients and healthy controls (HC) based on segmented cortical features via linear machine learning approaches have reported low accuracies. In this study, we used globally representative data from the ENIGMA-MDD working group containing an extensive sample of people with MDD (N=2,772) and HC (N=4,240), which allows a comprehensive analysis with generalizable results. Based on the hypothesis that integration of vertex-wise cortical features can improve classification performance, we compared Convolutional Neural Networks (CNN) with Support Vector

Machines (SVM), with the expectation that the former would outperform the latter. As we analyzed a multi-site sample, we additionally applied the ComBat harmonization tool to remove potential nuisance effects of site. We found that both classifiers exhibited close to chance performance (balanced accuracy CNN: 51%; SVM: 53%), when estimated on unseen sites. Slightly higher classification performance (balanced accuracy CNN: 58%; SVM: 55%) was found when the cross-validation folds contained subjects from all sites, indicating site effect. In conclusion, the integration of morphological features, the use of the non-linear classifier did not lead to the differentiability between MDD and HC. Future studies are needed to determine whether more sophisticated models or functional data integration will lead to a higher performance in this diagnostic task.

Introduction

Major depressive disorder (MDD) is a clinically heterogeneous psychiatric disorder manifested by low mood, anhedonia, impaired cognition, sleep disturbances, loss of energy, suicidal thoughts and appetite loss or gain. MDD dramatically impacts the daily functioning of patients and is currently the leading cause of disability worldwide [243]. Therefore, early diagnosis and optimal allocation of the proper treatment are critical. Unfortunately, the current treatment strategies present a response rate and remission as low as of 36.8% after a first treatment [20], [244], [245]. Thus, as proposed in the realms of systems medicine, we expect that by identifying brain patterns that classify patients at the individual level, we may open new biomarker-based avenues for the development of more personalized and effective treatments.

Neuroimaging techniques, such as magnetic resonance imaging (MRI), enable a non-invasive macro-scale view of human brain structure at the millimeter level of resolution. Initial neuroimaging studies used univariate approaches to reveal structural brain differences in MDD compared to healthy controls (HC) [52], [60], [63], identifying reduced hippocampal and frontal lobe volume. However, these studies had limited sample sizes and the more recent large sample studies have reported small effect sizes [71], [92], [93], [97], highlighting the absence of a single neuro-anatomical biomarker associated with MDD. The search for more complex biomarkers, which may include the interaction between different neuro-anatomical features,

can be conducted via machine learning (ML) algorithms (especially deep learning (DL) algorithms) applied to the MDD vs. HC classification task.

Like univariate approaches, ML and DL studies reported varying classification accuracies from 53% to 91% [127], [190]. The high variability of classification performances and the lack of consistent biomarkers can partly be explained by the small sample sizes, as it was demonstrated by Flint and colleagues [142]. Supplementing this, a study by Stolicyn based on cortical and subcortical morphological features, reported high accuracy of 75% in the small sample, which was not replicated in an independent large UK Biobank dataset, achieving only 54% [143].

Another factor that may inflate classification accuracies are related to study-site effects. The site-effect corresponds to site-specific characteristics other than diagnosis – such as scanner type, acquisition protocol, demographic differences, and inclusion and exclusion criteria, which may bias classification accuracies. Solanes demonstrated how site effect may contribute to both inflated and deflated classification accuracies [150]. There are numerous ways to tackle site-effect and improve model generalizability, from linear and non-linear ComBat harmonization tools [203], [205] to embedding site confounders directly to the model [215]. However, to overcome the difficult point of the heterogeneity of MDD and the lack of replicability and generalization of the models, the investigation of very large samples of participants with global representation is fundamental.

Using a large-scale dataset from the ENIGMA-MDD consortium, we compared the classification performance of commonly used ML models to predict diagnosis based on cortical and subcortical parcellations of morphological features (surface areas, thicknesses, volumes) [246]. Overall, results showed a trend that may highlight the contribution of site-effects to classification performance. Specifically, there was a clear difference in classification performance dependent on the cross-validation splitting techniques used in training. Site-splitting generally performed at close to chance level for all classifiers, while mixing sites across splits achieved up to 62% balanced accuracy with an SVM. Of note, data harmonization using ComBat removed the site effect and resulted in a balanced accuracy of 52% with SVM. Based on these findings, we concluded that most commonly used ML classification algorithms could not successfully discriminate MDD from HC individuals based on morphological features organized in pre-defined Desikan-Killiany atlas parcellations. It remains unclear whether more fine-grained information of morphological features, displayed in a vertex-wise organization, could outperform the classification based on parcellation atlas-distributed information.

There are few directions in improving classification based on morphological information. First, previous ML studies considered surface area, thickness, and volume characteristics only, while

the information on the cortical shape , such as gyral and sulcal shape patterns, was not integrated into analyses. Cortical gyrification modalities are affected by genetic and non-genetic factors [247], [248], alterations of which were associated with MDD [152], [153]. Multimodal morphological feature analysis, including myelination, gray matter, and curvature, revealed a correlation between cortical differences and MDD-associated genes [249]. Therefore, the addition of shape modalities, such as cortical curvature and sulcal depth, to cortical thickness could enhance the classification performance, as demonstrated for sex and autism classification [155].

Another direction to improve low classification performance is to deploy more advanced classification algorithms. DL methods have gained popularity in the neuroimaging field as a promising tool for cortical surface reconstruction [250], image preprocessing [251], and cortical parcellation [252]. Furthermore, DL is widely evaluated as a predictive tool in psychiatry, showing higher or at least the same classification performance compared to linear models [129], [149], [155], [164], [237], [253]. The analysis of cortical morphometric features can be conducted via convolutional neural network (CNN) [254], designed to reveal complex patterns in 2D images. In order to apply such 2D CNN in the classification, it requires 3D cortical features to be initially projected into 2D image space. Nevertheless, this step may inevitably create distortion in spatial properties such as shape, area, distance, and direction. Several approaches were implemented before, such as latitude/longitude projection [255] and optimal mass transport (OMT) projection [155], [256], which preserves area. However, the impact of these projection methods on classification performance were never directly compared in the neuroimaging field.

The main goal of this study was to distinguish MDD from HC individuals based on integrated cortical morphological features, including sulcal depth, curvature, and thickness. These features were analyzed via SVM with linear kernel and CNN architecture (pre-trained DenseNet [257]), which demonstrated its superiority over simpler models in autism vs. HC classification task [155]. SVM was chosen as it is a robust shallow ML model, frequently used in neuroimaging settings [178], [258], [259]. We compared classification performance of these methods to understand the role of complex non-linear patterns in MDD manifestation. We used balanced accuracy, sensitivity, specificity and AUC as the classification performance metrics. Higher classification performance of the CNN model will presume the presence of spatially complex patterns in brain morphology, which are relevant for classification. Furthermore, we aimed to estimate the relevance of integrating cortical thickness and shape characteristics (sulcal depth, curvature and thickness) into the analysis by training the models with all features combined and

by considering them separately. Similar to our previous study [246], different cross-validation (CV) approaches were evaluated: 1) splitting the data by balancing age and sex distribution across all CV folds (Splitting by Age/Sex) and 2) performing leave-sites-out CV in order to estimate the performance on the unseen during the training sites (Splitting by Site). This approach allowed us to estimate if the model's performance is biased towards site-related or demographic factors. The difference between results in both splitting strategies presumes the presence of the site effect we addressed by harmonizing the data in both splitting strategies via ComBat. In summary, we hypothesized that: 1) CNN can differentiate MDD from HC based on the provided features 2) Integration of thickness and shape brain characteristics will contribute positively to the classification performance. Site-effect, if present, will be addressed via ComBat. Additionally, we compared two projection methods, latitude/longitude and OMT projections by performing auxiliary single-site sex classification based on three of the largest cohorts to explore whether classification performance may vary according to 2D projection method. We had no *a priori* hypothesis for the projection methods' analysis.

Material and methods

Study participants and study design

We analyzed a large-scale multi-site sample provided by ENIGMA-MDD working group, comprising 2,772 MDD and 4,240 HC individuals, from 30 cohorts worldwide. Details on inclusion/exclusion criteria and sample characteristics can be found in Supplementary Table 1. Subjects with missing information on demographic data or any of cortical surface mesh files (l(r).sulc, l(r).curv, l(r).thickness) were excluded from the analysis (476 and 6 % excluded). All participating cohorts confirmed approval from their corresponding institutional review boards and local ethics committees as well as collected written consent of all participants.

Image processing and analysis

Each site acquired structural T1-weighted MRI scans of participants and preprocessed them according to ENIGMA Consortium protocol (http://enigma.ini.usc.edu/protocols/imaging-protocols/). This pipeline includes the segmentation of T1-weighted MRI volumes, tessellation, topology correction, and spherical inflation of the white matter surface. Detailed information on the acquisition protocols and scanner model in each cohort can be found in Supplementary Table 2. Cortical meshes were generated during FreeSurfer preprocessing in every site. Cerebral cortex meshes were then extracted from the FreeSurfer unsmoothed fsaverage6 template,

effectively removing intracranial volume (ICV) differences (see Supplementary Figure 1) and yielding 37,747 and 37,766 vertices for the left and right hemispheres, respectively. We analyzed vertex-wise features, such as sulcal depth, curvature, and thickness, both as integrated features and separately (Figure 1).

Considering the absence of well-established pre-trained on cortical meshes CNN models, we projected 3D cortical surfaces into 2D images and apply CNN model, which was pre-trained on natural images. There are few studies applying different projection methods such as latititude/longitude project and area-preserving maps (e.g., Seong et al., 2018; Gao et al., 2021). Of note, the latitude/logitude method, in which cortical mesh is first re-sampled to the sphere and consequently mapped to the 2D grid, creates strong area distortions in the edges and near the medial wall close to subcortical regions [255]. Both methods may (differentially) influence subsequent classification performances, but to the best of our knowledge, no studies to date have directly compared this in one study using the same samples. Thus, we applied both 2D projection methods to the cortical meshes, resulting in 224×224 pixels images for each hemisphere. The images were normalized to present mean of 0 and standard deviation of 1.

Data Splitting

To assess potential biases in the model's decision-making, we performed 10-fold cross-validation (CV) by splitting the data according to 1) demographic covariates, in which age and sex distribution were balanced and subjects from each site are equally distributed across all CV folds (*Splitting by Age/Sex*), and 2) site affiliation, where each site was contained only in one CV fold (*Splitting by Site*). In both strategies, 9 CV folds were used for training, while one remaining CV fold was used as a test set. This procedure was repeated iteratively until every CV fold was used as a test set. In the Splitting by Age/Sex strategy, effect of demographic factors on the classification performance is reduced, as the model is trained and tested on the same demographics. Nevertheless, the site-related differences may bias the decision-making of the classification models [246], which is directly addressed in Splitting by Site. This strategy demonstrates how well the model trained on one set of sites can be applied to the data from unseen sites. As the number of sites exceeds the number of folds, we distributed the sites across the folds to balance the number of subjects in every fold as close as possible by iteratively distributing the largest sites across all 10 folds. Smallest folds were added subsequently to further even the number of subjects in every fold. Overall, the difference in the classification

results between these two splitting strategies may indicate the existence of the site effect. More detailed description of both splitting strategies can be found elsewhere [246].

MDD vs HC classification

After the data-splitting step, the primary analysis was carried out. Firstly, we residualized all features normatively, removing linear age and sex dependencies. To avoid data leakage, age and sex regressors were estimated on the healthy subjects from the training set (9 CV folds) and then applied to the training and test set (1 CV fold) for patients and HC. Next, the classification algorithms were trained on the training folds, and classification performance was estimated on the test fold. As demonstrated by Dinga and colleagues, accuracy alone should be avoided as it does not account for class frequencies [137]. Thus, the algorithms were evaluated according to categorical measures, including balanced accuracy, sensitivity, specificity, and rank-based measure – AUC, allowing for a broad overview of performance. For model-level assessment [238], we performed the classification using all features combined and then using features separately to assess the final classification performance. We evaluated the classification performance of a robust shallow model - SVM with linear kernel, and deep learning model - DenseNet pre-trained on natural images from ImageNet dataset [260], which has been demonstrated as a robust convolutional neural network for image classification both for natural images as well as in neuroimaging [155], [257]. When DenseNet is trained on a single data domain, left and right hemisphere images are propagated through corresponding left and right DenseNets, the fully connected layers of which are concatenated. The resulting feature vectors are fed to the output layer. For the whole-brain all-features analysis, we combined the features extracted from every feature and hemisphere, concatenate them and feed them to the output layer. For SVM, all considered images were flattened and then concatenated to a single array. The analysis pipeline is presented in Figure 1. To mitigate site-related differences, which may potentially bias the classification results, we additionally performed the analysis with harmonizing all of the features via ComBat. Variance explained by age and sex was preserved during this harmonization step. Next, we residualized features normatively, as described above, and train/test the models. Application of ComBat differed for both splitting strategies. In short, ComBat parameters estimated on the training set were applied to the test set directly, in the splitting by Age/Sex. In splitting by Site, ComBat is applied twice. Firstly, we use ComBat to harmonize the training sites. Secondly, we apply ComBat to adjust test sites to the harmonized training sites, i.e. using the training sites as the reference batch [241]. A more detailed description of ComBat application can be found in our previous work [246].

Auxiliary analysis in projection methods

To explore and evaluate the potential impact of 2D projection methods on the classification performance, we compared both methods in their ability to classify healthy males from healthy females in 3 of the largest cohorts separately. The single-site classification was estimated via 10-fold CV on 411, 723, and 397 subjects, respectively. As usual, 9 CV folds were used for training, while one remaining CV fold was used as a test set. This procedure was repeated iteratively until every CV fold is used as a test set. In order to obtain an initial view of pre-trained CNN, we compared the balanced accuracies of two models: SVM with linear kernel and pre-trained DenseNet [257]. Furthermore, we used sex classification task to find the optimal hyperparameters for both SVM and DenseNet (see Supplementary Table 3). Finally, to examine the possible advantage of using SVM and the pre-trained CNN in the sex classification task, we compared the classification performance of both models.

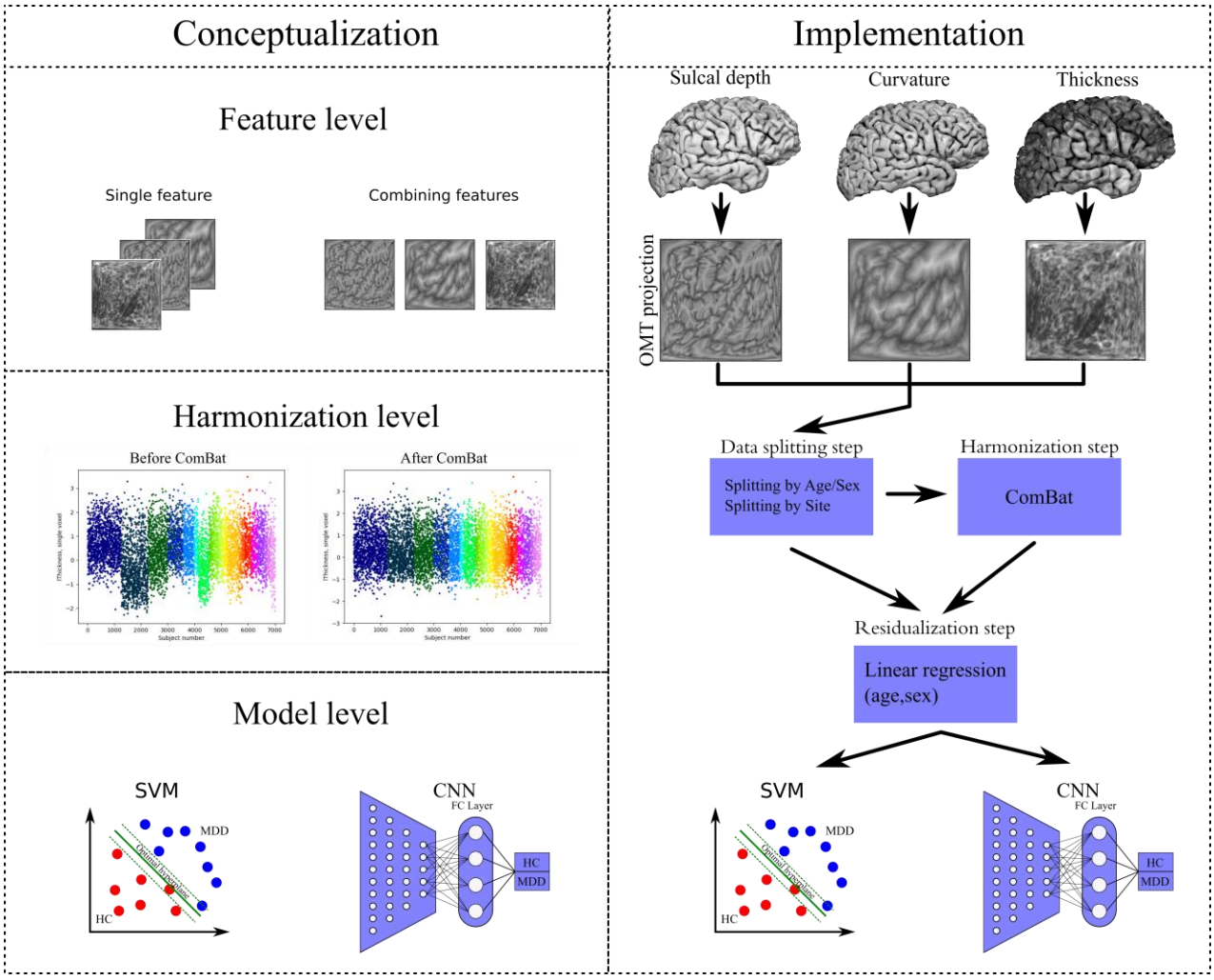


Figure 1: Proposed conceptualization levels and implementation of classification procedure. Left: Higher classification performance in MDD vs HC classification task can be achieved by implementing deep ML models, such as CNN, in comparison to a shallow ML model, for example, SVM. Furthermore, the analysis of integrated morphometric features can provide a more detailed description of cortical organization than separated features, leading to better differentiability of MDD from HC. The application of ComBat may improve the generalizability of results as site-related differences are removed. **Right:** Cortical sulcal depth, curvature, and thickness are first projected into the 2D grid and then transformed into 2D images using OMT projection. We split the data into 10 CV folds according to age and sex (Splitting by Age/Sex) and according to the site belonging (Splitting by Site). After the residualization step, where the age and sex effect are regressed out linearly, we train and test SVM and DenseNet on the diagnosis classification.

Results

Participants and Data Splitting

We detected substantial differences in age (78% of pairwise comparisons between cohorts were significant, t-test, $p<0.05$) and sex (47%, t-test, $p<0.05$) across cohorts. The full demographic profile is presented in the Table 1. As expected, Splitting by Age/Sex resulted in more balanced CV folds with respect to number of subjects, age and sex distributions, while folds created by Splitting by Site were more uneven on these characteristics (Table 2).

Table 1: Participating sites. The total number of subjects, number of MDD patients and number of HCs, as well as their mean age (in years) and sex (number and % of females) is presented.

Cohort		Number of subjects	Age Mean (SD)	Number of Females (%)
AFFDIS	**Total**	79	39.75 (14.67)	36(46)
	HC	46	39.87 (14.29)	22(48)
	MDD	33	39.58 (15.18)	14(42)
Barcelona-StPau	**Total**	94	46.66 (7.81)	72(77)
	HC	32	46.03 (8.00)	23(72)
	MDD	62	46.98 (7.68)	49(79)
CARDIFF	**Total**	40	46.55 (11.74)	27(68)
	HC	0	nan	nan
	MDD	40	46.55 (11.74)	27(68)
CSAN	**Total**	109	34.70 (12.88)	74(68)
	HC	49	33.20 (12.07)	34(69)
	MDD	60	35.92 (13.38)	40(67)
Calgary	**Total**	107	17.03 (4.12)	60(56)
	HC	52	15.81 (5.03)	29(56)
	MDD	55	18.19 (2.51)	31(56)
DCHS	**Total**	79	30.91 (6.71)	79(100)
	HC	61	31.49 (6.82)	61(100)
	MDD	18	28.94 (5.89)	18(100)
FIDMAG	**Total**	69	47.22 (12.29)	44(64)
	HC	34	45.94 (11.49)	22(65)
	MDD	35	48.46 (12.90)	22(63)
FOR2107Marburg	**Total**	738	36.30(13.39)	465 (63)
	HC	411	34.76(12.76)	257(63)
	MDD	327	38.24(13.91)	208(63)
FOR2107Munster	**Total**	395	31.66(12.09)	249 (63)
	HC	221	28.39(10.29)	140 (63)
	MDD	174	35.87(12.84)	109(63)
Houston	**Total**	290	28.72 (16.30)	169(58)
	HC	186	26.76 (15.91)	105(56)
	MDD	104	32.23 (16.39)	64(62)
Hiroshima	**Total**	319	41.93(12.36)	175(55)
	HC	169	39.87(12.36)	104(62)
	MDD	150	44.24(11.94)	71(47)
Jena	**Total**	107	46.76 (15.00)	52(49)
	HC	77	47.75 (15.93)	36(47)
	MDD	30	44.20 (11.92)	16(53)
MODECT	**Total**	42	72.71 (9.25)	28(67)
	HC	0	nan	nan
	MDD	42	72.71 (9.25)	28(67)
Melbourne	**Total**	245	19.42 (2.88)	130(53)
	HC	102	19.58 (2.97)	54(53)
	MDD	143	13.31 (2.80)	76(53)
Minnesota	**Total**	110	15.47 (1.89)	79(72)
	HC	40	15.68 (1.98)	26(65)
	MDD	70	15.36 (1.83)	53(76)
MOODS	**Total**	96	34.54(12.48)	65(68)
	HC	32	38.87(12.36)	104(62)
	MDD	64	44.25(11.95)	71(47)

Moraldilemma	**Total**	70	18.81 (1.94)	70(100)
	HC	46	18.50 (1.75)	46(100)
	MDD	24	19.42 (2.14)	24(100)
NESDA	**Total**	219	38.11 (10.32)	145(66)
	HC	65	40.29 (9.67)	42(65)
	MDD	154	37.19 (10.45)	103(67)
QTIM	**Total**	386	22.08 (3.25)	267(69)
	HC	284	22.11 (3.30)	190(67)
	MDD	102	22.01 (3.11)	77(75)
UCSF	**Total**	163	15.46 (1.31)	91(56)
	HC	88	15.32 (1.28)	42(48)
	MDD	75	15.63 (1.33)	49(65)
SHIP_START_2	**Total**	579	55.01 (12.57)	294(51)
	HC	443	55.44 (12.80)	198(45)
	MDD	136	53.59 (11.68)	96(71)
SHIP_TREND_0	**Total**	1229	50.15 (13.69)	607(49)
	HC	919	50.50 (14.18)	405(44)
	MDD	310	49.12 (12.04)	202 (65)
SanRaffaele	**Total**	45	49.07 (13.51)	32(71)
	HC	0	nan	nan
	MDD	45	49.07 (13.51)	32(71)
Sexpect	**Total**	40	36(9.69)	11 (27)
	HC	20	33.75(7.02)	3(15)
	MDD	20	38.25(11.34)	8(40)
Singapore	**Total**	38	39.50 (6.43)	18(47)
	HC	16	38.69 (4.59)	8(50)
	MDD	22	40.09 (7.43)	10(45)
Socat_dep	**Total**	179	37.85 (13.34)	161(90)
	HC	100	36.42 (13.57)	90 (90)
	MDD	79	39.66 (12.81)	71 (90)
StanfFAA	**Total**	32	32.71 (9.56)	32(100)
	HC	18	30.44 (9.96)	18(100)
	MDD	14	35.63 (8.14)	14(100)
StanfT1wAggr	**Total**	115	37.18 (10.27)	69(60)
	HC	59	37.24 (10.43)	36(61)
	MDD	56	37.11 (10.09)	33(59)
TAD	**Total**	39	16.03(1.14)	11(27)
	HC	0	nan	nan
	MDD	39	16.03(1.14)	11(27)
TIGER	**Total**	60	15.63 (1.34)	38(63)
	HC	11	15.18 (1.03)	5(45)
	MDD	49	15.73 (1.38)	33(67)
All sites	**Total**	7012	38.41(16.28)	4186(60)
	HC	4240	39.98(14.46)	2383(59)
	MDD	2772	39.57(15.28)	1803(61)

Table 2: Data splitting strategies. Differences manifested in age/sex distribution and number of subjects between corresponding folds per splitting strategy.

Splitting By Age/Sex				**Splitting by Site**			
Fold	**Number of subjects**	**Mean age (SD)**	**Number of Females (%)**	**Fold**	**Number of subjects**	**Mean age (SD)**	**Number of Females (%)**
0	708	38.34 (16.41)	434 (61)	**0**	1249	50.28 (13.78)	612 (49)
1	685	38.41 (16.51)	395 (58)	**1**	1005	36.01 (12.14)	577 (57)
2	692	38.59 (16.25)	441 (64)	**2**	738	36.30 (13.39)	465 (63)
3	709	37.99 (16.07)	428 (60)	**3**	579	55.00 (12.57)	294 (51)
4	704	38.74 (15.93)	417 (59)	**4**	563	33.06 (15.73)	374 (66)
5	708	38.90 (16.28)	415 (58)	**5**	596	26.42 (11.25)	370 (62)
6	693	38.09 (16.27)	423 (61)	**6**	559	36.89 (13.71)	372 (67)

7	716	38.3 (16.35)	431 (60)	**7**	589	35.71 (16.52)	356 (60)
8	689	38.55 (16.12)	396 (57)	**8**	546	28.70 (13.59)	359 (66)
9	708	38.14 (16.57)	406 (57)	**9**	588	33.99 (16.12)	407 (69)

MDD vs HC classification

First, we compared the performance of SVM and DenseNet for different splitting strategies (Table 3). In Splitting by Age/Sex, SVM achieved 0.551 ± 0.021 in balanced accuracy, while DenseNet yielded 0.578 ± 0.022. In Splitting by Site, both SVM and DenseNet models performed worse, yielding 0.528 ± 0.039 and 0.512 ± 0.019, respectively. The minor difference in classification performances for different splitting strategies indicated a potential site effect, which we addressed by applying ComBat. In Splitting by Age/Sex, the balanced accuracy of SVM with ComBat dropped to 0.478 ± 0.019, while the performance of DenseNet did not change and yielded 0.561 ± 0.015. In splitting by Site with ComBat, the performance of both models was similar and close to random chance, balanced accuracy yielded 0.520 ± 0.019 and 0.508 ± 0.020 for SVM and DenseNet respectively. Thus, we did not observe an improvement of models performances after data harmonization by ComBat.

Next, we explored if any of the considered feature modalities yields greater classification performance (Figure 2). In Splitting by Age/Sex, all data modalities yielded similar range of accuracies: thickness (SVM: 0.549 ± 0.020; DenseNet: 0.576 ± 0.019) compared to sulcal depth (SVM: 0.543 ± 0.022; DenseNet: 0.562 ± 0.019), and curvature (SVM: 0.531 ± 0.015; DenseNet: 0.567± 0.019), observed for both classification models. In Splitting by Site, sulcal depth (SVM: 0.523 ± 0.016; DenseNet: 0.515 ± 0.020), curvature (SVM: 0.513 ± 0.033; DenseNet: 0.516 ± 0.025) and thickness (SVM: 0.522 ± 0.038; DenseNet: 0.515 ± 0.022) also exhibited similar range of classification accuracies. Both models performed similarly for all feature types. These results demonstrate that integration of shape modalities with cortical thickness did not benefit the classification models. Results from explorative analyses for each hemisphere and for each feature modality per hemisphere showed no improvements in performance of the models (see Supplementary Table 4, Supplementary Figure 3).

Auxiliary sex prediction task

As an initial step, we also conducted a sex classification to explore, which projection method (Lat/Long, OMT) yields higher classification performance for both SVM and DenseNet

(Supplementary Figure 2). There was no clear difference between projection methods; however, we observed a consistently higher classification performance of DenseNet compared to SVM for all types of features and hemispheres. Considering previous success of OMT projection as a projection method applied on cortical surface and its property to preserve distances between vertices [155], we conducted our main analysis with OMT projection.

Table 3: Comparison of SVM and DenseNet classification performance on entire dataset using integrated whole brain feature modalities. The performance is evaluated via balanced accuracy, sensitivity, specificity, and AUC for each splitting strategy, with and without ComBat harmonization.

	Splitting by Age/Sex		**Splitting by Site**	
	No ComBat	With ComBat	No ComBat	With ComBat
SVM				
Balanced Acc	0.551 ± 0.021	0.478 ± 0.019	0.528 ± 0.039	0.520 ± 0.019
Sensitivity	0.477 ± 0.036	0.420 ± 0.024	0.490 ± 0.114	0.465 ± 0.033
Specificity	0.625 ± 0.030	0.536 ± 0.021	0.566 ± 0.124	0.574 ± 0.049
AUC	0.566 ± 0.021	0.490 ± 0.020	0.536 ± 0.062	0.520 ± 0.022
DenseNet				
	0.578 ± 0.022	0.561 ± 0.015	0.512 ± 0.019	0.508 ± 0.020
Balanced Acc	0.452 ± 0.102	0.401 ± 0.090	0.428 ± 0.172	0.466 ± 0.265
Sensitivity	0.704 ± 0.104	0.721 ± 0.092	0.596 ± 0.217	0.550 ± 0.241
Specificity	0.606 ± 0.026	0.595 ± 0.020	0.549 ± 0.076	0.544 ± 0.092
AUC				

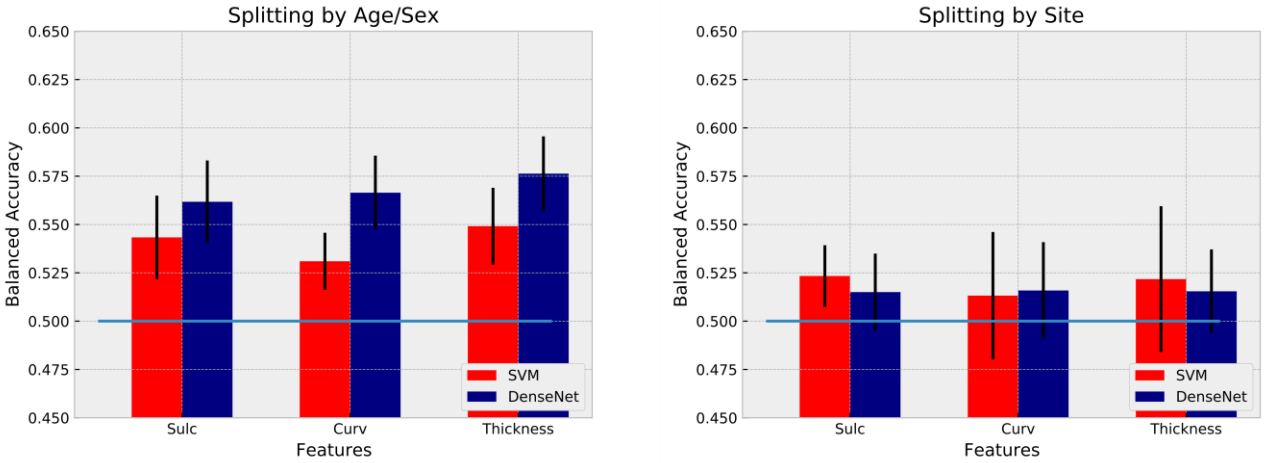


Figure 2: Single feature classification relevance. Balanced accuracy for both classification models when trained only on a subset of features: Sulcal depth only, curvature only, thickness only. We estimated the performance for both splitting strategies without harmonization step. Error bar represents standard deviation.

Discussion

In this work, we evaluated the diagnostic classification performance of SVM and DenseNet models, trained on cortical maps projected via OMT, including sulcal depth, curvature, and thickness, from a multi-site global dataset. Our analysis included 7,012 participants from 30 sites worldwide, allowing for a comprehensive and realistic overview of classification performances. Both models were evaluated in parallel using two different CV splitting strategies. In Splitting by Age/Sex, we obtained CV folds with comparable demographics; thus, the performance of the models should not be affected by these demographic variables. In Splitting by Site, sites were distributed across folds. Therefore, models were trained and tested on different sets of sites. This strategy is closer to application of diagnostic classification models in clinical practice, and allowed for realistic estimation of classification performance on unseen sites. Overall, the classification performances of both models were similar: In Splitting by Age/Sex, DenseNet achieved 58% vs. 55% for SVM. In Splitting by Site, the difference was even more negligible: 51% vs. 52%. Both models performed better in Splitting by Age/Sex, implying the presence of a confounding site effect, most likely arising from differences in image acquisition and settings. ComBat did not improve the accuracy of models and splitting strategies. The exploration of single-features revealed similar range of classification

performances of both models for all morphometric features, both in Splitting by Age/Sex and in Splitting by Site.

SVM vs DenseNet

Previous ML mega-analyses based on structural MDD vs HC classifications considered only shallow linear and non-linear ML models, such as SVM, penalized logistic regression and decision tree [142], [143], [246]. In this study, we extended the diagnostic classification approach by comparing the performance of shallow linear model - SVM with a linear kernel to a highly non-linear deep CNN classifier applied to vertex-wise cortical information. The explorative results of sex classification applied to HC revealed higher classification performance of the DenseNet compared to the SVM (see Supplementary Figure 2) for all data modalities. The higher accuracy suggests that DenseNet was able to capture non-linear sex dependencies that were present in the cortical maps. The superiority of CNN over SVM in the sex classification task was in line with previous study conducted on the same vertex-wise cortical maps [155]. On the contrary, another large sample study revealed no advantage of using any deep architectures over simpler models in predicting demographic factors [237]; therefore further tests in even bigger samples are required. Nevertheless, both models exhibited a similar range of accuracies, close to random chance, in the main task of MDD versus HC classification. Therefore, the application of CNN did not yield the expected improvement for detecting combined (or separated) structural cortical features that discriminate patients from controls.

Similar performance of the linear SVM and non-linear CNN model may be due to the absence of non-linear interactions between different cortical regions, significant for the MDD detection. Furthermore, the analyzed sample is highly heterogeneous in terms of demographic and clinical covariates, potentially interfering with the main task and lowering the main task performance. There are several directions for improving CNN performance. First, the considered model was pre-trained only on natural images from ImageNet. The model could be subsequently pre-trained on cortical projections from an independent large sample using immediate task, for example predicting sex as it was performed in Gao's study [155]. Furthermore, one could use more than one intermediate task to optimize the weights of the neural network, for example, predicting demographical or clinical covariates. This approach is broadly known as multi-task learning [261], the usefulness of which in the neuroimaging domain was already demonstrated [149], [215].

Secondly, the multi-task approach could be used to "unlearn" undesired biases. In our analysis, site-related differences were removed via ComBat. One could train the network to perform the main task while unlearning the scanner parameters, as was successfully demonstrated by Dinsdale and colleagues [262]. Furthermore, one could replace the residialization step in the same manner by making the network unlearn age and sex dependencies. In line with our previous analysis, we linearly regressed out age and sex dependencies from the cortical features using normative approach [246]. Considering the greater performance of the CNN model in predicting sex, we can speculate the presence of non-linear male-female differences in cortical morphology. Thus, unlearning age- and sex-related dependencies could improve classification performance.

Cortical morphological maps as diagnostic biomarkers for MDD

To the best of our knowledge, this the first study to combine cortical thickness, sulcal depth, and curvature in order to classify MDD vs HC. Furthermore, the previous large sample ML studies incorporated only low-resolution atlas-based thickness characteristics. In our approach, we analyzed vertex-wise information, providing a richer and more detailed description of brain characteristics than atlas-derived regional measures. Even so, the integration of complementary cortical characteristics did not lead to higher classification performances compared to the accuracies obtained from the single cortical features, regardless of the data splitting strategy and the classification model. In Splitting by Site, no feature yielded an accuracy substantially higher than random chance accuracy, indicating the failure of both models to capture MDD-specific alterations. Furthermore, the analysis of finer-grained cortical maps, even for thickness alone, did not result in higher classification performance, compared to ML performance levels observed in our previous study [246]. Thus, the assumption that higher resolution would lead to greater classification performance did not hold in this study, as all results were close to the chance level, in line with previous attempts in classifying MDD [142], [143], [246]. It suggests the general absence of prominent gray matter alterations that can serve as diagnostic criteria of MDD.

While we combined complementary characteristics in the analysis, the interaction between thickness and shape was not considered. According to recent evidence, local cortical shape correlates with thickness [263]. Combined thickness-shape patterns should be further explored as a potential structural predictors of MDD. Reduced myelination was associated with MDD

[264]–[266], which could lead to structural reorgnization of cortical features, making it a potential aspect to be investigated. Furthermore, subcortical morphological characteristics can enhance the classification by taking into account structural modifications in cortico-subcortical loops associated with MDD [93].

Integration of morphological characteristics with cytoarchtectonic and functional information may allow better contextualization of MDD-related alterations, as demonstrated in transdiagnostic study by Hettwer [267] with a potential to achieve higher classification performance [268], [269]. Brain topology can be described via connectome - a connectivity architecture of the brain. As nodes of brain connectome exhibited elevated susceptibility to brain disorders [270], graph analytical approaches could lead to stronger differentiability between MDD and HC. Moreover, subject-specific parcellation schemes could be applied to compute structural and functional connectomes [271], and further analyzed by sophisticated classification models suited taking into account neural architecture, for example graph neural network [272].

Data Splitting and Site Effect

Several multi-site psychiatric neuroimaging studies directly demonstrated how different splitting strategies might introduce unwanted biases in inflated classification performances [164], [207], [246]. In Splitting by Age/Sex, trained models are unbiased regarding demographic factors, while in Splitting by Site the site affiliation is controlled, therefore we addressed the generalizability of the models applied to unseen sites. Similar to the results from our previous study [246], the classification performance of both SVM and DenseNet was higher in Splitting by Age/Sex, up to 58%, compared to Splitting by Site, close to random chance. This discrepancy indicates the existence of hidden site-related biases influencing classification performance. As this trend is an appearing phenomenon in multi-site mega-analyses [159], [164], we strongly encourage the application of different splitting strategies in future multi-site machine learning studies.

The low accuracy of both models in Splitting by Site strategy is either due to the presence of a strong site-effect, hindering the ability of the models to capture diagnosis-related differences, or the general inability of both models to find meaningful alterations associated with MDD. We addressed site-effect via ComBat. Lastly, there is a possibility that subject-level prediction based on cortical features is in general impossible. As the Combat was never applied to cortical vertex-wise projections, we visually inspected the effect of ComBat on a single pixel for every

feature type (See Supplementary Figure 4). Application of ComBat resulted in more homogenous value distribution across cohorts, in line with previous studies analyzed atlas-based features [205], [246]. Nevertheless, it did not lead to improvement in accuracies in the Splitting by Site. While demographic covariates were preserved during harmonization step, ComBat could over-correct the data [273], causing a part of MDD-related associations to be removed with the site-effect. Against this, more careful consideration of the site-effect is required in the future studies.

In Splitting by Age/Sex, the accuracy of both models dropped (SVM: 55% - 48%; DenseNet: 58% - 56%) when ComBat was applied. The decrease of model's performances close to the levels in Splitting by Site indicates that initial higher results are most likely arose from site-related biases, as sites are evenly distributed across all CV folds. To further validate this assumption, we performed the classification with balanced ratio between HC and MDD in every site in Splitting by Age/Sex, resulting in close to random chance accuracies of both models. Noticeably, DenseNet was less affected by the application of ComBat, reflecting potential non-linear site-related differences remained in the dataset after ComBat, which is in line with findings by Solanes and colleagues [274]. Therefore, we recommend ComBat to be applied only in combination with more robust and simple models, such as SVM, while more sophisticated models should directly incorporate site information as an additional input.

Conclusion

In this study, we tested if more advanced classification algorithms applied to high-resolution morphometric shape characteristics can improve MDD vs. HC classification. Splitting the data according to demographic variables and according to site allowed a comprehensive analysis of model's performances and biases. We detected site effects, which we addressed with the ComBat harmonization tool. Both shallow and deep ML models exhibited low, close to chance accuracies. Furthermore, the integration of high-resolution cortical thickness and shape characteristics did not lead to greater classification performance than previously analyzed atlas-based cortical features. Lastly, the application of ComBat did not improve the accuracy in both splitting strategies. According to our and previous ML studies, it is unlikely that structural MRI will provide diagnostic biomarkers of depression. Thus, further investigation of other MRI modalities, including fMRI and DTI, is required.

Declaration of Competing Interest

PMT and NJ received a research grant from Biogen, Inc., for research unrelated to this manuscript. HJG has received travel grants and speakers honoraria from Fresenius Medical Care, Neuraxpharm, Servier and Janssen Cilag as well as research funding from Fresenius Medical Care unrelated to this manuscript. JCS has served as a consultant for Pfizer,Sunovion, Sanofi, Johnson & Johnson, Livanova, and Boehringer Ingelheim. The remaining authors declare no conflict of interest.

Data availability

Authors are not allowed to share the data of participating sites to third parties inside or outside the ENIGMA MDD consortium. Some sites may provide data upon request.

Acknowledgements

ENIGMA MDD: This work was supported by NIH grants U54 EB020403 (PMT), R01MH116147 (PMT), and R01 MH117601 (NJ & LS), and the NSFC grants 61722313 (LLZ), and 62036013 (LLZ). LH was funded by the Rubicon award (grant number 452020227) from the Dutch NOW. AFFDIS: this study was funded by the University Medical Center Goettingen (UMG Startfoerderung) and VB and RGM are supported by German Federal Ministry of Education and Research (Bundesministerium fuer Bildung und Forschung, BMBF: 01 ZX 1507, "PreNeSt - e:Med"). Calgary: Alberta Children's Hospital Foundation, Canadian Institutes for Health Research. Cardiff: This work was supported by a Medical Research Council (G 1100629) grant to DEJ Linden and a PhD studentshp by Health Research Wales (HS/14/20) for DMEM. CSAN: This work was supported by grants from Johnson & Johnson Innovation (S.E.), the Swedish Medical Research Council (S.E.: 2017–00875, M.H.: 2013–07434, 2019–01138), the ALF Grants, Region Östergötland M.H., J.P.H.), National Institutes of Health (R.D.: R01 CA193522 and R01 NS073939), MD Anderson Cancer Support Grant (R.D.: P30CA016672). Dep-arrest clin: BCD is supported by a NHMRC CJ Martin fellowship (app 1161356). DCHS: supported by the Medical Research Council of South Africa. The DCHS was funded by the Bill and Melinda Gates Foundation (OPP1017641), and received additional support from the South African Medical Research Council. ENIGMA Central/USC: This work was supported by the National Natural Science Foundation of China (61722313 and 62036013). ETPB: Funding for this work was provided by the Intramural Research Program at the National Institute of Mental Health, National Institutes of Health (IRP-NIMH-NIH; ZIA-MH002857),

by a NARSAD Independent Investigator to Dr. Zarate, and by a Brain & Behavior Mood Disorders Research Award to Dr. Zarate. Suported by the NIMH Intramural Research Program. FIDMAG: This work was supported by the Generalitat de Catalunya (2014 SGR 1573) and Instituto de Salud Carlos III (CPII16/00018) and (PI14/01151 and PI14/01148). PFC was supported by a Sara Borrell grant (CD19/00149, Instituto de Salud Carlos III) and a fellowship from "la Caixa" Foundation (ID 100010434, fellowship code LCF/BQ/PR22/11920017). FOR2107-Marburg: This work was funded by the German Research Foundation Igor Nenadic (NE 2254/1-2). This work was funded by the German Research Foundation (DFG grant FOR2107, KI588/14-1 and FOR2107, KI588/14-2 to Tilo Kircher, Marburg, Germany). FOR2107-Muenster: This work was funded by the German Research Foundation (DFG), Udo Dannlowski (co-speaker FOR2107, DA 1151/5-1, DA 1151/5-2, grant DA1151/9-1, DA1151/10-1 and DA1151/11-1) and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/022/22 to UD). Houston: JCS has received research grants from Compass, Alkermes, and Allergan. Melboune: This work was supported by National Health and Medical Research Council of Australia (NHMRC) Project Grants (1064643) to BJH and to CGD (1024570). Minnesota: The study was funded by the National Institute of Mental Health (K23MH090421), the National Alliance for Research on Schizophrenia and Depression, the University of Minnesota Graduate School, the Minnesota Medical Foundation, and the Biotechnology Research Center (P41 RR008079 to the Center for Magnetic Resonance Research), University of Minnesota, and the Deborah E. Powell Center for Women's Health Seed Grant, University of Minnesota. Modal dilemma: study was supported by the Brain and Behavior Research Foundation and by the National Health and Medical Research Council ID 1125504 to SLW. Munster: This work was funded by the German Research Foundation (SFB-TRR58, Project C09 to UD). NESDA: study was supported by the Brain and Behavior Research Foundation and by the National Health and Medical Research Council ID 1125504 to SLW. QTIM: The QTIM dataset was supported by the Australian National Health and Medical Research Council (Project Grants No. 496682 and 1009064) and US National Institute of Child Health and Human Development (RO1HD050735). Singapore: KS was supported by National Healthcare Group, Singapore (SIG/15012) for the project. Standford: IHG is supported in part by National Institutes of Health (R37MH101495). MDS is supported by the National Institute of Mental Health (Project Number R01MH125850), Dimension Giving Fund, Ad Astra Chandaria Foundation, Brain and Behavior Research Foundation (Grant Number 28972), BIAL Foundation (Grant Number 099/2020), Emergence Benefactors, The Ride for Mental Health, and Gatto Foundation. USFC, TIGER: UCSF Weill

Institute for Neurosciences Weill Award for Investigators in the Neurosciences Impacted by COVID-19 Setbacks to TTY; by the National Center for Complementary and Integrative Health (NCCIH) R21AT009173, R61AT009864, and 4R33AT009864-03 to TTY; by the National Center for Advancing Translational Sciences (CTSI), National Institutes of Health, through UCSF-CTSI UL1TR001872 to TTY; by the American Foundation for Suicide Prevention (AFSP) SRG-1-141-18 to TTY; by UCSF Research Evaluation and Allocation Committee (REAC) and J. Jacobson Fund to TTY; by the National Institute of Mental Health (NIMH) R01MH085734 and the Brain and Behavior Research Foundation (formerly NARSAD) to TTY. Support for the TIGER study includes the Klingenstein Third Generation Foundation the National Institute of Mental Health K01MH117442 the Stanford Maternal Child Health Research Institute and the Stanford Center for Cognitive and Neurobiological Imaging TCH receives partial support from the Ray and Dagmar Dolby Family Fund. TCH is supported in part by the National Institute of Mental Health (K01MH117442), Klingenstein Third Generation Foundation, Stanford Maternal Child Research Institute (Early Career Award, and K Support Award), UCSF Research Evaluation and Allocation Committee (REAC), Raschen-Tiedeman Fund, and Moffitt Memorial Fund

Supplementary Materials

Supplementary Table 1: ENIGMA MDD Instrument for diagnosing major depressive disorder and exclusion criteria by site

Cohort	Diagnosis measurment	Sample characteristics/Inclusion criteria	Exclusion criteria
AFFDIS	ICD-10/DSM-IV criteria	MDD subjects currently depressed and in day program or inpatient	All subjects exclusion criteria: current or history of neurological disorder or brain injury, current substance abuse or dependence (not including nicotine), pregnancy, MRI contraindications, inability to give consent. MDD specific: comorbid psychiatric diagnosis. Healthy control specific: current or history of psychiatric diagnosis.
Pharmo (AMC)	MINI Plus	48 subjects with lifetime diagnosis of either MDD and/or AD and 14 healthy controls. Patients were startified depending on exposure to SSRIs: early (before age 23) or late (after age 23) exposure to SSRI's, or no exposure at all (UN). 15 subjects were diagnosed with only MDD, 3 with only AD and 22 with both MDD and AD (8 subjects did not receive a diagnosis due to incomplete M.I.N.I. Plus assessment). According to the M.I.N.I. Plus, none of the HC subjects were ever diagnosed with MDD or AD	Less than three week medication-free interval before scanning, current psychotropic medication use, a history of chronic or neurological disorder, family history of sudden heart failure or epileptic attacks, pregnancy (tested via urine sampling prior to the assessment), breast feeding, alcohol dependence and contra-indications for an MRI scan (e.g., ferromagnetic fragments). Participants agreed to abstain from smoking, caffeine and alcohol use for 24 hours prior to the assessments.
Barcelona	DSM-IV-TR acc. to CIDI-interview and HAMD	Outpatients with MDD diagnosis (DSM-IV-TR), with a first episode, recurrent MDD or chronic MDD (TRD) age 18-65	The exclusion criteria for healthy participants were: lifetime psychiatric diagnoses, first-degree relatives with psychiatric diagnoses and clinically significant physical or neurological illnesses. Axis I comorbidity according to DSM-IV-TR criteria was an exclusion criteria for all participants.

Cardiff	Hamilton Depression Rating Scale (HDRS-17)	N= 40, MDD patients with a current moderate to severe depressive episode despite minimum three months of stable antidepressant treatment	Psychotic symptoms, current substance dependence, eating disorders, claustrophobia and other MRI contraindications, and ongoing non-pharmacological treatment.
CSAN (Adf)	MINI	Current MDD: Meets MINI criteria for depression; comorbid anxiety disorders are allowed; mood-congruent psychotic symptoms allowed.	Current MDD: a current DSM-5 diagnosis of substance use disorder, except nicotine; a psychotic disorder, except depression with mood-congruent psychotic features; new antidepressant medication during the month before study participation (two months for fluoxetine); change of the dose of psychotropic medications over the last month (antidepressant and antipsychotic medication) or the last two months (mood stabilizers and anticonvulsants).
Calgary	KSADS	First episode MDD and healthy controls (Dalhousie sample). Recurrent MDD and healthy controls, recruited via referral from clinicians in Calgary, Alberta and through advertisements in local clinics and at the University of Calgary (Calgary sample).	Dalhousie Sample: A history of neurological illness, medical illness, claustrophobia, >21 year of age, or the presence of a ferrous implant or pacemaker. University of Calgary: Left handed; history of seizures, epilepsy or other neurological or psychiatric diagnoses (specifically bipolar disorder, psychosis, pervasive developmental disorder, eating disorders, PTSD); pregnancy
DCHS	MINI	Women over the age of 18 years, who were between 20 and 28 weeks pregnant, who presented at either of the two recruitment clinics, and who had no intention of moving out of the area within the following year, and were able to give written consent	1) loss of consciousness longer than 30 minutes, 2) inability to speak English, 3) current/lifetime alcohol and/or substance dependence or abuse, 4) psychopathology other than PTSD and/or MDD, 5) traumatic brain injury, 6) standard MRI exclusion criteria
FIDMAG	DSM-IV-TR criteria	MDD patients within a current depressive episode (HDRS >= 17, only 1 patient was in remission), right-handed, age 18-65	Patients were excluded (i) if they were left-handed; (ii) if they were younger than 18 or older than 65 years; (iii) if they had a history of brain trauma or neurological disease; (iv) if they had shown alcohol/ substance abuse within 12 months prior to participation; and (v) if they had undergone electroconvulsive therapy in the previous 12 months.
FOR2107Marburg	SCID-1	Participants recruited by means of public advertisement and from the inpatient services. Inclusion criteria: age 18-65 years; patients were diagnosed with major depressive disorder by SCID-Interview, currently depressed or remitted.	Exclusion criteria all: any MRI contraindications; any neurological abnormalities. Exclusion criteria controls: any current or former psychiatric disorder; Exclusion criteria patients: substance dependence or current benzodiazepine treatment (wash out of at least three half-lives before study participation)"
FOR2107Munster	SCID-1	Participants recruited by means of public advertisement and from the inpatient services. Inclusion criteria: age 18-65 years; patients were diagnosed with major depressive disorder by SCID-Interview, currently depressed or remitted.	Exclusion criteria all: any MRI contraindications; any neurological abnormalities. Exclusion criteria controls: any current or former psychiatric disorder; Exclusion criteria patients: substance dependence or current benzodiazepine treatment (wash out of at least three half-lives before study participation)"
Houston	SCID interview	Outpatients	MDD subjects: age below 18; lifetime or current diagnosis of psychotic disorder, or bipolar I or II disorder; substance abuse/dependence in 6 months prior to study inclusion; current major medical problems. Control subjects: age below 18; current major medical problems; current psychiatric or neurologic disorder; history of psychiatric disorders in a first-degree relative; current major medical problems. Both groups: MRI contra-indications
Hiroshima	MINI	MDD Patients were recruited from local clinics, 20-80 years. Controls were recruited from local community by advertising in local papers.	MDD patients: comorbid psychiatric disorders other than MDD, Control subjects: any history of psychiatric disorder
TiPs (Jena, Germany)	SCID interview	Psychiatric inpatients and tinnitus patients with MDD or a disorder of the depressive spectrum (also adjustment disorders as pointed out in the data table); psychiatrically healthy controls were derived from community and tinnitus patients	MDD subjects: presence of axis-I disorders other than MDD or adjustment disorders. Control subjects: no Axis-I diagnosis, no medication use. Exclusion criteria for all subjects included history of neurological disease (e.g. tumour, head trauma, epilepsy) or untreated internal medical condtitions, intellectual and/or developmental disability.

			Only German native speakers were allowed to participate.
MODECT	MINI	Older adults, aged above 55, with severe depression admitted to be treated with ECT	Exclusion criteria were another major DSM-IV-TR diagnosis, such as schizophrenia, bipolar or schizoaffective disorder and a history of major neurological illness (including Parkinson's disease, stroke and dementia).
Melbourne	SCID interview	Youth depression sample: 15-25 years of age. Recruited as part of 2 large RCTs (incl. YoDA-C - Davey et al., 2014; Trials) and scanned prior to treatment randomisation. 60 patients unmedicated (YoDA-C).	MDD subjects: lifetime or current SCID-I diagnosis of psychotic disorder, or bipolar I or II disorder. Control subjects: any SCID-I diagnosis or medication use. Both groups: Acute or unstable medical disorder; general MRI contraindications
Minnesota	Schedule for Affective Disorders and Schizophrenia for School-Age Children–Present and Lifetime Version and the Children's Depression Rating Scale–Revised (CDRS-R).	Adolescents with MDD and HCs aged 12 to 19 years were recruited to participate through community postings and referrals from local mental health services. Adolescents with MDD were eligible if they had a primary diagnosis of MDD and had not received any psychotropic medication treatment for the past 2 months. Healthy adolescents were eligible if they had no current or past psychiatric diagnoses and were frequency matched to the MDD group on age and sex	Exclusion criteria for both groups included the presence of a neurologic or other chronic medical condition, mental retardation, pervasive developmental disorder, substance use disorder, bipolar disorder, or schizophrenia
MOODS / DEP-ARREST CLIN	MINI, DSM5	Patients aged 18-65 years with a current MDE diagnosis (MINI interview(Sheehan et al., 1998) and a minimum depression score of 18 on the Hamilton Depression Rating Scale-17 items (HDRS) in the context of MDD, as well as free of antidepressant drug use at least one month before the study beginning, were included. HCs were included based on the absence of current or past mental disorders or somatic conditions, particularly nasal polyposis and chronic or acute sinusitis or rhinitis	Patients suffering from bipolar disorder, psychotic disorder, eating disorder, and addictions, according to the DSM-5 criteria, or from nasal polyposis, chronic or acute sinusitis, chronic or acute rhinitis or pregnancy or breastfeeding, were not included. HCs were included based on the absence of current or past mental disorders or somatic conditions, particularly nasal polyposis and chronic or acute sinusitis or rhinitis
Moral Dilemma	SCID interview	Youth depression sample: 15-25 years of age; recruited from outpatient service. Controls recruited from general community.	MDD subjects: lifetime or current SCID-I diagnosis of psychotic disorder, or bipolar I or II disorder; current antidepressant medication use. Control subjects: any SCID-I diagnosis or medication use. Both groups: Acute or unstable medical disorder; general MRI contraindications
Munster	SCID interview	Participants recruited by means of public advertisement and from the inpatient services. Inclusion criteria: age 16-65 years; patients were diagnosed with major depressive disorder by SCID-Interview	MDD subjects: presence of bipolar disorder, schizoaffective disorders and schizophrenia; substancerelated disorders or current benzodiazepine treatment (wash out of at least three half-lives before study participation), and former electroconvulsive therapy. Control subjects: any current or former psychiatric disorder. Both groups: any neurological abnormalities, MRI contra-indications
NESDA	CIDI interview	DSM-4 based diagnosis of MDD (6 month recency), using CIDI interview. 93 (60%) MDD patients have a comorbid ANX diagnosis. Age range 18-65	N/A
QTIM	CIDI interview	Retrospective questionnaire about depression episodes combined with an MRI study. The best described MDD episode is defined as the worst one (according to individuals). We have up to 5 supplementary episodes (briefly) described. Sample composed of twins and relatives. Population-based sample	MDD subjects: presence of axis-I disorders other than MDD and anxiety disorders Control subjects: antidepressant use, psychiatric disorders All subjects: relatedness between subjects, left handedness, history of neurological or other severe medical illness, head injury or current or past diagnosis of substance abuse, use of cognition affecting medication and general MRI contraindications
San Francisco UCSF	KSADS (semi-structured interview based on DSM) for MDD, DISC/DPS for HCL	Outpatient/community-based sample with DSM diagnosis, mostly antidepressant-naive and approximately 60% of MDD have comorbid anxiety disorders	Exclusion criteria for all participants included: 1) use of pharmacotherapeutics for treating psychiatric conditions within the past 6 months, 2) misuse of drugs within two months prior to MRI scanning; 3) two or more alcoholic drinks per week within the previous month (as assessed by the Customary Drinking and Drug Use Record; CDDR) (Brown et al, 1998); 4) a full scale IQ score of less than 75 (as assessed

			by the Wechsler Abbreviated Scale of Intelligence; WASI) (Wechsler, 1999); 5) contraindications for MRI including ferromagnetic implants and claustrophobia; 6) pregnancy or the possibility of pregnancy; 7) left-handedness; 8) prepubertal status (as assessed as Tanner stages of 1 or 2) (Tanner, 1962); 9) inability to understand and comply with procedures; 10) neurological disorder (including meningitis, migraine, or HIV); 11) head trauma; 12) learning disability; 13) serious health problems; and 14) complicated or premature birth (i.e., birth before 33 weeks of gestation). The MDD group was subject to the additional exclusion criterion of a primary psychiatric diagnosis other than MDD. The HCL group was subject to the additional exclusion criteria of: 1) history of mood or psychotic disorders in a first- or second-degree relative (as assessed by the Family Interview for Genetics; FIGS) (Maxwell, 1992); and 2) current or lifetime DSM-IV-TR Axis I psychiatric disorder.
SHIP_START-2	M-CIDI interview	Population based longitudinal cohort study	MDD subjects: presence of axis-I disorders other than MDD, anxiety disorders, conversion, somatization and eating disorder. Control subjects: no lifetime diagnosis of depression, no antidepressiva, and severity index=0 All subjects: We removed subjects with medical conditions (e.g. a history of cerebral tumor, stroke, Parkinson's diseases, multiple sclerosis, epilepsy, hydrocephalus, enlarged ventricles, pathological lesions) or due to technical reasons (e.g. severe movement artifacts or inhomogeneity of the magnetic field).
SHIP_TREND-0	M-CIDI interview	Population based longitudinal cohort study	MDD subjects: no special exclusion criteria Control subjects: no lifetime diagnosis of depression, no antidepressiva, and severity index=0 All subjects: We removed subjects with due to medical conditions (e.g. a history of cerebral tumor, stroke, Parkinson's diseases, multiple sclerosis, epilepsy, hydrocephalus, enlarged ventricles, pathological lesions) or due to technical reasons (e.g. severe movement artifacts or inhomogeneity of the magnetic field).
Singapore	SCID interview	Inclusion: 1) DSM IV dx of MDD (Patients) 2) Age: 21-65 3) English speaking 4) Provision of informed written consent	Exclusion criteria 1) History of significant head injury 2)Neurological diseases such as epilepsy, cerebrovascular accident 3) Impaired thyroid function 4) Steroid use 5) DSM IV alcohol or substance use or dependence 6) Contraindications to MRI (e.g. pacemaker, orbital foreign body, recent surgery/procedure with metallic devices/implants deployed) using standard MRI Request Form from NNI 7)Pregnant women 8) Claustrophobia
SoCAT	SCID interview	Inclusion criteria: DSM IV dx for mdd patients Age: 18-65 right-handed currently depressed or remitted; Control subjects: any history of psychiatric disorder	Exclusion criteria 1) History of significant head injury 2)Neurological diseases such as epilepsy, cerebrovascular accident 3)Other diagnoses on Axis I disorders4)
Stanford FAA	SCID interview	Community-based DSM-diagnosed sample	MDD subjects: presence of axis-I disorders other than MDD, anxiety and eating disorders . Control subjects: control individuals did not meet diagnostic criteria for any current psychiatric. Both groups: alcohol / substance abuse or dependence within six months prior to MRI scanning, history of head trauma with loss of consciousness > 5 min, aneurysm, or any neurological or metabolic disorders that require ongoing medication or that may affect the central nervous system (including thyroid disease, diabetes, epilepsy or other seizures, or multiple sclerosis), MRI contraindications, or bad MRI data (e.g., extreme movement).
Stanford T1w	SCID interview	Community-based DSM-diagnosed sample	MDD subjects: presence of axis-I disorders other than MDD, anxiety and eating disorders .

Aggregate			Control subjects: control individuals did not meet diagnostic criteria for any current psychiatric. Both groups: alcohol / substance abuse or dependence within six months prior to MRI scanning, history of head trauma with loss of consciousness > 5 min, aneurysm, or any neurological or metabolic disorders that require ongoing medication or that may affect the central nervous system (including thyroid disease, diabetes, epilepsy or other seizures, or multiple sclerosis), MRI contraindications, or bad MRI data (e.g., extreme movement).
TAD			
TIGER	KSADS	Community-based DSM-diagnosed sample	All subjects: Exclusion criteria were premenarchal status (for females), history of concussion within the past 6 weeks or history of any lifetime concussion with loss of consciousness, contraindications to MRI scanning (e.g. braces, metal implants, or claustrophobia), serious neurological or intellectual disorders that could interfere with the participant's ability to complete study components. MDD subjects: meeting lifetime or current DSM-IV criteria for any Bipolar Disorder, Psychosis, or Alcohol Dependence, or DSM-5 criteria for Moderate Substance Use Disorder with substance-specific threshold for withdrawal. CTL subjects: any current or past DSM-IV Axis I Disorder and first-degree relative with confirmed or suspected history of depression, mania, psychosis, or substance dependence.

Supplementary Table 2: ENIGMA MDD Image acquisition and processing by cohort

Cohort	Scanner type	Sequence T1	FreeSurfer version	Slice orientation	Operating system
AFFDIS	3T Siemens Magnetom TrioTim	3D T1 (176 slices; TR = 2250 ms; TE = 3.26 ms; FOV 256; voxel size 1X1X1mm)	5,3	Sagittal	Linux CentOS
Barcelona	3T Philips Achieva	3D MPRAGE images (Whole-brain T1-weighted); TR=6.7ms, TE=3.2ms; 170 slices, voxel size 0.89X0.89X1.2 mm. Image dimensions 288X288X170; field of view: 256X256X204; slice thickness: 1.2 mm; with a sagittal slice orientation, T1 contrast enhancement, flip angle: 8º, grey matter as a reference tissue, ACQ matrix MXP = 256X240 and turbo-field echo shots (TFE) = 218.	6	Sagittal	Scientific Linux 5
Cardiff	A 3 Tesla whole body MRI system (General Electric, Milwaukee, USA) with an 8-channel head coil was used at the Cardiff University Brain Research Imaging Centre (CUBRIC).	High-resolution anatomical scan (Fast Spoiled Gradient-Recalled-Echo [FSPGR] sequence): 178 slices, TE=3 ms, TR=7.9 ms, voxel size=1.0×1.0×1.0 mm3, FA=15°, FOV=256×256	5,3		freesurfer-Linux-centos6_x86_64-stable-pub-v5.3.0

CSAN (Adf)	3T Siemens MAGNETOM PRISMA	Whole-head t1-weighted MPRAGE (TR = 2300 ms, TE = 2.34 ms, FOV 250 × 250 mm, voxel size = 0.9 × 0.868 × 0.868 mm, flip angle = 8°)	7.2	Sagittal	Ubuntu
Calgary	1.5T Siemens Magnetom Vision. 3T GE Discovery MR750	1.5T: A sagittal scout series was acquired to test image quality. 3D fast low angle shot (FLASH) sequence was used to acquire data from 124 1.5 mm-thick contiguous coronal slices through the entire brain (echo time = 5ms, repetition time = 25ms, acquisition matrix = 256 x 256 pixels, field of view = 24 cm and flip angle = 40°). 3T: Anatomical imaging acquisition parameters: axial acquisition, repetition time (TR), 2200 milliseconds (ms); echo time (TE), 3.04 ms; TI, 766, 780; flip angle, 13 degrees; 208 partitions; 256 × 256 matrix; and field of view, 256.	5,3	Dalhousie sample, coronal; Calgary sample, axial	MacOs Sierra
DCHS	3T Siemens Skyra	3D multi-echo MPRAGE, voxel size 1 mm x 1mm x 1.5mm, TR = 2530 ms, TE = 1.69 x 3.55 x 5.41 x 7.27ms, FOV: 256x256mm, flip angle = 7°	5.3	Sagittal	Linux-centos6_x86_64
FOR2107 - Marbourg	3T Siemens Magnetom TiroTim syngo MR B17	Sequence: 3D T1-weighted magnetization prepared rapid acquisition gradient echo (MPRAGE) - Sagittal Acquisition Direction, # of Slices 176, 0.5mm Slice Gap, 1.0x1.0x1.0 Voxel Size (mm3), TI 900 ms, TE 2.26 ms, TR 1900 ms, Flip Angle 9.	5,3	Sagittal	Red Hat Enterprise Linux Server release 5.11 (Tikanga)
FOR2017 - Münster	3T Philips	3D T1-weighted scan (170 slices; TR = 9ms; TE = 3.6ms; 256x231 matrix of 1×1×1 mm voxels)	5,3	Sagittal	Red Hat Enterprise Linux Server release 5.11 (Tikanga)
FIDMAG	1.5T, GE Signa	3D T1: matrix size = 512 × 512, 180 contiguous axial slices, voxel resolution = 0.47 × 0.47 × 1mm, no slice gap, TE = 3.93ms, TR = 2000ms and inversion time (TI) = 710ms, flip angle = 15 degrees	6	Axial	Linux-centos6_x86_64
Houston	subjects in 20000s: 1.5 T Philips Medical Systems Gyroscan Intera; subjects in 30000s: 3T Siemens Allegra	Subjects in the 20000s: Fast field echo sequence- repetition time (TR) = 24 ms, echo time (TE) = 4.99 ms, flip angle = 40°, slice thickness = 1 mm, matrix size = 256 × 256 and 150 slices. Subjects in 30000s: MPRAGE- repetition time (TR) = 1750 ms, echo time (TE) = 4.39 ms, flip angle = 8°, slice thickness = 1 mm, matrix size = 208 × 256 and 160 slices.	5,3	Subjects in 20000s: Sagittal; Subjects in 30000s: Transverse	Fedora 19
Hiroshima	3T Siemens (Spectra, Verio.Dot), 3T GE (Signa HDxt) Site 1 = GE Signa HDxt 3.0T 2= GE Signa HDxt 3.0T 3 = SIEMENS MAGNETOM Spectra 3.0T 4 = SIEMENS MAGNETOM Verio.Dot 3.0T	T1 256x256x256 matrix of 1x1x1mm voxels (Siemens: ADNI MPRAGE (tfl), GRAPPA, 192 slices, GE: SPGR, 184 slices) *Detailed scanning parameter sheets are available for all 4 scanners on request.	5,3	Sagittal	Linux_Ubuntu_18.04
TiPs (Jena, Germany)	3T Siemens MAGNETOM Prisma_fit	MPRAGE sequence: TR 2300 ms, TE 3.03 ms, α 9°, 192 contiguous sagittal slices, in-plane field of view 256 mm, voxel resolution 1Å~1Å~1 mm; acquisition time 5:21 min	5,3	Sagittal	Linux

MODECT	3T (General Electric Signa HDxt, Milwaukee, WI, USA)	T1-weigthed data set was acquired (flip angle=12°, repetition time=7.84 milliseconds, echo time=3.02 milliseconds; matrix 256x256, voxel size 0.94x0.94x1 mm; 180 slices).	5,3	Coronal	Linux
Melbourne	3T GE Signa Excite	3D BRAVO sequence 140; TR=7900 ms; TE=3000 ms; flip angle=13º; FOV=256 mm; matrix=256 x 256	5,3	Axial	Linux Debian x86 64
Minnesota	3.0 Tesla Tim Trio scanner; Siemens Corp	A 5-minute structural scan was acquired using a T1-weighted, high-resolution, magnetization-prepared gradient-echo sequence: repetition time, 2530 milliseconds; echo time, 3.65 milliseconds; inversion time, 1100 milliseconds; flip angle, 7°; field of view, 256 × 176 mm; voxel size, 1-mm isotropic; 224 slices; and generalized, autocalibrating, partially parallel acquisition acceleration factor, 2.	5,3	Coronal	Linux
MOODS / DEP-ARREST CLIN	3T Philips Achieva	3D T1-weighted image: TR=7, TE=3.5, FOV=352x352x180, Flip angle=8 degrees, number of slices : 180 slices, Slice gap 1 mm, voxel size: 0.8x0.8x1	6	Transverse (Axial)	CentOS Linux 7
Moral Dilemma	3T GE Signa Excite	3D BRAVO sequence: 140 contiguous slices; repetition time, 7900 ms; echo time, 3000 ms; flip angle, 13°; in a 25.6-cm field of view, with a 256 × 256 pixel matrix and a slice thickness of 1 mm (1 mm gap).	5,3	Axial	Linux Debian x86 64
Munster	3T Philips Gyroscan Intera	3D fast gradient echo sequence (turbo field echo), repetition time = 7.4 milliseconds, echo time = 3.4 milliseconds, flip angle = 9°, two signal averages, inversion prepulse every 814.5 milliseconds, acquired over a field of view of 256 (feet -head [FH]) × 204 (anterior -posterior [AP]) × 160 (right -left [RL]) mm, phase encoding in AP and RL direction, reconstructed to cubic voxels of .5 mm × .5 mm × .5 mm	5,3	Sagittal	Red Hat Enterprise Linux Server release 5.11 (Tikanga)
NESDA	3T Phillips Achieva/Inter a	3D gradient-echo T1-weighted sequence. TR=9 msec; TE=3.5 msec; flip angle 8º, FOV = 256 mm; matrix: 25x62x56; in plane voxel size = 1 mm × 1 mm x 1 mm; 170 slices.	5	Sagittal	SHARK HPC, Linux environment
QTIM	Bruker 4T Wholebody MRI	3D T1 weighted sequence. TR=1500 msec; TE=3.35 msec; flip angle=8°, 256 or 240 (coronal or sagittal) slices, FOV=240 mm, matrix 256x256x256 (or 256x256x240)	5,1	Coronal, then sagittal following software upgrade.	Linux-centos4_x86_64-stable-pub-v5.1.0
San Francisco UCSF	3T GE Discovery MR750	SPGR T1-weighted: TR=8.1 ms; TE=3.17 ms; TI=450 ms; flip angle=12°; 256x256 matrix; FOV=250x250 mm; 168 sagittal slices; slice thickness=1 mm; in-plane resolution=0.98x 0.98 mm	5,3	Sagittal	Linux-centos6_x86_64-stable-pub-v5.3.0.
SHIP _START-2	1.5T Siemens Avanto	3D T1-weighted (MP-RAGE/ axial plane); TR=1900 msec; TE=3.4 msec; Flip angle=15°; voxel size 1 mm x 1 mm x 1 mm	5.3 (cortical), 5.1 (subcortical)	Axial	Centos6_x86_64
SHIP-_TREND-0	1.5T Siemens Avanto	3D T1-weighted (MP-RAGE/ axial plane); TR=1900 msec; TE=3.4 msec; Flip angle=15°; voxel size 1 mm x 1 mm x 1 mm	5.3 (cortical), 5.1 (subcortical)	Axial	Centos6_x86_64
Singapore	Achieva 3T, Philips Medical Systems, Netherlands	Whole brain high resolution 3D MP-RAGE (magnetisation-prepared rapid acquisition with a gradient echo) volumetric scans (TR/TE/TI/flip angle 8.4/3.8/3000/8; matrix 256x204; FOV 240mm2) with axial orientation (reformatted to coronal)	5,3	Axial	Linux_Ubuntu12.04_6 4
SoCAT	3.0 T, Siemens Verio,Numaris/4,Syngo MR	3D T1 weighted MP-Rage/axial plane; TR=1900 msec; TE=3.4 msec; Flip angle=15°; Voxel size 1 mm x 1 mm x 1 mm	5,3	Axial	Ubuntu 18.04 LTS

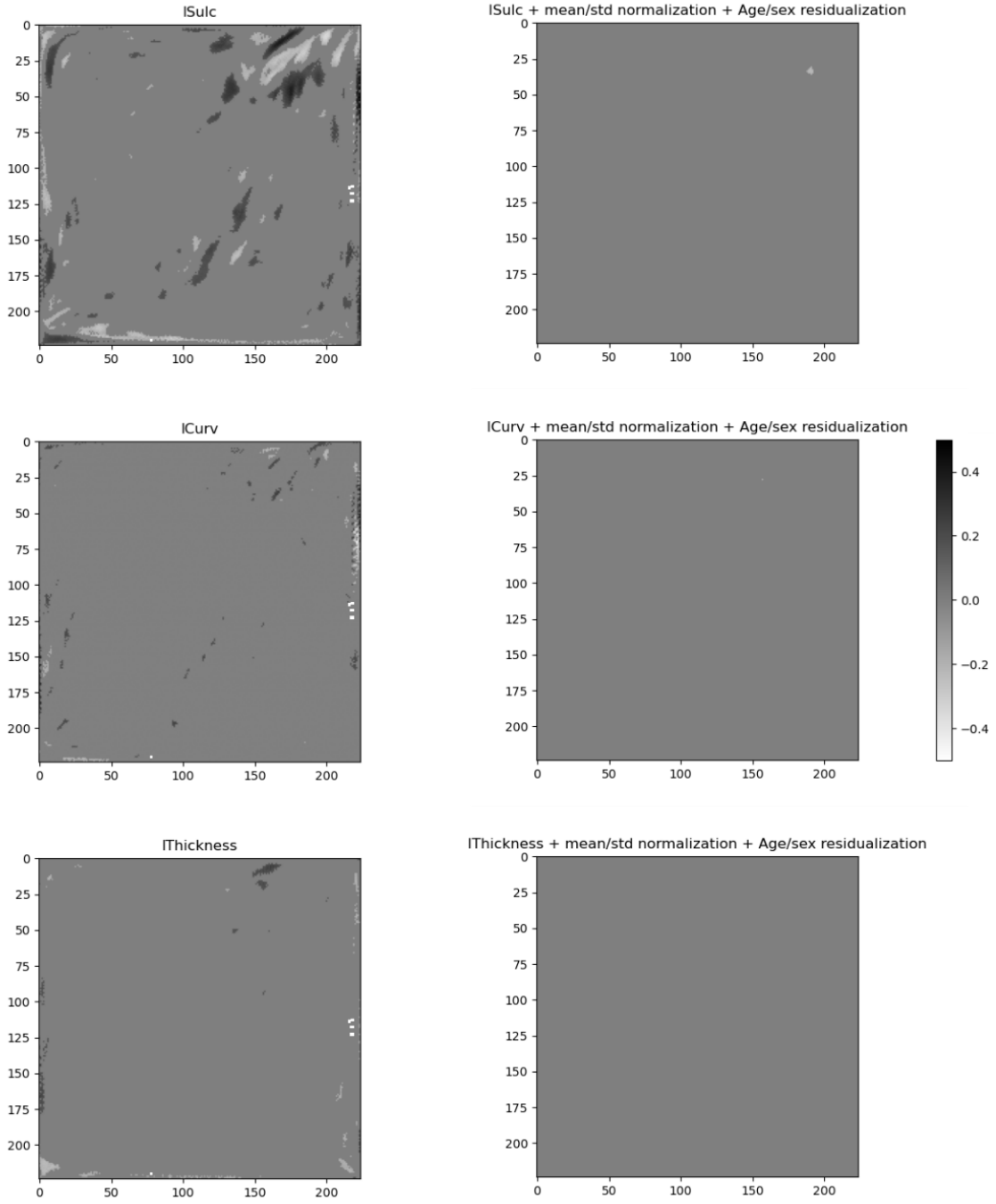
	B17,Erlangen ,Germany				
Stanford FAA	3.0T GE Discovery MR750	Whole-brain T1-weighted images were collected using a spoiled gradient echo (SPGR) pulse sequence (186 sagittal slices; resolution = 0.9 mm isotropic; flip angle = 12°; repetition time [TR] = 6,240 ms; echo time [TE] = 2.34 ms)	5,3	Sagittal	Linux-centos6_x86_64
Stanford T1w Aggregate	1.5T GE Signa Excite	Whole-brain T1-weighted images were collected using a spoiled gradient echo (SPGR) pulse sequence (116 sagittal slices; through-plane resolution = 1.5 mm; in-plane resolution = 0.86 x 0.86 mm; flip angle = 15 degrees; repetition time [TR] = 8.3-10.1 ms; echo time [TE] = 1.7-3.0; inversion time [TI] = 300 ms; matrix = 256 x 192).	5,3	Sagittal	Centos6_x86_64, Linux-based HPC
TAD					
TIGER	3T GE MR750	TR/TE/TI=8.2/3.2/600 ms; flip angle=12°; 156 axial slices; FOV=25.6 cm; matrix=256 mm x 256 mm, isotropic voxel=1 mm, total scan time: 3:40	6	Axial	Linux

Supplementary Table 3: List of hyperparameters of trained algorithms. Optimal hyperparameters were found by the grid search during the sex classification task. We followed a heuristic approach outlined in [275] to determine a range of values for C.

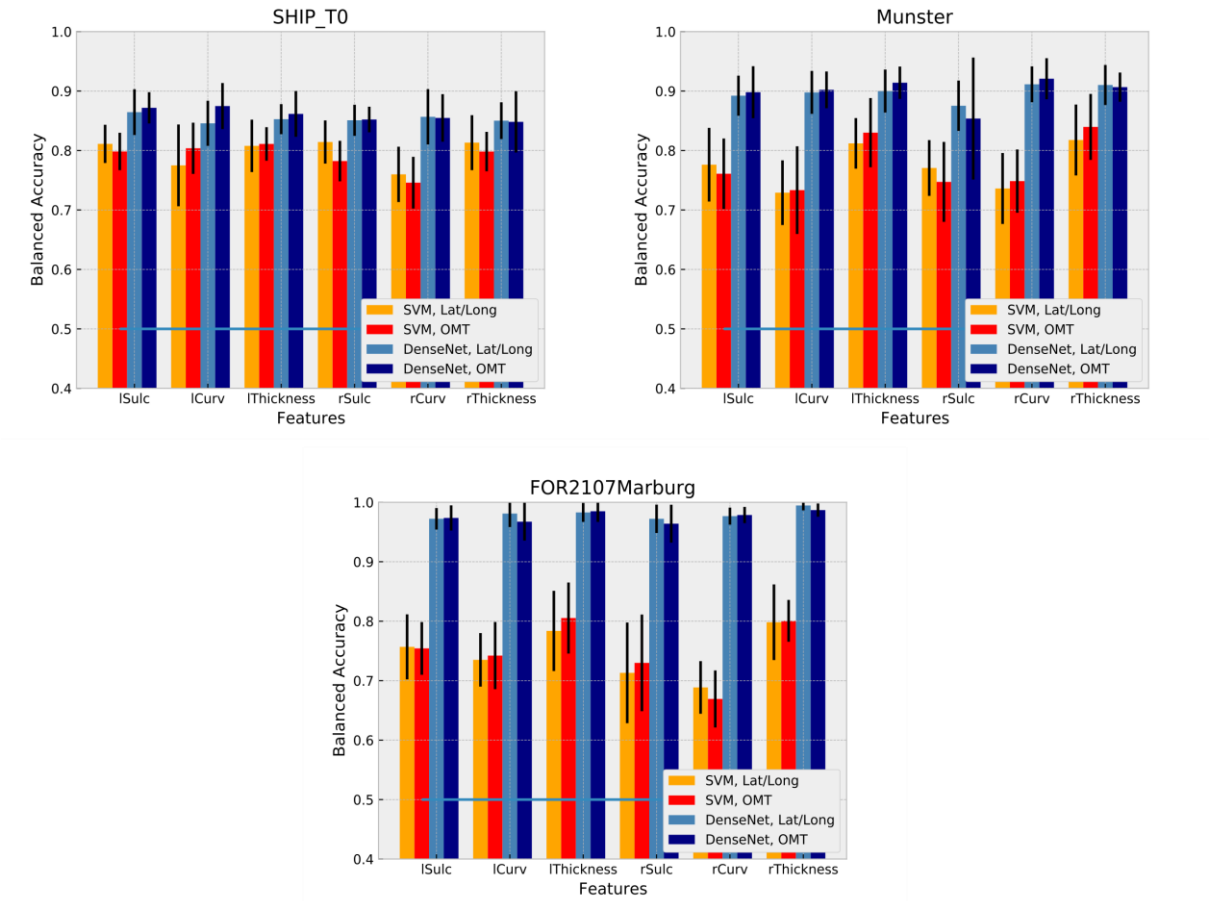
Classification algorithm	Feature Selection	Hyperparameters	Nested CV
SVM Linear	None	C = $[10^{-4}, 10^{-3}, \ldots, 10^{4}]$	10 fold
DenseNet	None	Number of dense layers = [1,2,3] Number of nodes in the dense layers = [10,100,200] Adam optimizer: learning rate [.01,.001,.0001] DropOut layer before dense layers (yes, no)	10 fold

Supplementary Table 4: MDD vs HC classification, hemisphere analysis.

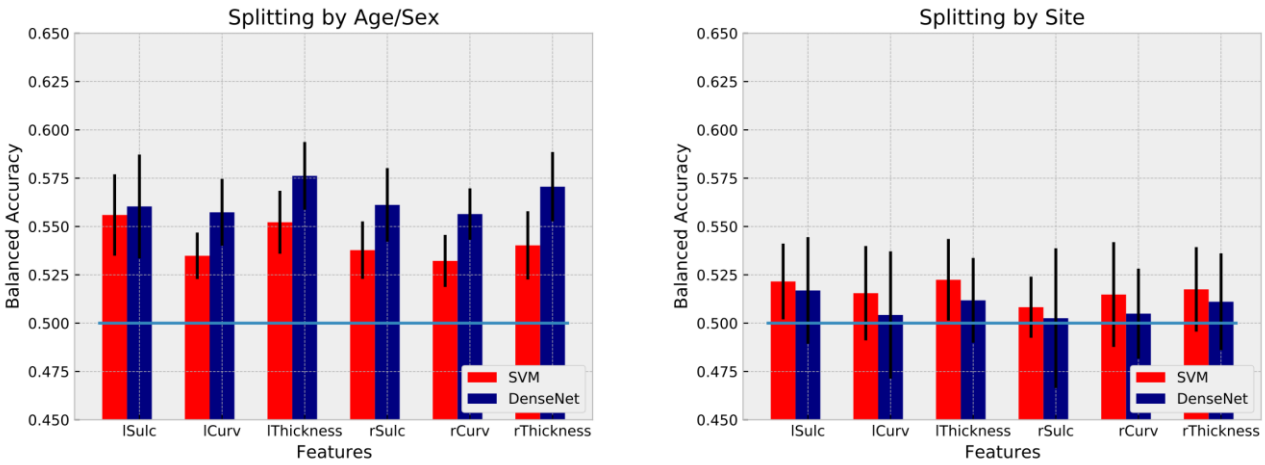
	Splitting by Age/Sex		Splitting by Site	
Hemisphere	Left	Right	Left	Right
SVM	0.546 ± 0.022	0.539 ± 0.017	0.514 ± 0.034	0.522 ± 0.036
DenseNet	0.569 ± 0.019	0.556 ± 0.024	0.513 ± 0.018	0.506 ± 0.017



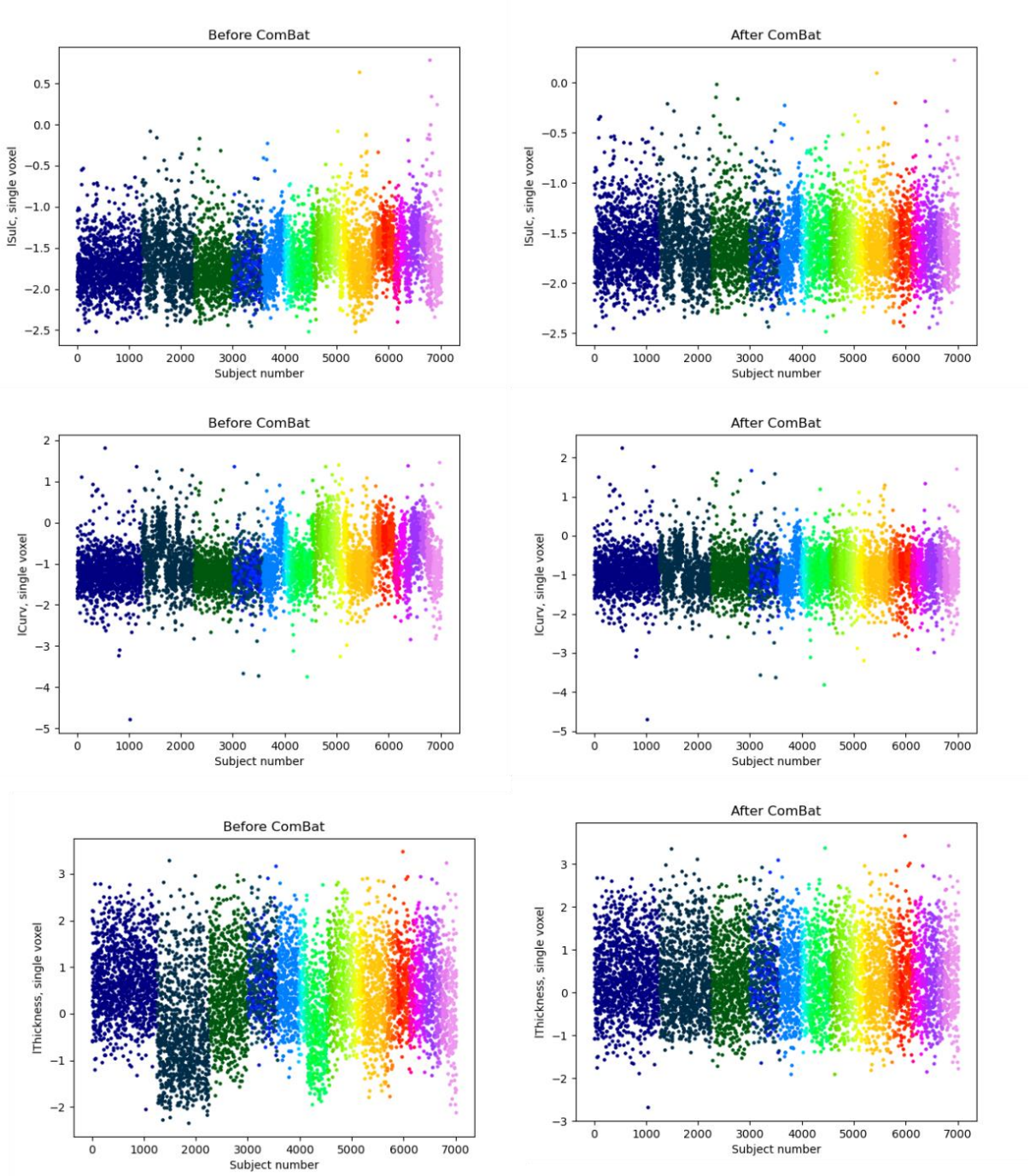
Supplementary Figure 1: ICV dependence stored in the cortical maps. To investigate if ICV is preserved in cortical features, we correlated ICV values with the pixels' intensities from SHIP_TREND_0 healthy controls (left). After standardizing the features to the mean of 0 and standard deviation of 1, and regressing out Age/Sex covariates, we effectively removed the effect of ICV from the features (right). Colormap represents the direction of the significant correlations.



Supplementary Figure 2: Sex classification. We estimated classification performance via balanced accuracy of SVM and DenseNet on three biggest cohorts: SHIP_TREND-0 (top left), Munster (top right) and FOR2107Marburg (bottom) for all features separately using 1) Latitude/Longitude projection and 2) OMT projection.



Supplementary Figure 3: MDD vs HC classification. We estimated classification performance via balanced accuracy of SVM and DenseNet for each hemisphere and feature type separately.



Supplementary Figure 4: Examples of ComBat harmonization for all data modalities. Color corresponds to the site affiliation.

Chapter 5 Subject-specific whole-brain parcellations of nodes and boundaries are modulated differently under 10Hz rTMS

Subject-specific whole-brain parcellations of nodes and boundaries are modulated differently under 10Hz rTMS

Vladimir Belov[1,#], Vladislav Kozyrev[1,2,3#], Aditya Singh[1], Matthew D. Sacchet[4], Roberto Goya-Maldonado[1,*]

Affiliations:

[1] Laboratory of Systems Neuroscience and Imaging in Psychiatry (SNIP-Lab), Department of Psychiatry and Psychotherapy, University Medical Center Göttingen (UMG), Göttingen, Germany

[2] Functional Imaging Laboratory, German Primate Center – Leibniz Institute for Primate Research, Göttingen, Germany

[3] Institute of Molecular and Clinical Ophthalmology Basel, Basel, Switzerland

[4] Meditation Research Program, Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Running title: rTMS modulates nodes and boundaries

 # equal contribution

***Corresponding author:**
PD Dr. Roberto Goya-Maldonado
Laboratory of Systems Neuroscience and Imaging in Psychiatry (SNIP-Lab)
Department of Psychiatry and Psychotherapy
University Medical Center Göttingen (UMG)
Von-Siebold Str. 5, 37075 Göttingen
e-mail: roberto.goya@med.uni-goettingen.de

Contributions

Conceptualization: RGM; Methodology: VB, VK; Data curation: AS; Investigation: RGM; Supervision: RGM; Writing (original draft): RGM, VB, VK; Writing (review and editing): RGM, MDS. Data and materials availability: Raw data analyzed during the current study are not publicly available due to absence of written consent from the participants of the study. Scripts and codes used in the analyses will be deposited in a public database.

Abstract

Repetitive transcranial magnetic stimulation (rTMS) has gained considerable importance in the treatment of neuropsychiatric disorders, including major depression. However, it is not yet understood how rTMS alters brain's functional connectivity. Here we report changes in functional connectivity captured by resting state functional magnetic resonance imaging (rsfMRI) within the first hour after 10Hz rTMS. We apply subject-specific parcellation schemes to detect changes (1) in network nodes, where the strongest functional connectivity of regions is observed, and (2) in network boundaries, where functional transitions between regions occur. We use support vector machines (SVM), a widely used machine learning algorithm that is robust and effective for the classification and characterization of time intervals of changes in node and boundary maps. Our results reveal that changes in connectivity at the boundaries are slower and more complex than in those observed in the nodes, but of similar magnitude according to accuracy confidence intervals. These results were strongest in the posterior cingulate cortex and precuneus. As network boundaries are indeed under-investigated in

comparison to nodes in connectomics research, our results highlight their contribution to functional adjustments to rTMS.

1. Introduction

Repetitive transcranial magnetic stimulation (rTMS) has become a popular method for the non-invasive modulation of brain function [276]. Recent neuroimaging studies have shown that functional changes induced by rTMS in a localized cortical region lead to selective and distinct modulation of activity and functional connectivity both within and between large-scale brain networks [277]–[282]. The mechanisms by which rTMS induces network modulations are still not well understood. Today, mapping whole-brain effects caused by local neural perturbations, including by rTMS, is a growing field of research. Well-established methods now allow for the assessment of connectome-level functional adjustments to high frequency rTMS in both node and boundary maps in sequential time intervals [283], [284].

Functional magnetic resonance imaging (fMRI) data obtained while participants are not engaged in any specific task is called resting state fMRI (rsfMRI). RsfMRI has been instrumental for advancing our understanding of the brain's macroscopic functional network architecture [285]–[287] as well as which regions might be most functionally altered in psychiatric disorders [288], [289]. However, fMRI data typically consists of functional time-courses in thousands of voxels, which on the one hand allows for accurate inference of correlations or "functional connectivities" between regions, but on the other hand has high dimensionality. Different approaches have been proposed to reduce data dimensionality of the data and to identify the most relevant patterns of spatiotemporal organization in fMRI data. This is the case for whole-brain functional regions that will be represented in our study as nodes [290] and boundaries [283], [291], [292]. Nodes are defined as the greatest strength of local or global connectivity, also known as concepts of modularity and integration respectively, which enabled many insights into dimensional organization of the healthy and diseased brain [293]. Boundaries are the counterparts of the nodes, identified where the connectivity strength is the lowest or absent, usually in the transition between neighboring functional regions [291]. In contrast to the investigation of boundaries, the scientific community has given disproportionate

attention to the nodes of functional networks. In node clustering approaches, spatiotemporal elements (i.e., voxels) may be grouped on the basis of the similarity versus dissimilarity of their functional connectivity [290]. An example of a node clustering approach is independent component analysis (ICA), which is used as a brain mapping method that efficiently segregates functional components based on their corresponding spatiotemporal distribution [294]. ICA has been widely applied to the identification of large-scale brain networks [280], [295], [296].

Group-based parcellation schemes use fMRI data from multiple individuals to map large-scale functional brain networks, that is, collections of widespread regions showing functional connectivity [292], [297]–[299]. While group-based parcellation captures major features of functional brain organization that are evident across individuals, such approaches may obscure certain person-specific features of brain organization. In contrast, subject-specific parcellation methods have been shown to effectively map aspects of functional organization that differ for particular individuals (e.g. Kong et al., 2019; Saxe et al., 2006). Several recent studies have demonstrated that extensive rsfMRI data collected across multiple sessions from the same individual can be used to delineate high-quality cortical parcellations at the individual level [301]–[303]. Subject-specific parcellation may enable increasingly precise planning and delivery of rTMS interventions.

Machine learning includes the development of algorithms that can detect spatially complex and often subtle patterns in highly dimensional data, and has been applied to neuroimaging. Machine learning may thus be useful for individual-level predictions that could ultimately be used in clinical contexts [158], [161], [304], [305]. The support vector machine (SVM) is a machine learning algorithm that constructs hyperplanes in multidimensional space to optimally separate data classes [306]. Taken together, in the current study SVM was used to identify the strongest rTMS-related changes in brain functional connectivity patterns in both nodes and boundary maps. SVM is one of the most commonly used machine learning algorithms due to its robustness and easy interpretability. A particular advantage of SVM in this study is that it often yields better classification in smaller datasets, that is, datasets in which the number of features greatly exceeds the number of training data samples [307]. As a commonly applied resource, the classification task can be simplified by using unbiased feature selection approaches. Features selection involves the identification of the most useful data features in the training dataset [308], which are then solely used for classification. Features selection has been shown to improve accuracy while also increasing the interpretability of identified multivariate models [309], [310].

Recent neuroimaging and predictive modeling findings suggest that a locally generated brain stimulation-induced perturbation in neural activity is gradually integrated by selective alterations of within- and between-network dynamics [277], [311]. In addition, animal studies have shown that 10 Hz rTMS creates a transient cortical functional state that is characterized by increased excitability and increased response variability [312], [313] 10 Hz rTMS applied to cat visual cortex resulted in a reduction of the inhibitory notch commonly seen in visual evoked activity, evidencing decreased inhibition during visual processing [313]. On the other hand, the findings by [312] implicate a reduction of specificity (decorrelation) close to the borders of the functionally distinct regions reflected in the widening of boundaries between them. This is plausibly happening due to rTMS-induced reduction of inhibition, predominantly in the boundaries but also in the nodes.

In the current project, we endeavored to understand rTMS modulation of whole-brain connectivity patterns. For that, we used individual-level comparisons to create sham-corrected maps for nodes and boundaries. This approach enhanced the sensitivity of detecting individual-level rTMS-induced variations in functional connectivity. Our SVM approach allowed for the data-driven identification of the most substantial functional changes as a whole and pairwise across time conditions (time intervals R). Based on the prior studies described above, we hypothesized that 10 Hz rTMS would affect functional connectivity both in the nodes and in the boundaries of distant regions that interact with the DLPFC.

2. Material and methods

2.1 Participants and study design

23 healthy subjects between the ages of 18-65 were recruited from a university environment to participate in a double-blind, sham-controlled, crossover design study that investigated the neural effects of 10 Hz rTMS using rsfMRI. Further details on the study design have been reported elsewhere [314]. Participants were screened with a self-report clinical questionnaire and the Symptom Checklist 90 Revised (SCL-90-R) to ensure that they had no current or previous history of neurological or psychiatric disorders. Additional exclusion criteria included recreational drug use in the past month, current or history of substance abuse or addiction, any contraindications to MRI or TMS (e.g., pregnancy, epilepsy), history of traumatic brain injury, participation in any TMS or ECT study in the past 8 weeks, and unwillingness to consent or to be informed of incidental findings. Informed consent was obtained from all subjects before their inclusion in the study. The study protocol was approved by the Ethics Committee of the

University Medical Center Göttingen (UMG). This study was conducted in accordance with the Declaration of Helsinki.

For each participant, experiments were conducted over the course of 3 visits with approximately one week in between each visit. As described in [314], on visit 1 a structural T1-weighted volume and rsfMRI were acquired for the identification of a subject-specific DLPFC site for rTMS stimulation. This target was then used for real and sham stimulation protocols on visit 2 and 3. At the beginning of visit 2 and 3, an rsfMRI scan (R0) was obtained pre-rTMS. Next the resting motor threshold (RMT) for each subject and session was determined, which was then used to set the stimulation intensity (i.e., 110% of the RMT). Thereafter, a 10 Hz rTMS clinical protocol of 3000 pulses was delivered over 37.5 min. This procedure was additionally controlled by sham rTMS in a double-blind counterbalanced crossover design. RTMS was precisely delivered to each subject at the pre-selected DLFPC target, guided by online neuronavigation. Three additional rsfMRI scans were obtained post- rTMS at 10-15 min (R1), 27-32 min (R2), and 45-50 min (R3) after the end of stimulation (non-continuous time slots of about 5 min each). These acquisitions allowed for the assessment of functional connectivity changes induced by rTMS.

2.2 Neuroimaging data acquisition and preprocessing.

Structural T1-weighted scans with 1-mm isotropic resolution and functional data were obtained using a 32-channel head coil and 3T MRI scanner (Magnetom TRIO, Siemens Healthcare, Erlangen, Germany). For rsfMRI, 125 volumes were acquired in approximately 5.5 minutes using a gradient EPI sequence with the following parameters: TR of 2.5 s, TE of 33 ms, 60 slices with a multiband factor of 3, FOV of 210 mm × 210 mm, 2×2×2 mm, with 10% gap between slices and anterior to posterior phase encoding.

RsfMRI preprocessing was conducted in Data Processing Assistant for Resting-State fMRI software (DPARSF V4.4, [315]). Initial steps of preprocessing included the slice timing and head motion correction [316], [317]. Afterwards corrected images were analyzed using the SPM12 gradient echo field map unwarping tool [318]. White matter, CSF, and global signal were then regressed out to additionally reduce nuisance effects [319]. Corresponding T1-weighted images were adjusted to the standard Montreal Neurological Institute (MNI) template and then used for spatial normalization of rsfMRI with a resampled voxel size of 3x3x3 mm. The preprocessed images were then spatially smoothed with a 4x4x4 mm full width at half

maximum (FWHM) Gaussian kernel. Data were then detrended and band-pass (0.01–0.1 Hz) filtered to remove bias from low-frequency drift and high frequency noise.

2.3 Parcellation of individual-level resting state functional correlations

Subject-related parcellations were obtained using two methods: (1) the "Snowballing" algorithm provided a connectivity node density map [284] and (2) Boundary Mapping generated an average spatial boundary map [283], [284]. Both methods use whole-brain resting state functional connectivity (RSFC) to create individualized three-dimensional node and boundary maps. Briefly, the Snowballing algorithm uses seed-based RSFC to identify locations that are correlated (the "neighbors") with a starting seed region of interest (ROI) location. These neighbors are then used as the new seed regions and the procedure is repeated iteratively. Each iteration of identified neighbors in this procedure is referred to as a "zone". The neighbors correlated with a given seed region are spatially clustered with their distinct local maxima. Averaged peak voxel locations over multiple zones results in a node density map that represents the number of times a voxel was identified as a node across all ROI correlation zones from starting seed locations. As in [284], nodes were identified as the peak distribution from three zones of Snowballing. In the original study the snowballing parcellation was performed on a cortical surface. Here, we extended it to the whole-brain volume space to include subcortical regions and to match the dimensionality of our feature selection methods. As this has not been shown before, we have validated 3D node density maps by showing that inter-individual variability exceeds intra-individual variability both in the current dataset as well as in an independent dataset (**Supplementary Figure 2, left panel**). Most importantly, these node distribution maps have been shown to be invariant to the locations of starting seeds as well as the seed sizes [284]. While in the original study 264 seeds were used, our 278 initiation points corresponded to 278 ROIs of 5mm radius, containing additional subcortical and subcallosal seeds [285]. The correlation maps were thresholded at $r>0.3$.

In contrast to the definition of nodes, RSFC-Boundary Mapping identifies locations where the patterns of RSFC correlations exhibit abrupt transitions, and therefore provides estimates of putative boundaries between functional regions [283]. In contrast to the initially proposed technique that flattens a given subject's cortex into a 2D surface [284], we performed all computations directly on the subject's full-volume 3D Cartesian grid. This allowed us to simultaneously overlay, for each subject, the 3D Snowballing node density images with the RSFC-Boundary Mapping output (**Figure 1**). Similar to full-volume nodes, full-volume

boundaries were validated by showing that inter-individual variability surpassed intra-individual variability (**Supplementary Figure 2, right panel**). In line with the original report, we also found that boundaries resulting from RSFC-Boundary Mapping (**Figure 1, regions in green**) typically surrounded the nodes of high peak count identified by the Snowballing algorithm (**Figure 1, regions in red**). Aside from this difference, we followed the RSFC-Boundary Mapping method as previously described [284]. To speed-up computation time, the full-resolution preprocessed fMRI data were first resampled to a coarser grid, resulting in 19,973 voxels (4.5x4.5x4.5 mm in size). The whole-brain RSFC map was then computed for each voxel by correlating the time series of the given voxel with every other voxel. Then, the similarity between each voxel's temporal correlation map and every other voxel's temporal correlation map was computed as the spatial correlation between them. That resulted in 19,973x19,973 symmetrical spatial correlation matrices. Every row of this matrix corresponds to a reference voxel and can be displayed in the volume of the brain where the value in every voxel is the similarity between the temporal correlation map at that position and the reference voxel. Every row was then mapped to the brain mask back, representing voxel's similarity map. Next the spatial gradient was computed and then averaged across all similarity maps. The average of those spatial gradient maps represents the probability with which each location is identified as a point of rapid change in the RSFC maps between two neighboring areas.

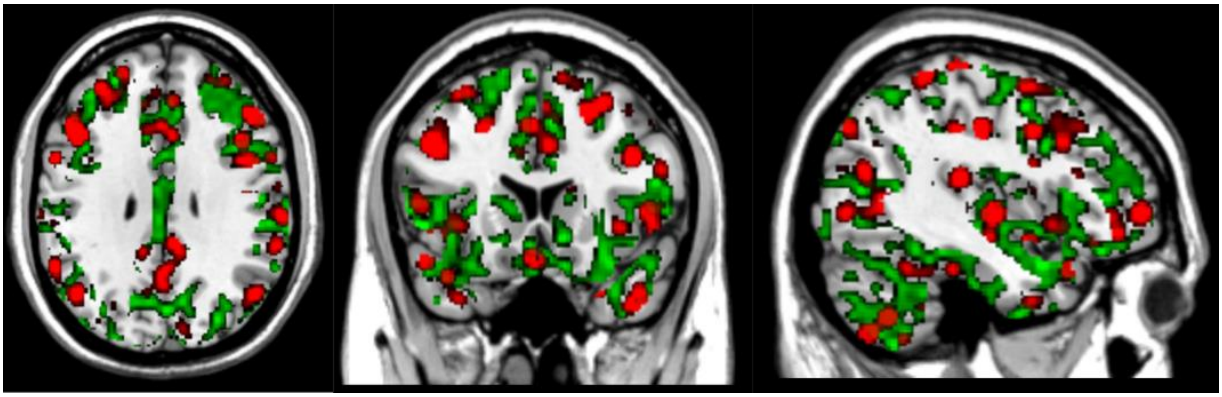


Figure 1: An example of individual-level node and boundary maps. It is noteworthy that calculated RSFC-Snowballing nodes (red areas, thresholded r=0.3) are often surrounded by transitional areas represented by RSFC-Boundary mapping boundaries (green areas thresholded r=0.08), revealing the complementary nature of both parcellation methods.

2.4 Feature selection

Machine learning models can be directly applied to both node and boundary maps. However, due to a large number of features (104-105 voxels) compared to the number of subjects in this study (n=23), that would lead to a substantial overfitting. Subsequently, overfitted classification models suffer from unsatisfactory interpretability and accuracy [238], [320]. In order to select a smaller number of physiologically plausible voxels located in node and boundary maps, we employed a novel data-driven feature selection approach based on the resting state networks (RSNs). RSNs were accessed in the same group of subjects but from an independent dataset (Visit 1) from the classification dataset (Visit 2 and 3). The RSNs were identified by group ICA [286], [320] using MELODIC tool of FSL 5.0.7. We temporally concatenated the fMRI data of 23 subjects recorded on visit1. Based on visual inspection and the power spectrum of the MELODIC output, we selected the nine best-fitting spatiotemporal independent components (IC). While the node regions were assumed to lie close to maxima of the selected ICs, the boundary areas were rather identified at the intersections of the ICs.

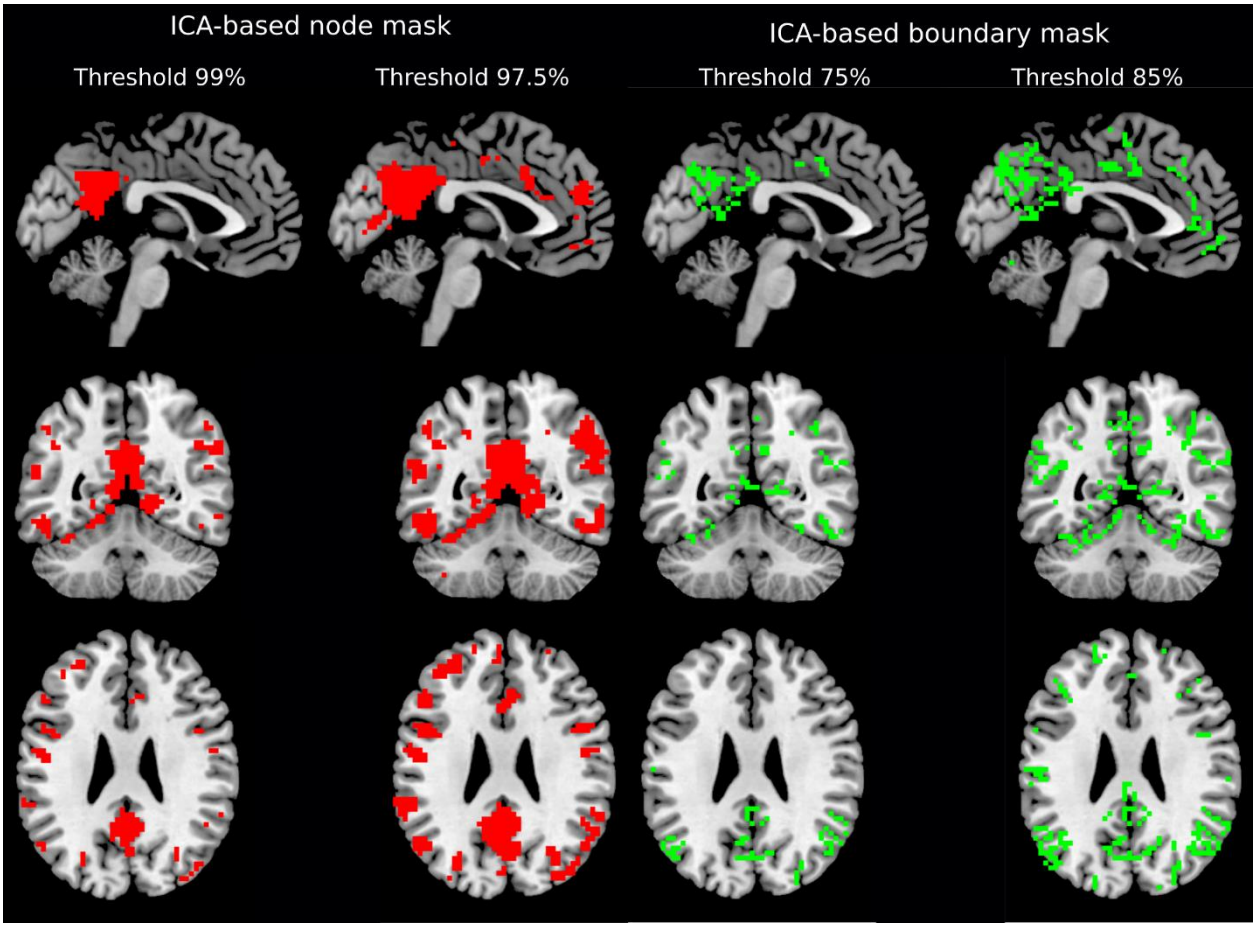


Figure 2: ICA-based binary masks of node (left) and boundary masks (right) from a dataset not used in machine learning classification. ICs node regions grow in size when one increases the

threshold value. At the moment two or more ICs meet in one particular voxel, that voxel is identified as the boundary. Changing the threshold value allows for the optimization of the number of voxels corresponding to both masks.

As the different ICs showed different strengths (distributions of values across the brain) and we intended to have multiple networks represented rather than a single "winner", the node areas were defined by overlaying all ICs. The percentage of voxels belonging to every IC was controlled by the threshold; voxels with lower strength were discarded. Next, the components were merged together resulting in an array of nodal voxels. The value of the node density (Snowballing) map of each subject/time interval was then extracted for each of those voxels. This process was designed to select the most important features that would then be modeled using the classification algorithm. An example set of node masks based on all subjects, but independent from the dataset used for classification, is presented in **Figure 2**, left panel (thresholded at 97.5% and 99% of discarded voxels).

In case of the boundary mask, we have developed an algorithm closely related to watershed transform [321]. The underlying idea of watershed transform is finding an optimal position for dams to be built, where the water coming from different basins, according to the surface shape, would meet. In our case, the surface is represented by all ICs and the basins are the strongest voxels according to ICs. Starting from the maximum of each component (100% threshold), and then reducing the threshold in 1% steps, the ICs increase in size. A voxel is identified as a boundary when two or more ICs intersect each other (i.e., the same voxel is included in multiple ICs). Single voxels, i.e. not surrounded by more voxels from the same IC, were not treated as boundaries to avoid spurious findings. The number of included voxels were dependent on the percentage level step that was used to descend from the maximum. An example set of boundary masks based on all subjects, but independent from the dataset used for classification, is presented in **Figure 2**, right panel (threshold values of 75% and 85%).

2.4 SVM classification

The capacity of SVM to predict outcomes of the rTMS intervention was first assessed using leave-one-group-out stratified cross-validation (CV), where the group was defined as all of stimulation conditions (time intervals R) of the same subject. Our main analysis employed SVM for multiclass classification with "one-versus-one" approach of all sham-subtracted

conditions, i.e. R0 vs R1 vs R2 vs R3. SVM classification was performed using a linear kernel and the default scaling factor of 1. Sex and age were not regressed out because in every comparison both groups were represented by the same subjects, and thus automatically balanced.

The number of features (voxels) used as an input for SVM varied by changing the threshold value of the extracted ICA-based node and gradient masks (**Figure 2**). SVM assigns weight to every voxel, which can be interpreted as an importance of the voxel to separate conditions. The SVM was trained for every pair of conditions starting with a mask threshold corresponding to 10,000 voxels. By changing the threshold for both ICA-based node and gradient masks, we removed the "weakest" voxels and trained the SVM again from scratch. This procedure was repeated until the total number of voxels surviving thresholding reached zero. This recursive process allowed for the assessment of a model accuracy curve that was defined by the percentage of included voxels. To access the confidence interval of the accuracy curve, we ran SVM on 1000 bootstrap samples. The global maximum of this model mean across bootstraps accuracy curve represents the most informative set of voxels for classifying all of the conditions. In case the accuracy curve (or a certain part of it) is consistently higher than the chance level (>25%), we perform pairwise classification, i.e. R0 vs R1, R0 vs R2, R0 vs R3, R1 vs R2, R1vs R3 and R2 vs R3, to identify the time and the direction of the most significant changes happening after 10Hz rTMS. The same algorithm was applied to both the node density and the gradient maps across voxel thresholds. This procedure is schematized in **Figure 3**.

CV is well known as an effective method in machine learning for testing generalizability of a trained model. Model performance was further tested using leave-two and leave-three-out stratified CV (7 fold and 11 fold CV) in which 2 or 3 subjects were withheld from training and assigned to the test sample. Our goal with this approach was to investigate the effect of the amount of data provided to a classifier, and thereby assess the impact of CV strategy. By providing more training data, as in leave-one-out CV, the model has more generalizable performance compared to leave-two-out and leave-three-out CV [322]. The computations were completed in Python using custom-written scripts that used functions from the Nilearn v0.7.0 (https://nilearn.github.io/) and Sklearn v0.23.2 (https://scikit-learn.org/stable/) libraries.

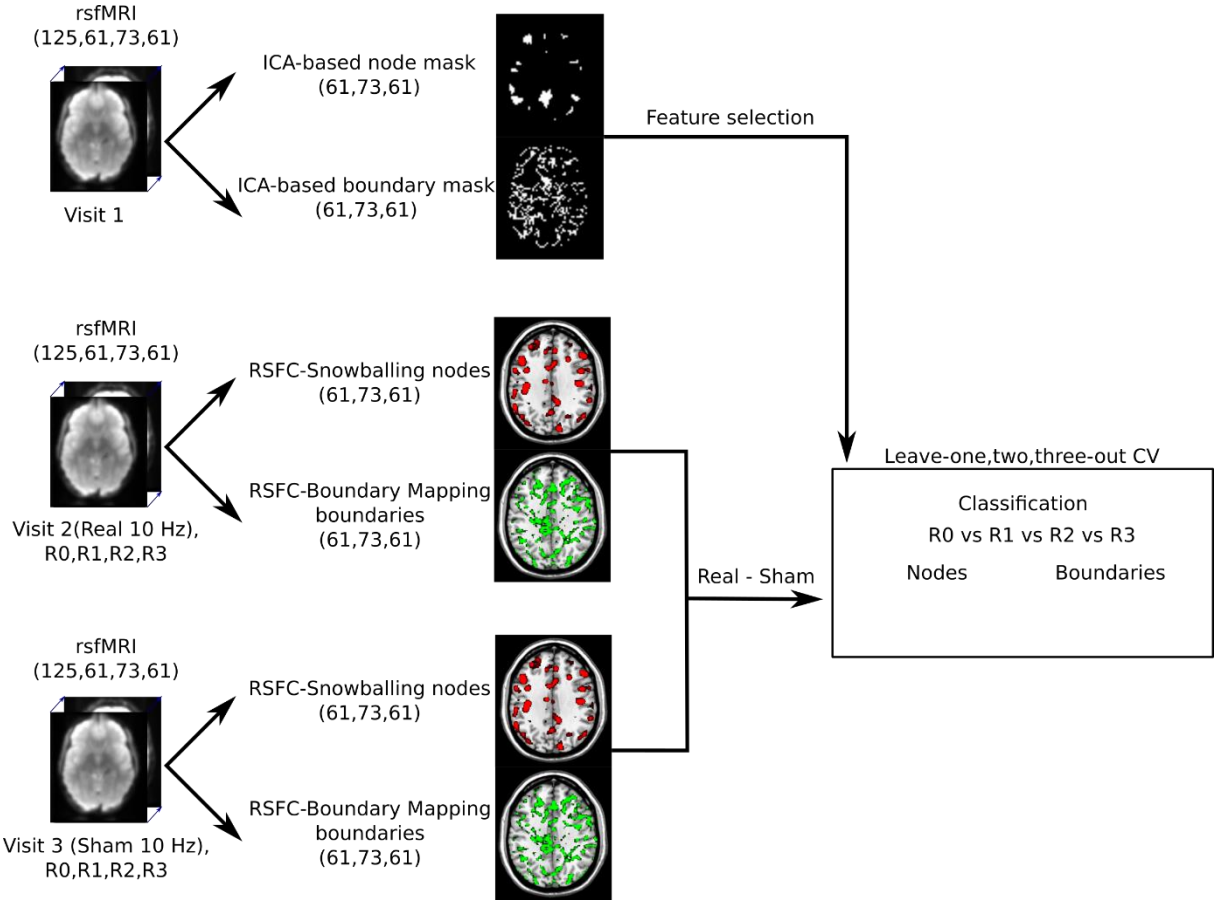


Figure 3: Schematic diagram of the analysis pipeline. RsfMRI data from 4 sessions (R0, R1, R2, R3). Both rsfMRI from Visit 2 (real rTMS) and Visit 3 (sham rTMS) were used to compute RSFC-Snowballing density nodes and RSFC-Boundary Mapping gradients. Next sham stimulation maps were subtracted from the corresponding real stimulation maps. The extracted ICA-based masks derived from the visit1 measurements were then applied to the corresponding node and gradient maps for feature selection. Finally, the remaining voxels were used for machine learning classification.

2.5 Effect of head motion on the separation of conditions

To exclude the possibility that condition-related differences are caused by head motion [323], we performed pairwise classification of mean frame-to-frame head displacement across real, sham, and real-sham conditions (R0, R1, R2, R3). The resulting SVM performance accuracies are presented in **Supplementary Table 2**. As the performance of algorithms based on head motion are close to chance, this analysis confirmed irrelevant influences of head motion on the results of classification based on RSFC nodes and boundary maps. Additionally, a cutoff to

remove the subjects with high-motion frames according to the threshold (mean FD = 0.5) was set, yet no session surpassed this value.

3. Results

3.1 RSFC nodes' density maps

We performed SVM classification on RSFC nodes density maps across all conditions and applied an iterative ICA-based feature selection step to spatially identify nodes that were strongly modulated by 10 Hz rTMS. The threshold for the ICA-based node mask, corresponding to the percentage of voxels removed, varied between 99.9% (154 voxels) to 94.5% (10,799 voxels) in 0.1% increments. The highest accuracy of 33.2±0.8% was achieved for a threshold of 99.1%, yielding 1690 voxels fed to the SVM (**Figure 4a**). Results indicated that by increasing the threshold, many informative voxels for the classification were discarded, yielding the massive accuracy drop between 99.9% and 99.7%. Lower thresholds led to a decrease in accuracy, which may be due to model overfitting (i.e., inclusion of uninformative voxels). Additionally, we compared the performance of the SVM model with a different number of CV folds. All three CV schemes showed similar accuracies across thresholds ranging from 99.9% to 98.7%. Lower thresholds for leave-three-out CV resulted in slightly higher accuracies compared to leave-one-out and leave-two-out CV. The most informative voxels were located in the posterior cingulate cortex, angular gyrus, anterior insula, and visual cortex (**Figure 5, top panel**).

Since the accuracy was consistently higher than by chance, we aimed to discriminate the sessions in which connectivity was most strongly modulated by 10 Hz rTMS by performing pairwise classification of conditions. The highest accuracy of 74.2±1.1% was achieved for R1 vs R2 comparison with a threshold value of 99.7%, yielding 764 voxels used in SVM-based classification (**Figure 6a**). All three CV schemes yielded similar accuracies for R1 vs R2 classification across thresholds ranging from 99.9% to 98.7% (**Supplementary Figure 3**). Majorly, the same set of voxels with both positive and negative weights was found modulated in R1 vs R2 classification (**Figure 7, top panel**) as in multiclass classification.

3.2 RSFC boundary maps

For boundaries extracted from RSFC-Boundary Mapping, the threshold for ICA-based boundary masks was varied from 50% (23 voxels) to 94% (10861 voxels) in 1% increments.

The highest accuracy of 32.4±0.9% was achieved by the threshold of 59% corresponding to 113 voxels being fed into the SVM (**Figure 4b**). These voxels were predominantly located in the ventral posterior cingulate cortex, precuneus, angular gyrus, and fusiform gyrus (**Figure 5, bottom panel**). Consistent with the analysis on Snowballing nodes described above, increasing the number of voxels extracted from the ICA-based boundary mask caused a drop in the overall accuracy of the SVM. All three CV routines showed a similar pattern in accuracy across the threshold range of 50%-67%.

Similar to the analysis based on RSFC nodes' density maps, we performed pairwise classification of conditions to detect the timing of the strongest changes in boundaries. The highest SVM accuracy of 74.5±1.1% was achieved for R1 vs R3 (**Figure 6b**) for the threshold value of 63% (174 voxels, **Figure 7**), Using leave-one-out CV resulted in the performance improvement from 72.3% to 74.5% compared to leave-two-out CV for the ICA-based boundary mask threshold of 63% (**Supplementary Figure 4**). The learning curve of R1 vs R3 across CV approaches (**Supplementary Figure 4**) indicates the highest masking threshold at which leave-one-out CV strategy yields the highest (or at least equal accuracy compared to leave-two and leave-three-out CV) was 76%. This masking threshold corresponded to 1194 voxels. For higher threshold values, leave-two-out CV exhibited slightly higher accuracy than the other approaches.

The second highest accuracy of 68.5±1.3% was achieved by R1 vs. R2 classification (**Figure 6b**) using a threshold value of 67% that corresponds to 346 included voxels. For threshold values from 60% to 86%, both leave-one-out and leave-two-out CV strategies were associated with higher accuracy compared to leave-three-out CV (**Supplementary Figure 4**). The sign of feature weights in the majority of identified voxels was in the opposite direction to the R1 vs R2 and R1 vs R3 comparison (**Figure 8a and b**). The highest classification results are presented in **Supplementary Table 1**.

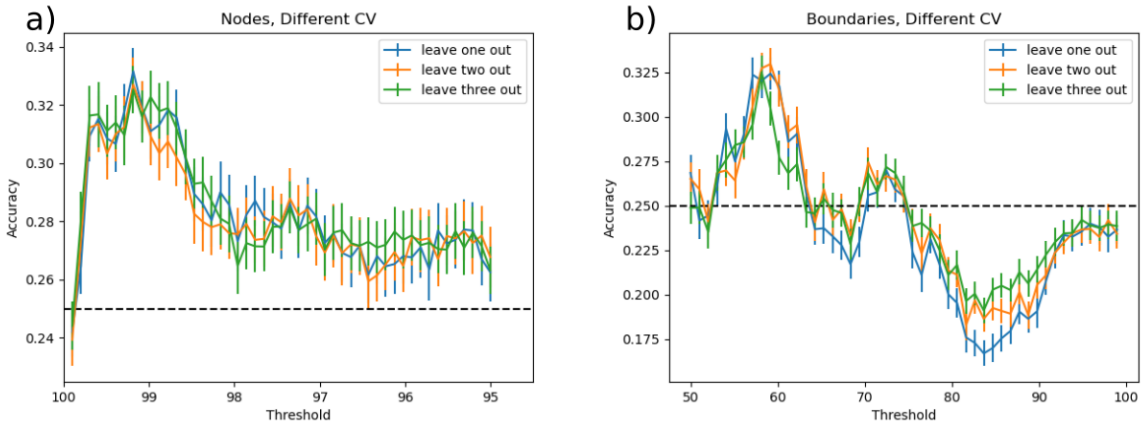


Figure 4: Multiclass classification accuracy of SVM using node density map (**a**) and boundaries (**b**) across all threshold levels. The error bar indicates 95% confidence intervals over 1000 bootstraps. To test the stability of classification results, classification was also performed for leave-one-, -two- and -three-out cross-validation (CV) routines. The highest performance for peaks (Acc =33.2%) was achieved for the threshold value of 99.1% corresponding to 1690 informative voxels. For boundaries, the threshold of 59% yielded the highest performance (Acc=32.4%) resulting in 113 voxels to be strongly modulated via 10Hz rTMS. Dashed line represents random-choice accuracy, one of four R categories.

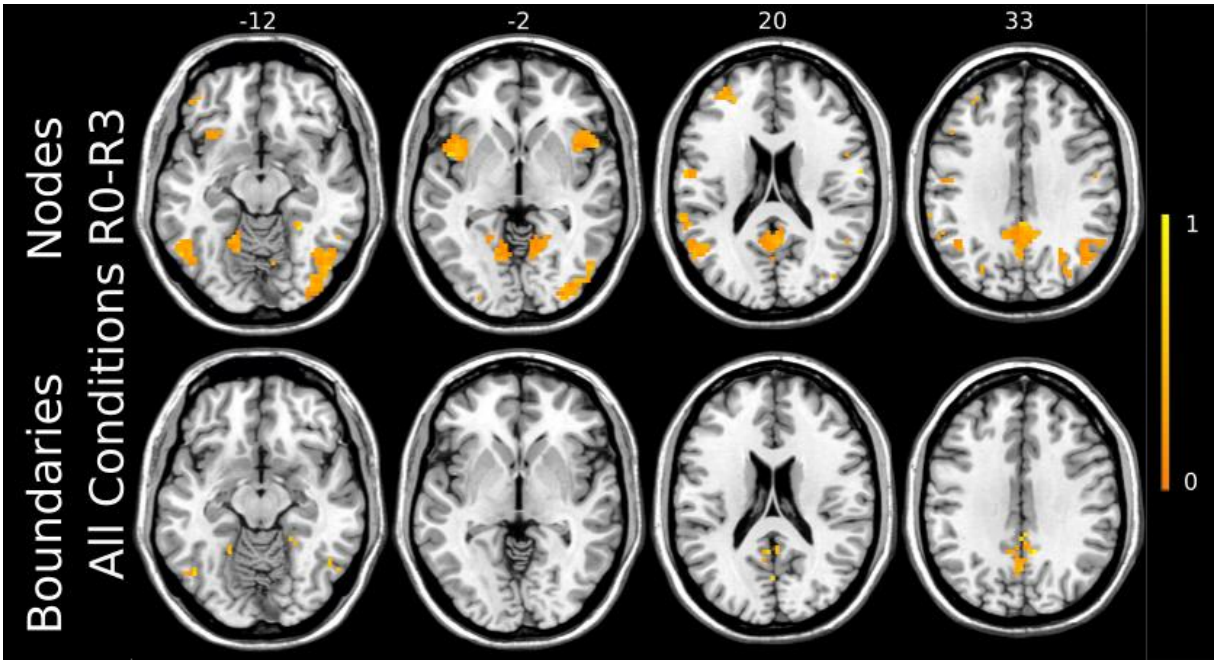


Figure 5: Most strongly modulated voxels corresponding to the highest multiclass classification accuracy of SVM for both nodes and boundaries. The strength of modulation is color-coded by

a warm colormap over the individual template. The top numbers refer to axial plane z-coordinates in MNI space.

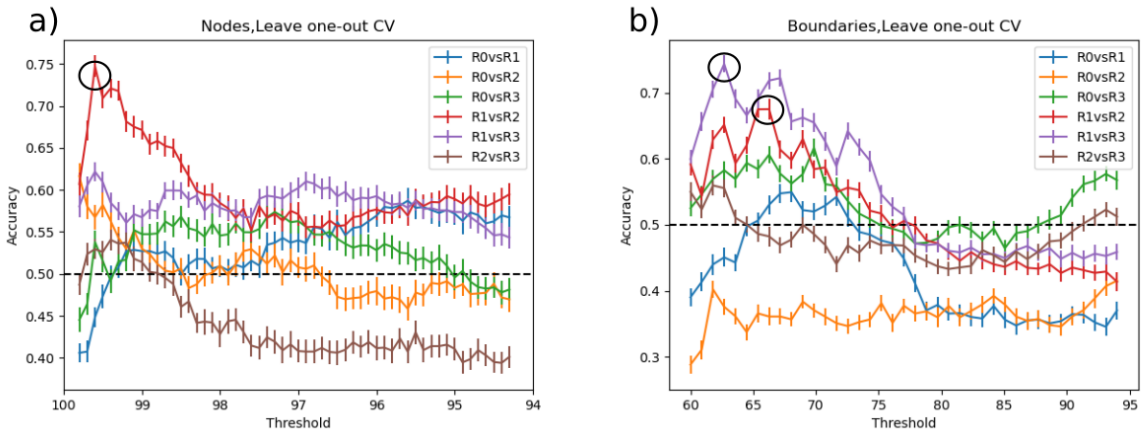


Figure 6: Pairwise classification accuracy of SVM using nodes (a) and boundaries (b) across all threshold levels. The highest accuracy for nodes was achieved by R1 vs. R2 comparison for the threshold value of 99.7%, corresponding to 764 voxels. In case of boundaries, R1 vs. R3 yielded the highest accuracy for the threshold value of 63% (174 voxels). The second highest accuracy was obtained by R1 vs. R2 comparison for the threshold value of 67% resulted in 346 voxels being highly modulated by 10 Hz rTMS. Dashed line represents random-choice accuracy, one of two R categories.

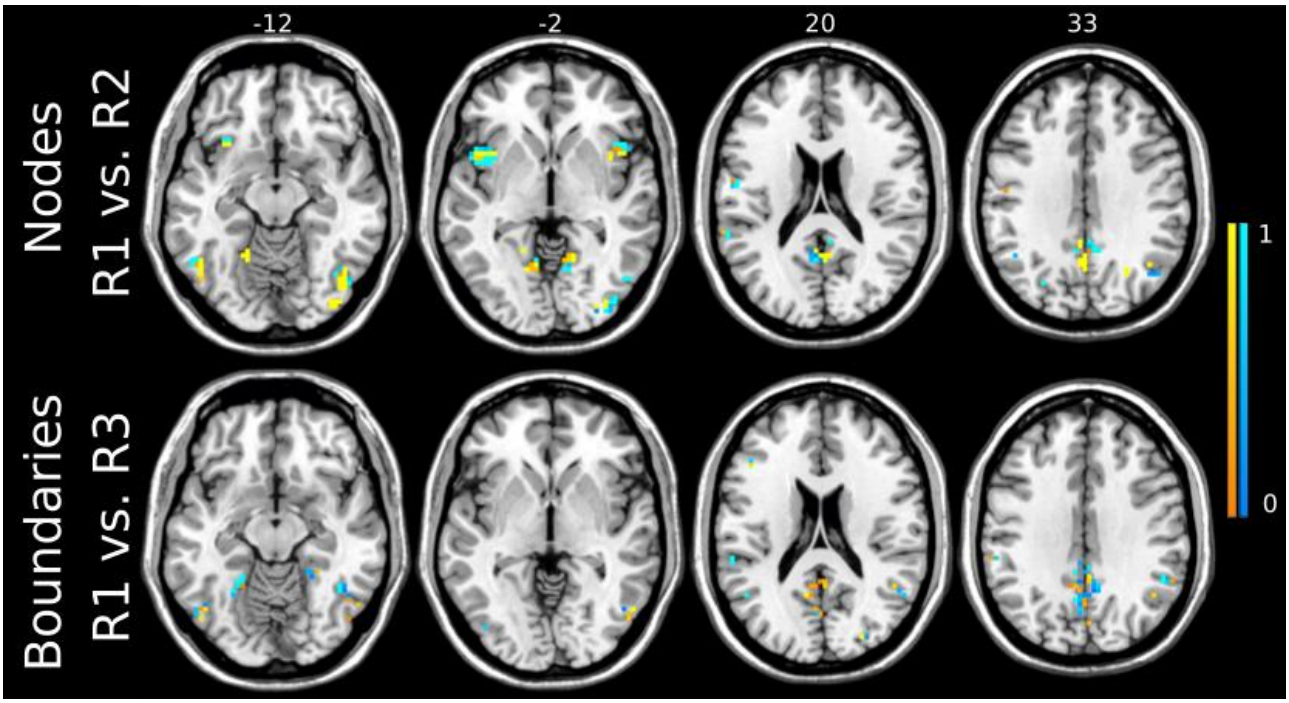


Figure 7: Voxels corresponding to the highest performance of the SVM using the Snowballing nodes for R1 vs R2 comparison (top) and the Boundary-Mapping boundaries for R1 vs R3 comparison (bottom). The sign and strength of modulation is color-coded by warm colormap (connectivity increase) and cold colormap (connectivity decrease). The top numbers refer to axial plane z-coordinates in MNI space.

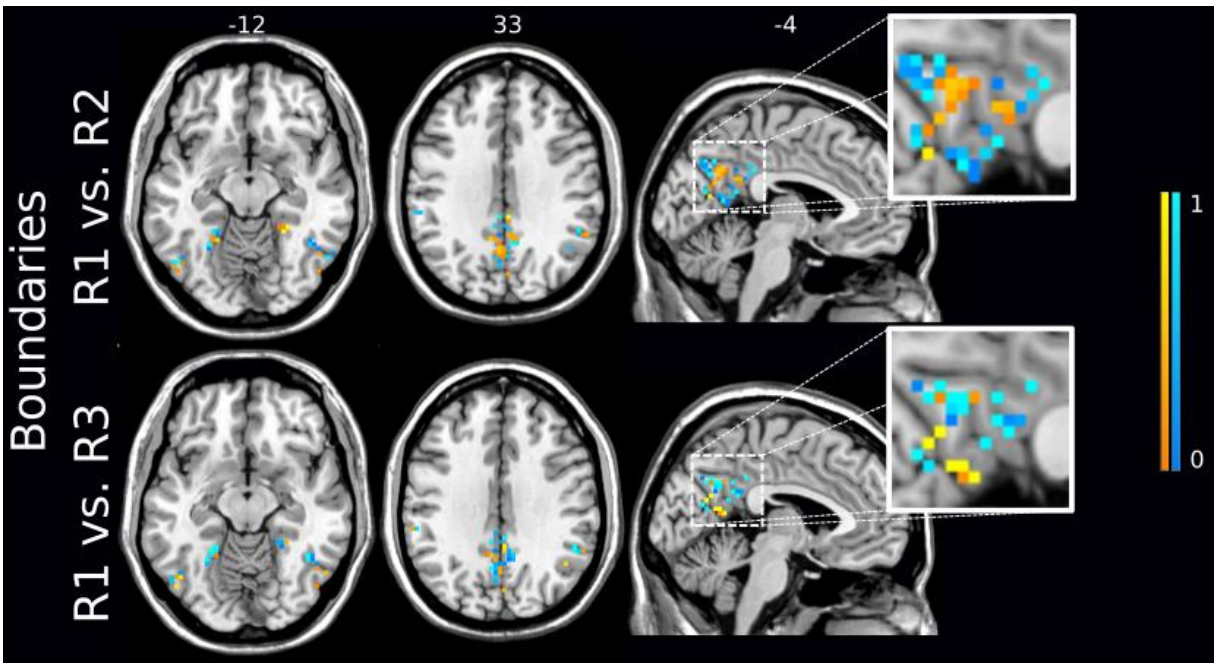


Figure 8: Two highest classification accuracies based on the Boundary-Mapping boundaries were achieved for comparisons R1 vs R2 (top) and R1 vs R3 (bottom) with 329 and 174 stable voxels respectively. The sign and strength of modulation is color-coded by a by warm colormap (connectivity increase) and cold colormap (connectivity decrease). The zoomed-in image shows that the majority of discriminative voxels in both comparisons were located in the posterior cingulate cortex (PCC) and precuneus. The top numbers refer to axial plane z-coordinates in MNI space.

4. Discussion

There is considerable interest in understanding the changes in functional connectivity driven by 10 Hz rTMS, both in the field of neuroscience and in applied clinical practice. Here we have used a predictive SVM model approach to identify the locations and time intervals after stimulation that exhibit the most substantial whole-brain functional changes in large-scale network nodes and boundaries. To overcome the issue of high-dimensional data, that is, the number of voxels in node and boundary maps being much greater than the number of subjects in the study, we have applied feature selection using multiple threshold ICA-based masks created from separate fMRI data. Confirming our hypothesis, we have identified connectivity changes related to 10 Hz rTMS in both network node and boundary maps. Of note, as indicated by cross-validated SVM, changes majorly comprise the PCC, angular gyrus, insula and fusiform gyrus. Accuracy confidence intervals in the classification of boundaries are similarly as substantial as those occurring in nodes. A complex pattern of changes was observed in boundaries alternating between decreases and increases in functional connectivity, which was particularly evident but not limited to the PCC and precuneus.

4.1 Individual nodes and boundaries

This work extends to the concept of modularity and integration frequently addressed in the field of connectomics [293],, both of which focus on the perspective of nodes in the formation and reorganization of networks in cognitive functions and in clinical disorders. Most connectomics studies to date have been focused on the strongest points of functional connectivity, “hubs”, and the way that such hubs are organized to efficiently propagate information across regions [324], [325]. Transition gradients have recently received more attention in the literature [326]–[328]. To the best of our knowledge, the current study is the first to show which changes occur in both functional connectivity nodes and boundaries after 10 Hz rTMS. Indeed, based on their accuracy and confidence intervals, we see that rTMS effects related to boundaries are similarly

as substantial as those seen in nodes. Cytoarchitectural divisions of the PCC into dorsal and ventral parts have been previously shown to exhibit distinct functional connectivity patterns [329]. According to findings from a recent study [171], both PCC and precuneus contain predominantly transient nodes - an entropy class that the authors used to classify changes in network assignments across subjects and brain states. Our findings are consistent with this notion, highlighting that further functional heterogeneity is largely elicited in these regions, particularly in boundaries, after 10 Hz rTMS.

The exact function of boundaries is not yet well understood, which might have discouraged its systematic evaluation. It may be speculated that boundaries may act to segregate information within functional regions, or that they may support network stability [330]. Another possible role of boundaries may be supporting functional adaptation of a given region during plastic changes in the mature primate brain. Evidence for this has been shown in primary motor and sensory regions (for review see Florence et al., 1997). In support of this theory are several studies that have shown that rTMS-induced changes in neuronal inhibition can prime cortical networks for the expression of subsequent experience-dependent plasticity [312], [332], [333]. The high frequency rTMS potentially creates a cortical state with enhanced plasticity, opening a time window for targeted re-learning of connectivity patterns [312]. A complex pattern of changes in the functional connectivity of boundaries, which was particularly evident but not limited to the PCC and precuneus, is reminiscent of the neuronal population dynamics in the cat visual cortex after 10 Hz rTMS perturbations [313]. In the animal experiments, stimulation induced initial suppression of on-going cortical activity, followed by an increase in cortical excitability that lasted about 2-3 hours, but was prone to slow activity fluctuations. Worth mentioning that this animal study as well as clinical TMS/EEG findings reported by [334] suggest a different mechanism than excitatory LTP. Indeed, rTMS may instead reduce the local intracortical inhibition leading to long-lasting neuromodulatory effects in both the boundaries and the nodes.

Causal effects of 1 and 5 Hz rTMS on global functional connectivity have been explored recently using fMRI [335], [336]. These studies have found that distant effects, that is, effects relatively far from DLPFC, are determined by connectivity profiles of the stimulation target with macroscopic networks. Excitatory 10 Hz rTMS in the left DLPFC, as we used in the current study, resulted in multivariate patterns of increases and decreases in functional connectivity. These fluctuations occurred primarily in the PCC, angular gyrus, fusiform gyrus, and insula regions. Each of these regions have been previously shown to be functionally related

to the DLPFC [337]–[340], reinforcing the notion of distant effects of this stimulation protocol. In the current study, the SVM most substantially identified functional connectivity changes in nodes that occur about 30 min after stimulation, in line with a previous study performed on the same dataset, but using factorial design ANOVA to find group differences [314]. In contrast, effects related to boundaries were temporally extended up to 45 min. Previous studies that investigated the effects of 10 Hz rTMS with rsfMRI constrained their analyses to particular seeds or networks of interest. Another important advancement provided by the current study is an assessment that started from a global evaluation of 10 Hz rTMS effects, considering maps of nodes and boundaries at the subject level at different intervals. The importance of individualized characterization of functional brain networks has been highlighted in the literature [301]–[303]. These developments may inform future clinical applications based in "precision" or "systems" medicine. Toward this goal, methods applied in the current study might have been advantageous, considering that individual subject analysis of node and boundary maps closely correspond to the original study [284]. In our study though functional boundaries were not only extracted for the cortical surface. To match the resolution and dimensionality of the node maps, we calculated boundaries on the whole brain. Furthermore, our study is unique in that we used pairwise comparisons of each individualized map type, that is, nodes and boundaries, in the context of a double-blind design controlling for placebo effects. This may have allowed us to identify the contribution of boundaries to the functional changes caused by 10 Hz rTMS. This individualized approach was followed by SVM, which might have contributed to the identification of the most important effects related to the time intervals and areas.

4.2 SVM classification and feature selection

SVM is a machine learning approach with clear advantages over univariate models. With that said, important preconditions had to be fulfilled to avoid potential biases. One of the main concerns when we applied SVM was to prevent overfitting. The full-resolution rsfMRI was first resampled to 19,973 voxels to reduce the computation time of the spatial correlation matrix – the most computationally demanding step in the algorithm. Splitting the rsfMRI signal into two complementary maps also had the practical benefit of enabling independent assessment using SVM. This also reduced the number of input variables to the model, and thus further prevented overfitting. Secondly, the large number of voxels in both individualized node and boundary maps required additional feature selection. While there are several masking algorithms that can locate functional nodes of the brain, to our knowledge, there is no complementary boundary

parcellation method associated with any of them. For this reason, we have used group ICs to create both nodal and boundary masks.

Voxels included in the masks are controlled by the threshold. While these masks were built in a complementary manner, these masks included common voxels in several regions, including ventral PCC, the boundary area between angular gyrus and fusiform gyrus, and the boundary area between fusiform gyrus and associative visual cortex. These regions are consistent with findings from a study that reported on low stability of connectivity regions [171]. Some of these regions have been associated with myelin content, particularly the ones that myelinate latest during development [341], [342]. As some of the implicated regions overlap with those with late myelination, this suggests a role for flexible connectivity in learning [336].

Our approach is closely related to the recursive feature elimination technique that has been previously applied in neuroimaging [258], [322]. In our method, voxels/features are not removed based on their rank according to SVM, but rather based on the information obtained from the independent dataset (not included in SVM modeling). Therefore, our approach is less prone to overfitting and inflating performance accuracy. On the other hand, a drawback of our approach is that some fraction of the voxels not included might be potentially informative for classification.

Considering our small sample size, the performance of the classifier was low – close to chance – when all the features were used. Since in the multiclass classification SVM has shown the accuracy consistently higher than by chance, we performed post-hoc pairwise classification of conditions to pinpoint the timing of the most significant changes. For most of the pairwise comparisons, there was a substantial drop in the accuracy when more than 1000 voxels were used in the model. This observation may be explained by the presence of too many insignificant voxels and a small sample size. After the optimal number of voxels was reached from unbiased feature selection, further removal of voxels resulted in a drop of the accuracy. This is consistent with previous studies that have used similar machine learning approaches [322], [343]. Moreover, we investigated estimated accuracy curves for different CV methods as the number of CV folds may influence the accuracy of machine learning. That is, threshold values that yielded higher accuracy for leave-three-out, in contrast to leave-one-out CV, are more likely to have inflated accuracy as the training set is smaller. In general, the trained model should underperform or at least yield similar accuracy when trained with less data, otherwise the performance of the model is ambiguous.

4.3 Limitations

Our work has important limitations that should be considered. As mentioned above, with feature selection not all information derived from both node and boundary maps was considered for classification. Therefore, we cannot exclude that additional areas might have been modulated by 10 Hz rTMS and, due to the feature selection procedure, are not identified in our results. While systematic evaluation of thresholds is a common approach in machine learning, we acknowledge that nodes and boundary masks behave differently when changing threshold values. This may have contributed to the exclusion of regions as part of boundaries at a faster rate than nodes. One possible explanation for this is non-uniform distance between networks, which may cause proximal areas in boundaries to interact/overlap at lower thresholds, compared to more distant regions in nodes. In addition, a larger sample size may have enabled further methodological improvements such as: (1) increasing the resolution in the masking procedure; (2) consideration of more voxels from the masks; (3) application of non-linear predictive algorithms; (4) better fine-tuning of classifier hyperparameters; and (5) an independent validation dataset to strengthen the reliability of our results.

5. Conclusion

Our findings provide evidence that SVM classifiers using ICA-based feature selection can identify different spatial definition, direction, and timing in the pattern of fMRI-based brain connectivity changes in functional nodes and boundaries derived from 10 Hz rTMS applied to the left DLPFC. Identified nodes and boundaries are located predominantly in the ventral PCC, precuneus, insula, and fusiform gyrus and appear approximately 30 minutes after the stimulation is performed. By dividing the signal into two complementary parts (nodes and boundaries), we highlight the contribution of boundaries to modulatory effects of high frequency rTMS. These findings provide new insights into the currently unknown role of boundaries in network organization, motivating future, related investigations for the advancement of connectomics.

Declaration of Competing Interest

The authors declare no conflicts of interest.

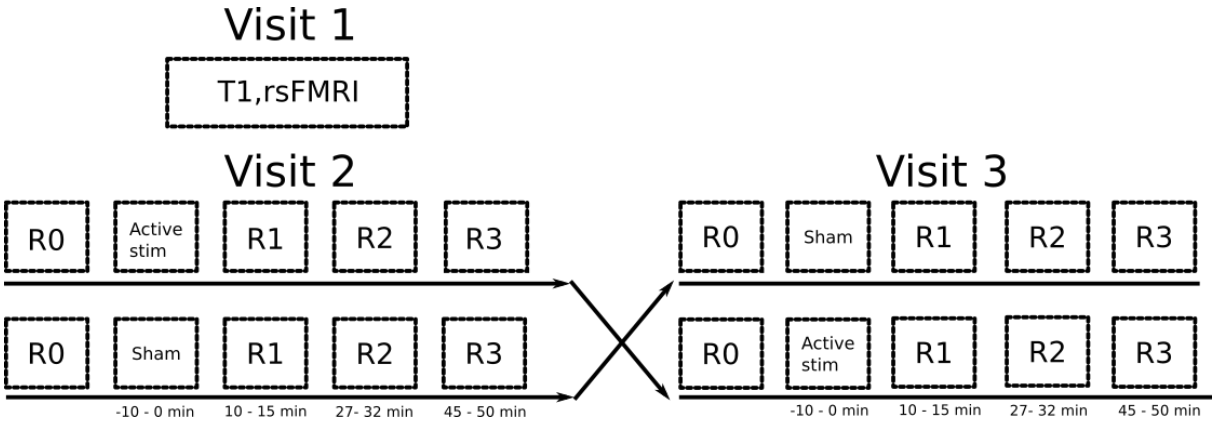
Data Availability Statement

The algorithms and codes that were used in this study are available from the corresponding author upon reasonable request.

Acknowledgements

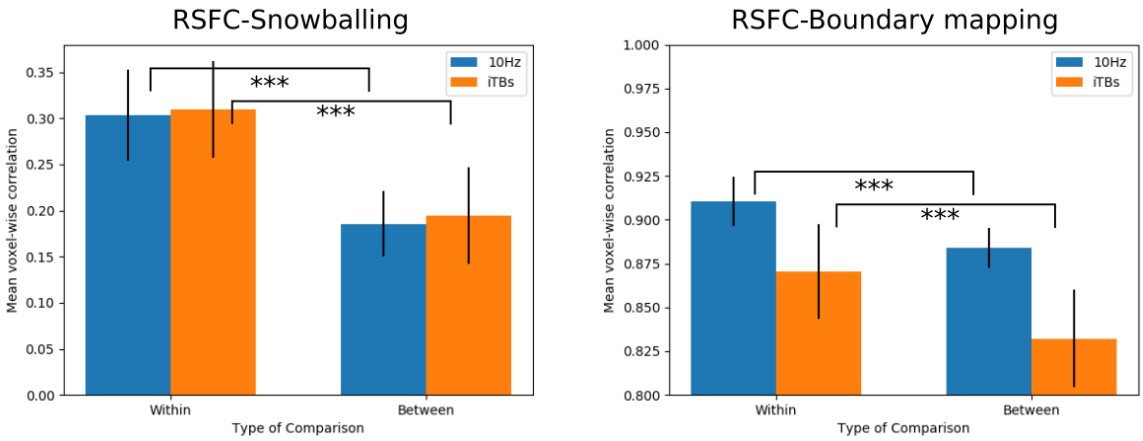
Funding: This work was supported by the German Federal Ministry of Education and Research (Bundesministerium fuer Bildung und Forschung, BMBF: 01 ZX 1507, ‘‘PreNeSt - e:Med’’). Author contributions Conceptualization: RGM; Methodology: VB, VK; Data curation: AS; Investigation: RGM; Supervision: RGM; Writing (original draft): RGM, VB, VK; Writing (review and editing): RGM, MDS. Data and materials availability: Raw data analyzed during the current study are not publicly available due to absence of written consent from the participants of the study. Scripts and codes used in the analyses will be deposited in a public database.

Supplementary Material

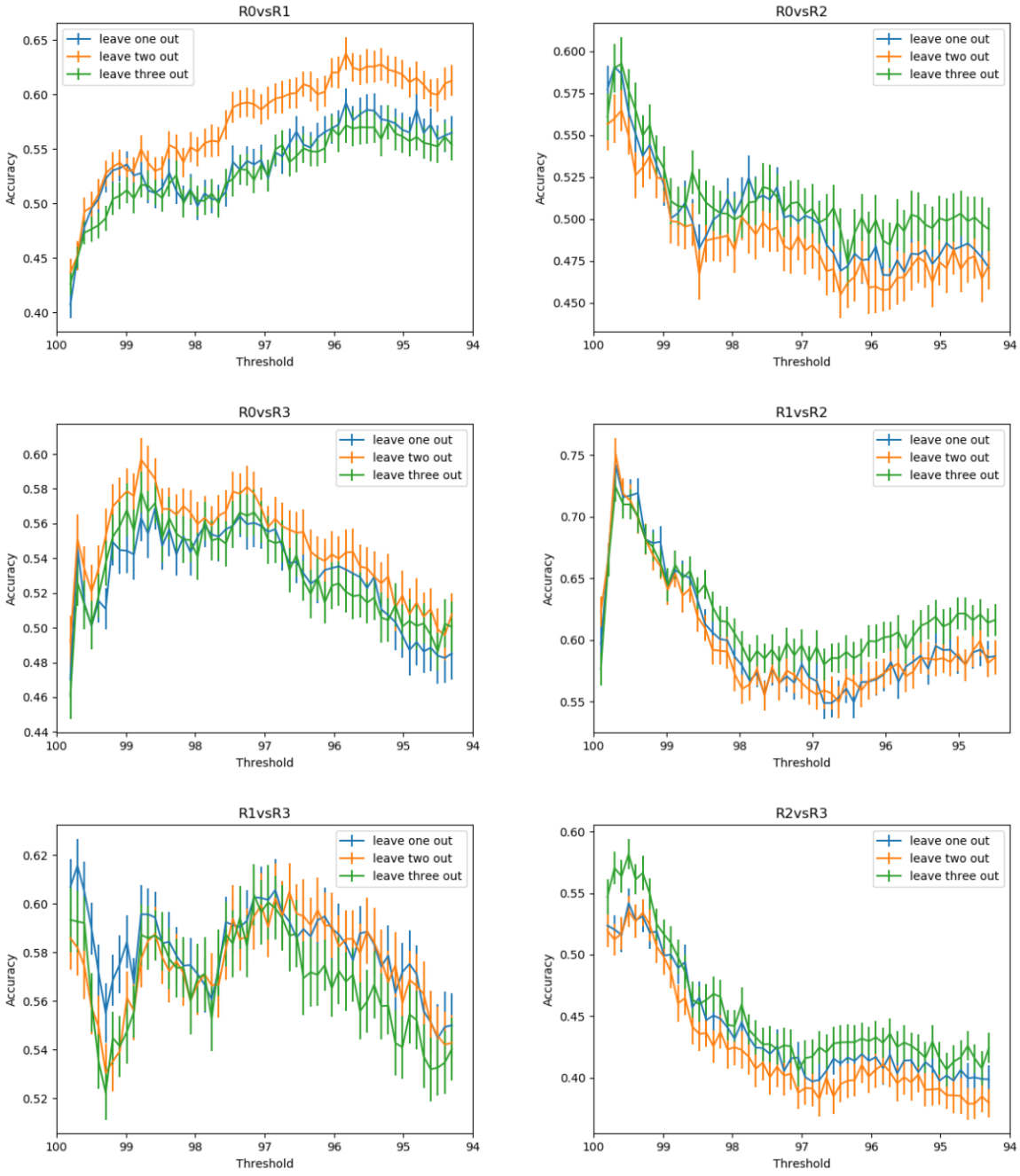


Supplementary Figure 1: Study design – We acquired T1 and rsfMRI images at visit 1 that were used for personalized target selection. The found target was then located in T1 image for stimulation one week after (Visit 2) and two weeks after (Visit 3) via online neuronavigation. Subjects were assigned to an arm of the study receiving both real and sham in a counterbalanced crossover manner. At the beginning of the sessions on Visit 2 and Visit 3, we obtained a baseline

rsfMRI scan (R0). After, the 10 Hz rTMS was delivered at the selected target. Three following scans (R1, R2 R3) were obtained after the stimulation.

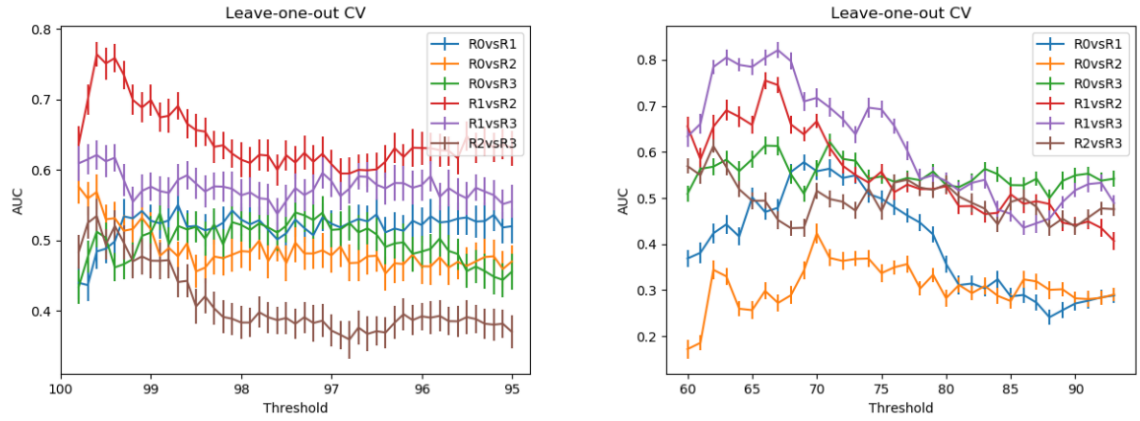


Supplementary Figure 2: Validation of 3D maps – Spatial correlation within and between node density maps (left) and boundaries (right) in healthy control subjects. Two datasets of baseline rsfMRI separated by about 1 week from independent cohorts of healthy controls ("10Hz" with 23 subjects and "iTBS" with 26 subjects in blue and orange, respectively). For both cohorts, within subject correlation was significantly higher (***p<0.001) than between subjects correlation in both nodes and boundaries. Bars represent standard deviation.



Supplementary Figure 3: Pairwise classification accuracies of nodes' density maps using three different cross-validation (CV) strategies

Supplementary Figure 4: Pairwise classification accuracies of boundary maps using three different cross-validation (CV) strategies



Supplementary Figure 5: Pairwise classification displayed as area under the curve (AUC) for node (left) and boundary mapping (right)

	Top ranking	Number of voxels	Accuracy (95% CI) [%]	AUC (95% CI) [%]
Snowballing peak density	R1 vs. R2	764	74.2 (73.0 – 75.4)	76.4 (74.6 – 78.1)
Boundary mapping	R1 vs. R2	346	68.5 (67.2 – 69.8)	74.5 (72.7 – 86.3)
	R1 vs. R3	174	74.5 (73.4 – 75.6)	80.5 (78.7 – 82.2)

Supplementary Table 1: Highest classification accuracies and area under the curve (AUC) for leave-one-out cross-validation (CV) of nodes and boundaries

Condition	Real	Sham	Real - Sham
	Mean (SD)	Mean (SD)	Mean (SD)
R0 vs R1	0.50 (0.15)	0.50 (0.29)	0.57 (0.27)
R0 vs R2	0.52 (0.10)	0.54 (0.20)	0.52 (0.31)
R0 vs R3	0.52 (0.10)	0.50 (0.00)	0.35 (0.23)
R1 vs R2	0.54 (0.14)	0.54 (0.20)	0.46 (0.20)
R1 vs R3	0.57 (0.17)	0.54 (0.14)	0.52 (0.23)
R2 vs R3	0.50 (0.00)	0.50 (0.00)	0.50 (0.00)

Supplementary Table 2: Classification accuracies (%) on head mean frame-to-frame displacement of real and sham rTMS

Chapter 6 Summary and future perspectives

The complexity of depression pathophysiology is one of the biggest hurdles to advancing treatment against depression. Despite numerous efforts to tackle this question at different levels, from molecular signatures to large-scale structural and functional brain alterations, the existing hypotheses of the pathophysiology of depression have not revealed robust biomarkers. The discovery of neurobiological biomarkers may lead to personalized therapies, allowing for precise targeting of the symptoms and potentially leading to a higher response/remission rate. Machine and deep learning became popular and practical approaches in neuroimaging studies, as they can reveal complex patterns in brain organization. However, these methods demand a large amount of available data to yield generalizable and reproducible results. My main goal was to approach MDD with shallow and deep machine learning methods using brain-imaging data of different spatial resolutions.

I analyzed the large multi-site sample from ENIGMA MDD Consortium. Chapter 3 presents the performance benchmark of MDD vs. HC classification of the shallow linear and non-linear models applied to atlas-based volumetric cortical and subcortical features. The extensive analysis revealed a poor classification performance, close to random chance level, of all models when tested on unseen sites, indicating a low accuracy in the most generalizable and realistic settings. By splitting data into CV folds using two different strategies (Splitting by Age/Sex and Splitting by Site), I identified the influence of the site-effect on the classification performance. ComBat removed site-related differences of the models in Splitting by Age/Sex. Further analysis demonstrated that the primary source of the site-effect was the scanner model and acquisition protocol. Lastly, I stratified the data according to demographic (sex) and clinical

(number of recurrent episodes, age of onset, and antidepressant use) factors in order to evaluate if higher classification accuracy can be achieved in more homogeneous samples, revealing no significant improvement in all subgroups.

This analysis was further expanded in Chapter 4, in which I applied machine and deep learning models to a more detailed description of cortical volumetric features – vertex-wise meshes. Furthermore, I analyzed cortical shape characteristics, such as cortical curvature and sulcal depth, integrated with the cortical thickness as the model's inputs. In accordance with the first study, the data was split by using Splitting by Age/Sex and Splitting in Site to estimate the site-effect and to obtain generalizable performance. The performance of both machine learning (SVM) and deep learning (pre-trained Dense-Net) models was close to the random level when data was split according to Splitting by Site. In splitting by Age/Sex, the accuracy reached 58%, revealing the site-effect. This site-effect was addressed via ComBat, resulting in chance accuracy levels of the SVM and DenseNet. Integration of volumetric and shape characteristics did not result in higher accuracies compared to the case when models were trained only on the subset of features, i.e., only curvature, only sulcal depth, and only thickness.

Chapter 5 presents a proof-of-concept demonstration of subject-level parcellation schemes as predictive data domains applied to healthy subjects to predict the global effect of 10 Hz rTMS. The brains were parcellated via the RSFC-Snowballing algorithm and RSFC-Boundary mapping to obtain functional nodes and boundaries. According to our findings, the effect in nodes of a single session of 10 Hz rTMS is the strongest in PCC and precuneus ~30 minutes after the stimulation. The most substantial effects in boundaries occurred in the same regions with a relative to node delay– up to 45 minutes after the stimulation. It is the first study to reveal global changes in functional boundaries after 10 Hz rTMS, emphasizing their role in the organization of the networks.

6.1 Multi-site classification of MDD via machine learning methods on cortical and subcortical measures

In line with other large-sample ML studies by Stolicyn and Flint [142], [143], I observed low, close to random chance classification performance. Current study is the largest in terms of the included sites, allowing for systematic evaluation of the site-related differences and existing biases and their effect on the classification performance. When all sites were equally distributed in CV folds, the balanced accuracy was up to 10% higher than in the Splitting by Site strategy, suggesting that site-related factors mainly drive the highest accuracy. That was additionally confirmed by performing classification with a 1:1 MDD to HC ratio in every site for Splitting

by Age/Sex. In this scenario, the balanced accuracy dropped to random chance. By applying ComBat for both splitting strategies, the classification performances equalized and, unfortunately, were close to the chance level. The results are closer to the classification performance reported by Stolicyn on the large unseen site (UK Biobank). Flint reported a balanced accuracy of up to 60%. However, the site-effect was not directly regressed out in the Flint's analysis; therefore, the accuracy observed in the Flint study could be inflated by the site-effect. This study is the first structural neuroimaging study to address site-effect for different CV splitting. Despite the low classification performance, the most significant classification features majorly overlapped with the findings from large-sample univariate analyses by Schmaal [71]. Crucially, a significant portion of analyzed sites was also included in Schmaal's study.

The consistency of our study with two other large-sample multi-site studies advocates for the application of more complex deep non-linear classification models, such as deep learning models, as shallow machine learning models could not distinguish MDD from HC. Here I encourage the application of deep learning models on cortical and subcortical volumetric features. Furthermore, the consideration of the high-resolution structural data and inclusion of other structural morphometrical characteristics, such as curvature of the brain regions, could be beneficial, as the current resolution and features may not contain enough information to detect MDD. These approaches were considered in Chapter 4 of this thesis.

6.2 Discriminating major depressive disorder on cortical surface-based features: A deep learning approach

A natural response to unsatisfactory results presented in Chapter 3 is to extend the analysis to high-resolution brain morphometric data, including cortical thickness and shape characteristics (sulcal depth and curvature), to achieve higher classification performances. Furthermore, I incorporated a highly non-linear deep classification model – pre-trained DenseNet, hypothesizing further improvement in classification performances. Both linear SVM and DenseNet exhibited a similar range of accuracies when trained on all data modalities combined (sulcal depth, curvature, and thickness): up to 58% in Splitting by Age/Sex and 52% in Splitting by Site. These results indicate an inability of both shallow and deep models to differentiate MDD from HC based on cortical vertex-wise maps. In line with our previous analysis, we observed a trend of higher accuracies of both models in Splitting by Age/Sex compared to Splitting by Site, indicating the presence of the site-effect biasing the decision-

making of classification algorithms. Site-effect was addressed via ComBat, leading to no improvement in classification performance.

The choice of the CNN model was predicated on its previous success in improving the classification accuracy of autism vs. HC by up to 67%, compared to the shallow linear model (SVM) with an accuracy of 58% when applied to similar features [155]. Nevertheless, in this study, I observed no difference between the SVM and the current implementation of the CNN model in the main MDD vs. HC task, suggesting the absence of non-linear patterns of cortical organizations to be altered in the MDD group. However, the considered sample may not be large enough for deep models to outperform shallow ones, as was demonstrated previously for sex and age prediction in an even bigger sample [237].

Overall, the accuracies obtained by both models did not exceed our previous study. Importantly, I analyzed the data provided by ENIGMA MDD Consortium; thus, we had a substantial overlap between the samples from previous and current studies, potentially contributing to a similar range of results. Although CNN was able to outperform SVM in the auxiliary sex classification task in the single-site scenario, the difference in performance between both models can be inflated by the small number of subjects in three considered sites (SHIP_T0, FOR2017 Marburg, and Munster) [142]. When ComBat was applied, the accuracy of CNN in Splitting by Age/Sex did not drop considerably, indicating the presence of non-linear site differences, which CNN detected even after the application of ComBat. This is in line with a recent study by Solanes [274], revealing the limitation of ComBat when applied in combination with non-linear models. It is still an open question if more sophisticated deep learning models may yield higher performance. The application of the ComBat yielded a lower classification performance for both data-splitting strategies and models. The biases and site-related differences could be mitigated better by applying deep learning domain adaptation techniques [344].

Furthermore, the integration of cortical thickness with shape cortical data modalities did not result in substantially higher classification performance than when models were trained on every data type separately. That highlights the absence of shape-thickness interactions in MDD manifestation. As the classification performance of both models was close to random chance in both CV splitting strategies with and without ComBat, there might be a general lack of information in cortical morphometric characteristics to detect depression. Against this, the integration of grey matter morphometric characteristics with cytoarchitecture and functional brain network organization could lead to better differentiability between MDD and HC, as evidenced in a multi-site transdiagnostic study by Hettwer [267].

6.3 Subject-specific whole-brain parcellations of nodes and boundaries are modulated differently under 10Hz rTMS

Functional brain organizations of all individuals are unique. A functional connectome can be considered a 'fingerprint', allowing identifying subjects from a large group [345]. Part of this uniqueness can be lost in atlas-based segmentation methods. I applied two complementary subject-specific parcellation methods to pinpoint the timing and location of rTMS-induced effects. Interestingly, while the location of both nodal and boundary maps was similar, the nodes responded ~15 minutes faster compared to boundaries. As the functions of boundaries are not well understood, I can only speculate if the response in boundaries is associated with the delayed response in nodes or if they have unique properties independent of nodes. Boundary regions may contribute to brain network stability [330] or functional adaptation within hubs during plastic changes [331]. The strongest boundary regions lay in the exterior of functional hubs. Therefore, the response to rTMS should occur in nodal areas first, compared to boundary regions. One could assume that changes in neuroplasticity appear in a similar order, i.e., initially in hubs and further propagating to periphery areas. To test this hypothesis, a high-resolution rs-fMRI study is required to precisely measure the timing and location of TMS-induced neuroplastic changes. In both cases, the alterations in nodes and boundaries for MDD diagnosis may reveal time-related associations, such as the duration of the current depressive episode, or be correlated with the age of onset.

A small sample size demanded a reduction in the number of analyzed features to prevent overfitting. Therefore, I developed the novel ICA-based complementary feature selection models. There are numerous studies applying ICA to extract RSNs and analyze them to unfold MDD-related alterations [111], [346]–[348]. Complementary to RSNs, one could analyze ICA-based boundary masks, representing the transitional areas between different brain networks. In the current approach, resulting ICA-based boundary masks are binary, i.e., the information on which networks intersect is not considered. This algorithm can be easily extended to incorporate network-specific intersections. I validated the stability of both RSFC-Snowballing and RSFC Boundary Mapping, as the inter-subject exceeded intra-subject variability. Similarly, validation schemes are required for ICA-based boundary masks.

6.4 Outlook

In the first part of my work, I investigated how well one can distinguish MDD from HC from structural brain features using shallow and deep machine learning models. I analyzed a sparse set of atlas-based features and a more fine-grained description of the structural brain organization. Furthermore, a broad spectrum of brain morphometric characteristics was investigated, including cortical surface area, thickness, curvature, sulcal depth, and subcortical volumes. To obtain reliable and conclusive results, I analyzed a large multi-site sample provided by ENIGMA MDD Consortium, which provided a highly heterogeneous sample in demographic and clinical factors. Site-related differences were detected by performing the classification analysis by splitting the data according to demographic factors (Splitting by Age/Sex) and site affiliation (Splitting by Site). Site-effect was addressed via ComBat for both strategies.

Overall, neither shallow nor deep machine learning models could differentiate MDD from HC, regardless of the provided datatype and resolution. Considering the amount of analyzed data and careful consideration of demographic and clinical factors, these results, in addition to the results from previous large sample studies [142], [143], indicate the absence of any significant alterations in grey matter structure in the MDD group. However, until now, only a sparse set of non-linear models has been investigated. Further development of deep learning models applied to structural imaging is required to validate the absence of any non-linear structural patterns associated with MDD.

Most deep learning models constructed for image analysis consider natural images with high-level features [349], [350]. Due to established processing steps in image acquisition and brain registration protocols, every location in the pre-processed image corresponds to the same brain area for every subject. While beneficial for natural images, the in-built property of CNN of translationally invariant feature detection could be fruitless for brain images. Applying a graph neural network [272] directly on vertex-wise meshes could yield higher classification performance for the considered task, as it may allow capturing of the unique and meaningful brain organization.

Another possible direction in improving the model's performance and finding clinically significant biomarkers is to subtype depression based on the clinical profiles and structural brain features. To my knowledge, this has not been attempted before. Neurophysiological data-driven subtyping of depression can potentially be succeeded via unsupervised-learning algorithms. The most prominent approach for detecting biotypes of MDD was performed by Drysdale and

colleagues [158]. Using resting-state fMRI, they correlated a linear combination of connectivity features with symptom profiles via canonical correlation analysis (CCA). Two strongest connectivity components were identified: 1) anhedonia combined with psychomotor retardation correlating with frontostriatal and orbitofrontal features, and 2) anxiety combined with insomnia correlating with limbic areas. They identified four clusters based on these two components via hierarchical clustering algorithms. However, these four clusters were not identified in the replication study by Dinga and colleagues [351]. Dinga and colleagues suggested that the lack of reproducibility was due to the absence of a direct evaluation of the cluster stability via cluster significance test and the absence of cross-validation implemented in the analysis, which must include both feature selection and CCA.

A similar study by Tokuda and colleagues for the subtyping of depression found three clusters based on FC features, clinical questionnaire scores, and biological data, such as BDNF, cortisol level, single nucleotide polymorphisms (SNPs), and DNA methylation [352]. The lack of evaluation of the cluster stability suggests that more replication studies should be conducted to validate these findings. A more recent study by Liang and colleagues identified two clusters based on connectivity measures only, with one cluster characterized by decreased FC in DMN, and the second one by increased FC in DMN [353]. Similar to the Drysdale study, there was no direct evaluation of the cluster stability, and feature selection was not explicitly validated in the replication dataset, lowering the probability of the results being replicated in future studies. In general, none of the proposed subtypes of depression have been rigorously validated, nor have such subgrouping been replicated in other studies. Further development in the neurobiologically-driven depression subtyping is required to find stronger structural and functional biomarkers related to MDD.

In the second part of my work, I validated the subject-specific parcellations calculated via RSFC-Snowballing and RSFC-Boundary Mapping as predictive features in a proof-of-concept study. This study revealed the global effect of single session 10Hz rTMS on healthy subjects' functional nodes and boundaries. Both nodal and boundary maps can encapsulate different characteristics of the functional brain organization, as boundaries responded to rTMS with 15 minutes of delay compared to nodes. While nodes are characterized as locations with the highest local and global connectivities, the role and function of boundaries are still underexplored. Our study inspires further extensive investigation of boundaries in particular and the use of subject-specific parcellations in clinical studies.

As I demonstrated the usefulness of subject-specific parcellations as predictive data domains, I encourage the application of the studied parcellations to unfold depression-related functional

brain alterations. They can be performed on the whole-brain level, similar to our study, or the cortical surface, as was first demonstrated in Wig's work [174]. The surface-based analysis allows a straightforward application of 3D-to-2D projection methods to incorporate pre-trained deep learning models, as it was performed on the structural images (Chapter 5). Considering the promising results from the Qin study [164] in differentiating MDD from HC, in which they parceled the whole brain into ROIs according to Doesnbach's atlas [354], subject-specific parcellations may boost the accuracy even further by capturing individual variability in functional brain morphology. Moreover, a fusion of structural and functional data modalities could enhance the analysis even further, as was demonstrated before [267], [268], which may lead to higher classification accuracies and, thus, more reliable biomarkers of depression.

From the molecular level to the large-scale neuroanatomical and functional brain network alterations, the exact mechanisms of the pathophysiology of depression are far from being completely deciphered. The heterogeneity of depression in terms of its clinical manifestations demands the development of new big-data analytical tools able to account for genetic, clinical, and sociodemographic factors. Recently established worldwide consortiums just started to reveal a more realistic picture of depression-related neurobiological alterations. Large collections of samples collected worldwide will enable the integration of different data modalities, which will provide a more holistic view of depression pathophysiology and, hopefully the development of more successful therapies.

Chapter 7 Bibliography

[1] W. H. Organization, "Depression and other common mental disorders: global health estimates," *Depress. Common Ment. Disord. Glob. Health Estim.*, 2017, Accessed: Jun. 20, 2022. [Online]. Available: https://apps.who.int/iris/handle/10665/254610

[2] I. H. Gotlib and J. Joormann, "Cognition and depression: current status and future directions," *Annu. Rev. Clin. Psychol.*, vol. 6, pp. 285–312, 2010, doi: 10.1146/annurev.clinpsy.121208.131305.

[3] C. Otte *et al.*, "Major depressive disorder," *Nat. Rev. Dis. Primer*, vol. 2, no. 1, Art. no. 1, Sep. 2016, doi: 10.1038/nrdp.2016.65.

[4] M. B. First, *DSM-5® Handbook of Differential Diagnosis*. American Psychiatric Publishing, 2013. doi: 10.1176/appi.books.9781585629992.

[5] H. Cai *et al.*, "Prevalence of Suicidality in Major Depressive Disorder: A Systematic Review and Meta-Analysis of Comparative Studies," *Front. Psychiatry*, vol. 12, 2021, Accessed: Feb. 06, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyt.2021.690130

[6] E. Chesney, G. M. Goodwin, and S. Fazel, "Risks of all-cause and suicide mortality in mental disorders: a meta-review," *World Psychiatry Off. J. World Psychiatr. Assoc. WPA*, vol. 13, no. 2, pp. 153–160, Jun. 2014, doi: 10.1002/wps.20128.

[7] S. Seedat *et al.*, "Cross-national associations between gender and mental disorders in the World Health Organization World Mental Health Surveys," *Arch. Gen. Psychiatry*, vol. 66, no. 7, pp. 785–795, Jul. 2009, doi: 10.1001/archgenpsychiatry.2009.36.

[8] S. E. Son and J. T. Kirchner, "Depression in Children and Adolescents," *Am. Fam. Physician*, vol. 62, no. 10, pp. 2297–2308, Nov. 2000.

[9] D. L. Hare, S. R. Toukhsati, P. Johansson, and T. Jaarsma, "Depression and cardiovascular disease: a clinical review," *Eur. Heart J.*, vol. 35, no. 21, pp. 1365–1372, Jun. 2014, doi: 10.1093/eurheartj/eht462.

[10] L. E. Egede and C. Ellis, "Diabetes and depression: Global perspectives," *Diabetes Res. Clin. Pract.*, vol. 87, no. 3, pp. 302–312, Mar. 2010, doi: 10.1016/j.diabres.2010.01.024.

[11] C.-T. Li *et al.*, “Major Depressive Disorder and Stroke Risks: A 9-Year Follow-Up Population-Based, Matched Cohort Study,” *PLoS ONE*, vol. 7, no. 10, p. e46818, Oct. 2012, doi: 10.1371/journal.pone.0046818.
[12] G. M. Goodwin, “Depression and associated physical diseases and symptoms,” *Dialogues Clin. Neurosci.*, vol. 8, no. 2, pp. 259–265, Jun. 2006.
[13] S. G. Hofmann, J. Curtiss, J. K. Carpenter, and S. Kind, “Effect of Treatments for Depression on Quality of Life: A Meta-Analysis,” *Cogn. Behav. Ther.*, vol. 46, no. 4, pp. 265–286, Jun. 2017, doi: 10.1080/16506073.2017.1304445.
[14] P. Cuijpers, A. Stringaris, and M. Wolpert, “Treatment outcomes for depression: challenges and opportunities,” *Lancet Psychiatry*, vol. 7, no. 11, pp. 925–927, Nov. 2020, doi: 10.1016/S2215-0366(20)30036-5.
[15] D. Souery, G. I. Papakostas, and M. H. Trivedi, “Treatment-resistant depression,” *J. Clin. Psychiatry*, vol. 67, no. SUPPL. 6, pp. 16–22, Jul. 2006.
[16] S. G. Kornstein and R. K. Schneider, “Clinical features of treatment-resistant depression,” *J. Clin. Psychiatry*, vol. 62 Suppl 16, pp. 18–25, 2001.
[17] S. H. Lisanby, “Electroconvulsive therapy for depression,” *N. Engl. J. Med.*, vol. 357, no. 19, pp. 1939–1945, Nov. 2007, doi: 10.1056/NEJMct075234.
[18] A. Somani and S. K. Kar, “Efficacy of repetitive transcranial magnetic stimulation in treatment-resistant depression: the evidence thus far,” *Gen. Psychiatry*, vol. 32, no. 4, p. e100074, Aug. 2019, doi: 10.1136/gpsych-2019-100074.
[19] E. J. Cole *et al.*, “Stanford Neuromodulation Therapy (SNT): A Double-Blind Randomized Controlled Trial,” *Am. J. Psychiatry*, vol. 179, no. 2, pp. 132–141, Feb. 2022, doi: 10.1176/appi.ajp.2021.20101429.
[20] J. Mendlewicz, “Towards achieving remission in the treatment of depression,” *Dialogues Clin. Neurosci.*, vol. 10, no. 4, pp. 371–375, Dec. 2008.
[21] E. I. Fried and R. M. Nesse, “Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study,” *J. Affect. Disord.*, vol. 172, pp. 96–102, Feb. 2015, doi: 10.1016/j.jad.2014.10.010.
[22] G. Parker, “Classifying depression: should paradigms lost be regained?,” *Am. J. Psychiatry*, vol. 157, no. 8, pp. 1195–1203, Aug. 2000, doi: 10.1176/appi.ajp.157.8.1195.
[23] R. Musil *et al.*, “Subtypes of depression and their overlap in a naturalistic inpatient sample of major depressive disorder,” *Int. J. Methods Psychiatr. Res.*, vol. 27, no. 1, p. e1569, Jun. 2017, doi: 10.1002/mpr.1569.
[24] K. K. Jain, “Personalized medicine,” *Curr. Opin. Mol. Ther.*, vol. 4, no. 6, pp. 548–558, Dec. 2002.
[25] D. L. Jardim *et al.*, “Impact of a Biomarker-Based Strategy on Oncology Drug Development: A Meta-analysis of Clinical Trials Leading to FDA Approval,” *J. Natl. Cancer Inst.*, vol. 107, no. 11, p. djv253, Nov. 2015, doi: 10.1093/jnci/djv253.
[26] M. Schwaederle *et al.*, “Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 33, no. 32, pp. 3817–3825, Nov. 2015, doi: 10.1200/JCO.2015.61.5997.
[27] N. Matussek, “Biochemie der Depression,” *J. Neural Transm.*, vol. 33, no. 3, pp. 223–234, Sep. 1972, doi: 10.1007/BF01245319.
[28] B. W. Dunlop and C. B. Nemeroff, “The Role of Dopamine in the Pathophysiology of Depression,” *Arch. Gen. Psychiatry*, vol. 64, no. 3, pp. 327–337, Mar. 2007, doi: 10.1001/archpsyc.64.3.327.
[29] H. Takano, “Cognitive Function and Monoamine Neurotransmission in Schizophrenia: Evidence From Positron Emission Tomography Studies,” *Front. Psychiatry*, vol. 9, 2018, Accessed: Nov. 29, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyt.2018.00228

[30] G. Hache, F. Coudore, A. M. Gardier, and B. P. Guiard, “Monoaminergic Antidepressants in the Relief of Pain: Potential Therapeutic Utility of Triple Reuptake Inhibitors (TRIs),” *Pharmaceuticals*, vol. 4, no. 2, Art. no. 2, Feb. 2011, doi: 10.3390/ph4020285.
[31] P. Willner, “The mesolimbic dopamine system as a target for rapid antidepressant action,” *Int. Clin. Psychopharmacol.*, vol. 12 Suppl 3, pp. S7-14, Jul. 1997, doi: 10.1097/00004850-199707003-00002.
[32] C. H. Lammers, J. Diaz, J. C. Schwartz, and P. Sokoloff, “Selective increase of dopamine D3 receptor gene expression as a common effect of chronic antidepressant treatments,” *Mol. Psychiatry*, vol. 5, no. 4, pp. 378–388, Jul. 2000, doi: 10.1038/sj.mp.4000754.
[33] L. Rampello, G. Nicoletti, and R. Raffaele, “Dopaminergic hypothesis for retarded depression: a symptom profile for predicting therapeutical responses,” *Acta Psychiatr. Scand.*, vol. 84, no. 6, pp. 552–554, Dec. 1991, doi: 10.1111/j.1600-0447.1991.tb03193.x.
[34] G. Racagni and M. Popoli, “Cellular and molecular mechanisms in the long-term action of antidepressants.,” *Dialogues Clin. Neurosci.*, vol. 10, no. 4, pp. 385–400, Dec. 2008, doi: 10.31887/DCNS.2008.10.4/gracagni.
[35] K. S. Kendler, L. M. Karkowski, and C. A. Prescott, “Causal relationship between stressful life events and the onset of major depression,” *Am. J. Psychiatry*, vol. 156, no. 6, pp. 837–841, Jun. 1999, doi: 10.1176/ajp.156.6.837.
[36] G. Asadikaram *et al.*, “Assessment of hormonal alterations in major depressive disorder: A clinical study,” *PsyCh J.*, vol. 8, no. 4, pp. 423–430, Dec. 2019, doi: 10.1002/pchj.290.
[37] E. J. Susman, K. H. Schmeelk, B. K. Worrall, D. A. Granger, A. Ponirakis, and G. P. Chrousos, “Corticotropin-releasing hormone and cortisol: longitudinal associations with depression and antisocial behavior in pregnant adolescents,” *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 38, no. 4, pp. 460–467, Apr. 1999, doi: 10.1097/00004583-199904000-00020.
[38] R. T. Rubin, R. E. Poland, I. M. Lesser, and D. J. Martin, “Neuroendocrine aspects of primary endogenous depression. V. Serum prolactin measures in patients and matched control subjects,” *Biol. Psychiatry*, vol. 25, no. 1, pp. 4–21, Jan. 1989, doi: 10.1016/0006-3223(89)90142-x.
[39] H. M. Burke, M. C. Davis, C. Otte, and D. C. Mohr, “Depression and cortisol responses to psychological stress: A meta-analysis,” *Psychoneuroendocrinology*, vol. 30, no. 9, pp. 846–856, Oct. 2005, doi: 10.1016/j.psyneuen.2005.02.010.
[40] Y. Zhou *et al.*, “Comparison of Thyroid Hormone Levels Between Patients With Major Depressive Disorder and Healthy Individuals in China,” *Front. Psychiatry*, vol. 12, p. 750749, Oct. 2021, doi: 10.3389/fpsyt.2021.750749.
[41] S. Terbeck, F. Akkus, L. P. Chesterman, and G. Hasler, “The role of metabotropic glutamate receptor 5 in the pathogenesis of mood disorders and addiction: combining preclinical evidence with human Positron Emission Tomography (PET) studies,” *Front. Neurosci.*, vol. 9, 2015, Accessed: Jun. 22, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2015.00086
[42] R. Aronson, H. J. Offman, R. T. Joffe, and C. D. Naylor, “Triiodothyronine Augmentation in the Treatment of Refractory Depression: A Meta-analysis,” *Arch. Gen. Psychiatry*, vol. 53, no. 9, pp. 842–848, Sep. 1996, doi: 10.1001/archpsyc.1996.01830090090013.
[43] J. T. Gordon, D. M. Kaminski, C. B. Rozanov, and M. B. Dratman, “Evidence that 3,3’,5-triiodothyronine is concentrated in and delivered from the locus coeruleus to its noradrenergic targets via anterograde axonal transport,” *Neuroscience*, vol. 93, no. 3, pp. 943–954, 1999, doi: 10.1016/s0306-4522(99)00146-3.

[44] M. Bloch, P. J. Schmidt, M. Danaceau, J. Murphy, L. Nieman, and D. R. Rubinow, "Effects of gonadal steroids in women with a history of postpartum depression," *Am. J. Psychiatry*, vol. 157, no. 6, pp. 924–930, Jun. 2000, doi: 10.1176/appi.ajp.157.6.924.
[45] R. E. Noble, "Depression in women," *Metabolism*, vol. 54, no. 5, Supplement, pp. 49–52, May 2005, doi: 10.1016/j.metabol.2005.01.014.
[46] L. A. Mamounas, M. E. Blue, J. A. Siuciak, and C. A. Altar, "Brain-derived neurotrophic factor promotes the survival and sprouting of serotonergic axons in rat brain," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 15, no. 12, pp. 7929–7939, Dec. 1995.
[47] F. Karege, G. Perret, G. Bondolfi, M. Schwald, G. Bertschy, and J.-M. Aubry, "Decreased serum brain-derived neurotrophic factor levels in major depressed patients," *Psychiatry Res.*, vol. 109, no. 2, pp. 143–148, Mar. 2002, doi: 10.1016/s0165-1781(02)00005-7.
[48] T. Hartley *et al.*, "The hippocampus is required for short-term topographical memory in humans," *Hippocampus*, vol. 17, no. 1, pp. 34–48, 2007, doi: 10.1002/hipo.20240.
[49] M. Remondes and E. M. Schuman, "Role for a cortical input to hippocampal area CA1 in the consolidation of a long-term memory," *Nature*, vol. 431, no. 7009, Art. no. 7009, Oct. 2004, doi: 10.1038/nature02965.
[50] J. R. Whitlock, A. J. Heynen, M. G. Shuler, and M. F. Bear, "Learning Induces Long-Term Potentiation in the Hippocampus," *Science*, vol. 313, no. 5790, pp. 1093–1097, Aug. 2006, doi: 10.1126/science.1128134.
[51] R. S. Duman, G. R. Heninger, and E. J. Nestler, "A molecular and cellular theory of depression," *Arch. Gen. Psychiatry*, vol. 54, no. 7, pp. 597–606, Jul. 1997, doi: 10.1001/archpsyc.1997.01830190015002.
[52] Y. I. Sheline, M. H. Gado, and H. C. Kraemer, "Untreated depression and hippocampal volume loss," *Am. J. Psychiatry*, vol. 160, no. 8, pp. 1516–1518, Aug. 2003, doi: 10.1176/appi.ajp.160.8.1516.
[53] O. Berton *et al.*, "Essential role of BDNF in the mesolimbic dopamine pathway in social defeat stress," *Science*, vol. 311, no. 5762, pp. 864–868, Feb. 2006, doi: 10.1126/science.1120972.
[54] H. Miyanishi and A. Nitta, "A Role of BDNF in the Depression Pathogenesis and a Potential Target as Antidepressant: The Modulator of Stress Sensitivity 'Shati/Nat8l-BDNF System' in the Dorsal Striatum," *Pharmaceuticals*, vol. 14, no. 9, p. 889, Sep. 2021, doi: 10.3390/ph14090889.
[55] A. Bird, "Understanding the Replication Crisis as a Base Rate Fallacy," *Br. J. Philos. Sci.*, vol. 72, no. 4, pp. 965–993, Dec. 2021, doi: 10.1093/bjps/axy051.
[56] E. Lin and A. Alessio, "What are the basic concepts of temporal, contrast, and spatial resolution in cardiac CT?," *J. Cardiovasc. Comput. Tomogr.*, vol. 3, no. 6, pp. 403–408, 2009, doi: 10.1016/j.jcct.2009.07.003.
[57] J. R. Haaga and D. Boll, *CT and MRI of the Whole Body*. Elsevier Health Sciences, 2016.
[58] B. P. Thomas *et al.*, "High-resolution 7T MRI of the human hippocampus in vivo," *J. Magn. Reson. Imaging*, vol. 28, no. 5, pp. 1266–1272, 2008, doi: 10.1002/jmri.21576.
[59] D. A. Axelson *et al.*, "Hypercortisolemia and hippocampal changes in depression," *Psychiatry Res.*, vol. 47, no. 2, pp. 163–173, May 1993, doi: 10.1016/0165-1781(93)90046-J.
[60] C. E. Coffey *et al.*, "Quantitative Cerebral Anatomy in Depression: A Controlled Magnetic Resonance Imaging Study," *Arch. Gen. Psychiatry*, vol. 50, no. 1, pp. 7–16, Jan. 1993, doi: 10.1001/archpsyc.1993.01820130009002.
[61] G. S. Alexopoulos, R. C. Young, and R. D. Shindledecker, "Brain computed tomography findings in geriatric depression and primary degenerative dementia," *Biol. Psychiatry*, vol. 31, no. 6, pp. 591–599, Mar. 1992, doi: 10.1016/0006-3223(92)90245-U.

[62] J. C. Soares and J. J. Mann, “The anatomy of mood disorders--review of structural neuroimaging studies,” *Biol. Psychiatry*, vol. 41, no. 1, pp. 86–106, Jan. 1997, doi: 10.1016/s0006-3223(96)00006-6.
[63] X. Zhang, S. Yao, X. Zhu, X. Wang, X. Zhu, and M. Zhong, “Gray matter volume abnormalities in individuals with cognitive vulnerability to depression: A voxel-based morphometry study,” *J. Affect. Disord.*, vol. 136, no. 3, pp. 443–452, Feb. 2012, doi: 10.1016/j.jad.2011.11.005.
[64] M. Serra-Blasco *et al.*, “Effects of illness duration and treatment resistance on grey matter abnormalities in major depression,” *Br. J. Psychiatry J. Ment. Sci.*, vol. 202, pp. 434–440, Jun. 2013, doi: 10.1192/bjp.bp.112.116228.
[65] H. S. Mayberg *et al.*, “Cingulate function in depression: a potential predictor of treatment response,” *NeuroReport*, vol. 8, no. 4, pp. 1057–1061, Mar. 1997.
[66] J. L. Phillips, L. A. Batten, P. Tremblay, F. Aldosary, and P. Blier, “A Prospective, Longitudinal Study of the Effect of Remission on Cortical Thickness and Hippocampal Volume in Patients with Treatment-Resistant Depression,” *Int. J. Neuropsychopharmacol.*, vol. 18, no. 8, p. pyv037, Mar. 2015, doi: 10.1093/ijnp/pyv037.
[67] H. Järnum *et al.*, “Longitudinal MRI study of cortical thickness, perfusion, and metabolite levels in major depressive disorder,” *Acta Psychiatr. Scand.*, vol. 124, no. 6, pp. 435–446, 2011, doi: 10.1111/j.1600-0447.2011.01766.x.
[68] A. Bechara, H. Damasio, D. Tranel, and S. W. Anderson, “Dissociation Of working memory from decision making within the human prefrontal cortex,” *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 18, no. 1, pp. 428–437, Jan. 1998.
[69] A. Bechara, S. Dolan, N. Denburg, A. Hindes, S. W. Anderson, and P. E. Nathan, “Decision-making deficits, linked to a dysfunctional ventromedial prefrontal cortex, revealed in alcohol and stimulant abusers,” *Neuropsychologia*, vol. 39, no. 4, pp. 376–389, 2001, doi: 10.1016/s0028-3932(00)00136-6.
[70] H. A. Berlin, E. T. Rolls, and U. Kischka, “Impulsivity, time perception, emotion and reinforcement sensitivity in patients with orbitofrontal cortex lesions,” *Brain J. Neurol.*, vol. 127, no. Pt 5, pp. 1108–1126, May 2004, doi: 10.1093/brain/awh135.
[71] L. Schmaal *et al.*, “Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group,” *Mol. Psychiatry*, vol. 22, no. 6, pp. 900–909, 2017, doi: 10.1038/mp.2016.60.
[72] M. J. Lan, B. T. Chhetry, C. Liston, J. J. Mann, and M. Dubin, “Transcranial Magnetic Stimulation of Left Dorsolateral Prefrontal Cortex Induces Brain Morphological Changes in Regions Associated with a Treatment Resistant Major Depressive Episode: An Exploratory Analysis,” *Brain Stimulat.*, vol. 9, no. 4, pp. 577–583, Aug. 2016, doi: 10.1016/j.brs.2016.02.011.
[73] X. Yang *et al.*, “Increased prefrontal and parietal cortical thickness does not correlate with anhedonia in patients with untreated first-episode major depressive disorders,” *Psychiatry Res.*, vol. 234, no. 1, pp. 144–151, Oct. 2015, doi: 10.1016/j.pscychresns.2015.09.014.
[74] Z. Chen *et al.*, “High-field magnetic resonance imaging of structural alterations in first-episode, drug-naive patients with major depressive disorder,” *Transl. Psychiatry*, vol. 6, no. 11, p. e942, Nov. 2016, doi: 10.1038/tp.2016.209.
[75] C.-G. Yan *et al.*, “Reduced default mode network functional connectivity in patients with recurrent major depressive disorder,” *Proc. Natl. Acad. Sci.*, vol. 116, no. 18, pp. 9078–9083, Apr. 2019, doi: 10.1073/pnas.1900390116.
[76] M. Smith, “Chapter 3 - Brain mapping,” in *Mechanisms and Genetics of Neurodevelopmental Cognitive Disorders*, M. Smith, Ed. Academic Press, 2021, pp. 49–76. doi: 10.1016/B978-0-12-821913-3.00004-4.

[77] W. Peng *et al.*, “Brain structural abnormalities in emotional regulation and sensory processing regions associated with anxious depression,” *Prog. Neuropsychopharmacol. Biol. Psychiatry*, vol. 94, p. 109676, Aug. 2019, doi: 10.1016/j.pnpbp.2019.109676.
[78] L. Sun *et al.*, “Human anterior thalamic nuclei are involved in emotion–attention interaction,” *Neuropsychologia*, vol. 78, pp. 88–94, Nov. 2015, doi: 10.1016/j.neuropsychologia.2015.10.001.
[79] Y. D. Van Der Werf *et al.*, “Thalamic volume predicts performance on tests of cognitive speed and decreases in healthy aging. A magnetic resonance imaging-based volumetric analysis,” *Brain Res. Cogn. Brain Res.*, vol. 11, no. 3, pp. 377–385, Jun. 2001, doi: 10.1016/s0926-6410(01)00010-6.
[80] K. E. Krout, R. E. Belzer, and A. D. Loewy, “Brainstem projections to midline and intralaminar thalamic nuclei of the rat,” *J. Comp. Neurol.*, vol. 448, no. 1, pp. 53–101, Jun. 2002, doi: 10.1002/cne.10236.
[81] S. Tekin and J. L. Cummings, “Frontal-subcortical neuronal circuits and clinical neuropsychiatry: an update,” *J. Psychosom. Res.*, vol. 53, no. 2, pp. 647–654, Aug. 2002, doi: 10.1016/s0022-3999(02)00428-2.
[82] L. Qiu *et al.*, “Regional increases of cortical thickness in untreated, first-episode major depressive disorder,” *Transl. Psychiatry*, vol. 4, p. e378, Apr. 2014, doi: 10.1038/tp.2014.18.
[83] Y.-J. Zhao *et al.*, “Brain grey matter abnormalities in medication-free patients with major depressive disorder: a meta-analysis,” *Psychol. Med.*, vol. 44, no. 14, pp. 2927–2937, Oct. 2014, doi: 10.1017/S0033291714000518.
[84] Y. Lu *et al.*, “The volumetric and shape changes of the putamen and thalamus in first episode, untreated major depressive disorder,” *NeuroImage Clin.*, vol. 11, pp. 658–666, 2016, doi: 10.1016/j.nicl.2016.04.008.
[85] M. Bellani, M. Baiano, and P. Brambilla, “Brain anatomy of major depression II. Focus on amygdala,” *Epidemiol. Psychiatr. Sci.*, vol. 20, no. 1, pp. 33–36, Mar. 2011, doi: 10.1017/S2045796011000096.
[86] Y. I. Sheline, M. H. Gado, and J. L. Price, “Amygdala core nuclei volumes are decreased in recurrent major depression,” *NeuroReport*, vol. 9, no. 9, pp. 2023–2028, Jun. 1998.
[87] I. B. Hickie *et al.*, “Serotonin transporter gene status predicts caudate nucleus but not amygdala or hippocampal volumes in older persons with major depression,” *J. Affect. Disord.*, vol. 98, no. 1, pp. 137–142, Feb. 2007, doi: 10.1016/j.jad.2006.07.010.
[88] Y. Tang *et al.*, “Reduced ventral anterior cingulate and amygdala volumes in medication-naïve females with major depressive disorder: A voxel-based morphometric magnetic resonance imaging study,” *Psychiatry Res. Neuroimaging*, vol. 156, no. 1, pp. 83–86, Oct. 2007, doi: 10.1016/j.pscychresns.2007.03.005.
[89] E. Mervaala *et al.*, “Quantitative MRI of the hippocampus and amygdala in severe depression,” *Psychol. Med.*, vol. 30, no. 1, pp. 117–125, Jan. 2000, doi: 10.1017/S0033291799001567.
[90] C. Lange and E. Irle, “Enlarged amygdala volume and reduced hippocampal volume in young women with major depression,” *Psychol. Med.*, vol. 34, no. 6, pp. 1059–1064, Aug. 2004, doi: 10.1017/S0033291703001806.
[91] G. Weniger, C. Lange, and E. Irle, “Abnormal size of the amygdala predicts impaired emotional memory in major depressive disorder,” *J. Affect. Disord.*, vol. 94, no. 1, pp. 219–229, Aug. 2006, doi: 10.1016/j.jad.2006.04.017.
[92] L. Schmaal *et al.*, “Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group,” *Mol. Psychiatry*, vol. 21, no. 6, pp. 806–812, Jun. 2016, doi: 10.1038/mp.2015.69.

[93] T. C. Ho *et al.*, “Subcortical shape alterations in major depressive disorder: Findings from the ENIGMA major depressive disorder working group,” *Hum. Brain Mapp.*, vol. 43, no. 1, pp. 341–351, 2022, doi: 10.1002/hbm.24988.
[94] Z. Yao *et al.*, “Morphological changes in subregions of hippocampus and amygdala in major depressive disorder patients,” *Brain Imaging Behav.*, vol. 14, no. 3, pp. 653–667, Jun. 2020, doi: 10.1007/s11682-018-0003-1.
[95] Y. Liu *et al.*, “Morphometry of the Hippocampus Across the Adult Life-Span in Patients with Depressive Disorders: Association with Neuroticism,” *Brain Topogr.*, vol. 34, Sep. 2021, doi: 10.1007/s10548-021-00846-0.
[96] S. Marek *et al.*, “Reproducible brain-wide association studies require thousands of individuals,” *Nature*, vol. 603, no. 7902, Art. no. 7902, Mar. 2022, doi: 10.1038/s41586-022-04492-9.
[97] N. R. Winter *et al.*, “Quantifying Deviations of Brain Structure and Function in Major Depressive Disorder Across Neuroimaging Modalities,” *JAMA Psychiatry*, vol. 79, no. 9, pp. 879–888, Sep. 2022, doi: 10.1001/jamapsychiatry.2022.1780.
[98] G. H. Glover, “Overview of Functional Magnetic Resonance Imaging,” *Neurosurg. Clin. N. Am.*, vol. 22, no. 2, pp. 133–139, Apr. 2011, doi: 10.1016/j.nec.2010.11.001.
[99] S. Ogawa, T.-M. Lee, R. Stepnoski, W. Chen, X.-H. Zhu, and K. Ugurbil, “An approach to probe some neural systems interaction by functional MRI at neural time scale down to milliseconds,” *Proc. Natl. Acad. Sci.*, vol. 97, no. 20, pp. 11026–11031, Sep. 2000, doi: 10.1073/pnas.97.20.11026.
[100] K. J. Friston, “Functional and effective connectivity in neuroimaging: A synthesis,” *Hum. Brain Mapp.*, vol. 2, no. 1–2, pp. 56–78, 1994, doi: 10.1002/hbm.460020107.
[101] S. E. Petersen and O. Sporns, “Brain Networks and Cognitive Architectures,” *Neuron*, vol. 88, no. 1, pp. 207–219, Oct. 2015, doi: 10.1016/j.neuron.2015.09.027.
[102] M. M. Mesulam, “Large-scale neurocognitive networks and distributed processing for attention, language, and memory,” *Ann. Neurol.*, vol. 28, no. 5, pp. 597–613, Nov. 1990, doi: 10.1002/ana.410280502.
[103] B. T. Thomas Yeo *et al.*, “The organization of the human cerebral cortex estimated by intrinsic functional connectivity,” *J. Neurophysiol.*, vol. 106, no. 3, pp. 1125–1165, Sep. 2011, doi: 10.1152/jn.00338.2011.
[104] B. Biswal, F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde, “Functional connectivity in the motor cortex of resting human brain using echo-planar mri,” *Magn. Reson. Med.*, vol. 34, no. 4, pp. 537–541, 1995, doi: 10.1002/mrm.1910340409.
[105] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman, “A default mode of brain function,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 2, pp. 676–682, Jan. 2001, doi: 10.1073/pnas.98.2.676.
[106] C. L. Grady, A. R. McIntosh, S. Beig, M. L. Keightley, H. Burian, and S. E. Black, “Evidence from functional neuroimaging of a compensatory prefrontal network in Alzheimer’s disease,” *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 23, no. 3, pp. 986–993, Feb. 2003.
[107] D. Jandric *et al.*, “Mechanisms of Network Changes in Cognitive Impairment in Multiple Sclerosis,” *Neurology*, vol. 97, no. 19, pp. e1886–e1897, Nov. 2021, doi: 10.1212/WNL.0000000000012834.
[108] A. K. Rehme and C. Grefkes, “Cerebral network disorders after stroke: evidence from imaging-based connectivity analyses of active and resting brain states in humans,” *J. Physiol.*, vol. 591, no. Pt 1, pp. 17–31, Jan. 2013, doi: 10.1113/jphysiol.2012.243469.
[109] R. H. Kaiser, J. R. Andrews-Hanna, T. D. Wager, and D. A. Pizzagalli, “Large-Scale Network Dysfunction in Major Depressive Disorder: A Meta-analysis of Resting-State Functional Connectivity,” *JAMA Psychiatry*, vol. 72, no. 6, pp. 603–611, Jun. 2015, doi: 10.1001/jamapsychiatry.2015.0071.

[110] G. Wagner, C. Schachtzabel, G. Peikert, and K.-J. Bär, "The neural basis of the abnormal self-referential processing and its impact on cognitive control in depressed patients," *Hum. Brain Mapp.*, vol. 36, no. 7, pp. 2781–2794, Jul. 2015, doi: 10.1002/hbm.22807.
[111] M. D. Greicius *et al.*, "Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus," *Biol. Psychiatry*, vol. 62, no. 5, pp. 429–437, Sep. 2007, doi: 10.1016/j.biopsych.2006.09.020.
[112] Y. Zhou *et al.*, "Increased neural resources recruitment in the intrinsic organization in major depression," *J. Affect. Disord.*, vol. 121, no. 3, pp. 220–230, Mar. 2010, doi: 10.1016/j.jad.2009.05.029.
[113] C. G. Connolly *et al.*, "Resting-State Functional Connectivity of Subgenual Anterior Cingulate Cortex in Depressed Adolescents," *Biol. Psychiatry*, vol. 74, no. 12, pp. 898–907, Dec. 2013, doi: 10.1016/j.biopsych.2013.05.036.
[114] F. Collette and M. Van der Linden, "Brain imaging of the central executive component of working memory," *Neurosci. Biobehav. Rev.*, vol. 26, no. 2, pp. 105–125, Mar. 2002, doi: 10.1016/S0149-7634(01)00063-X.
[115] M. Hampson, N. Driesen, J. K. Roth, J. C. Gore, and R. T. Constable, "Functional connectivity between task-positive and task-negative brain areas and its relation to working memory performance," *Magn. Reson. Imaging*, vol. 28, no. 8, pp. 1051–1057, Oct. 2010, doi: 10.1016/j.mri.2010.03.021.
[116] M. Koenigs and J. Grafman, "The functional neuroanatomy of depression: Distinct roles for ventromedial and dorsolateral prefrontal cortex," *Behav. Brain Res.*, vol. 201, no. 2, pp. 239–243, Aug. 2009, doi: 10.1016/j.bbr.2009.03.004.
[117] L. Bartova *et al.*, "Reduced default mode network suppression during a working memory task in remitted major depression," *J. Psychiatr. Res.*, vol. 64, pp. 9–18, May 2015, doi: 10.1016/j.jpsychires.2015.02.025.
[118] V. Menon and L. Q. Uddin, "Saliency, switching, attention and control: a network model of insula function," *Brain Struct. Funct.*, vol. 214, no. 5, pp. 655–667, Jun. 2010, doi: 10.1007/s00429-010-0262-0.
[119] V. Menon, "Large-scale brain networks and psychopathology: a unifying triple network model," *Trends Cogn. Sci.*, vol. 15, no. 10, pp. 483–506, Oct. 2011, doi: 10.1016/j.tics.2011.08.003.
[120] D. Sridharan, D. J. Levitin, and V. Menon, "A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks," *Proc. Natl. Acad. Sci.*, vol. 105, no. 34, pp. 12569–12574, Aug. 2008, doi: 10.1073/pnas.0800005105.
[121] N. Goulden *et al.*, "The salience network is responsible for switching between the default mode network and the central executive network: Replication from DCM," *NeuroImage*, vol. 99, pp. 180–190, Oct. 2014, doi: 10.1016/j.neuroimage.2014.05.052.
[122] E. T. Rolls, "Limbic systems for emotion and for memory, but no single limbic system," *Cortex J. Devoted Study Nerv. Syst. Behav.*, vol. 62, pp. 119–157, Jan. 2015, doi: 10.1016/j.cortex.2013.12.005.
[123] A. M. B. Milne, G. M. MacQueen, and G. B. C. Hall, "Abnormal hippocampal activation in patients with extensive history of major depression: an fMRI study," *J. Psychiatry Neurosci. JPN*, vol. 37, no. 1, pp. 28–36, Jan. 2012, doi: 10.1503/jpn.110004.
[124] T. Shen *et al.*, "Altered spontaneous neural activity in first-episode, unmedicated patients with major depressive disorder," *Neuroreport*, vol. 25, no. 16, pp. 1302–1307, Nov. 2014, doi: 10.1097/WNR.0000000000000263.
[125] Z. Y. Hao *et al.*, "Abnormal resting-state functional connectivity of hippocampal subfields in patients with major depressive disorder," *BMC Psychiatry*, vol. 20, no. 1, p. 71, Feb. 2020, doi: 10.1186/s12888-020-02490-7.

[126] N. Javaheripour *et al.*, "Altered resting-state functional connectome in major depressive disorder: a mega-analysis from the PsyMRI consortium," *Transl. Psychiatry*, vol. 11, no. 1, Art. no. 1, Oct. 2021, doi: 10.1038/s41398-021-01619-w.
[127] S. Gao, V. D. Calhoun, and J. Sui, "Machine learning in major depression: From classification to treatment outcome prediction," *CNS Neurosci. Ther.*, vol. 24, no. 11, pp. 1037–1052, 2018, doi: 10.1111/cns.13048.
[128] N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan, and C. I. Baker, "Circular analysis in systems neuroscience: the dangers of double dipping," *Nat. Neurosci.*, vol. 12, no. 5, Art. no. 5, May 2009, doi: 10.1038/nn.2303.
[129] J. Wen *et al.*, "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Med. Image Anal.*, vol. 63, p. 101694, Jul. 2020, doi: 10.1016/j.media.2020.101694.
[130] J. Dockès, G. Varoquaux, and J.-B. Poline, "Preventing dataset shift from breaking machine-learning biomarkers," arXiv, arXiv:2107.09947, Jul. 2021. Accessed: May 20, 2022. [Online]. Available: http://arxiv.org/abs/2107.09947
[131] C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager, "Building better biomarkers: brain models in translational neuroimaging," *Nat. Neurosci.*, vol. 20, no. 3, Art. no. 3, Mar. 2017, doi: 10.1038/nn.4478.
[132] M. A. Little *et al.*, "Using and understanding cross-validation strategies. Perspectives on Saeb et al.," *GigaScience*, vol. 6, no. 5, p. gix020, May 2017, doi: 10.1093/gigascience/gix020.
[133] D. M. Hawkins, "The Problem of Overfitting," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 1–12, Jan. 2004, doi: 10.1021/ci0342472.
[134] D. Bashir, G. D. Montañez, S. Sehra, P. S. Segura, and J. Lauw, "An Information-Theoretic Perspective on Overfitting and Underfitting," in *AI 2020: Advances in Artificial Intelligence*, Cham, 2020, pp. 347–358. doi: 10.1007/978-3-030-64984-5_27.
[135] M. Balasubramanian and E. L. Schwartz, "The Isomap Algorithm and Topological Stability," *Science*, Jan. 2002, doi: 10.1126/science.295.5552.7a.
[136] van der Maaten, L.J.P. and Hinton, G.E., "Visualizing High-Dimensional Data Using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. nov, pp. 2579–2605, 2008.
[137] R. Dinga, B. Penninx, D. Veltman, L. Schmaal, and A. Marquand, *Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines*. 2019. doi: 10.1101/743138.
[138] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[139] M. Seeger, "Gaussian processes for machine learning," *Int. J. Neural Syst.*, vol. 14, no. 02, pp. 69–106, Apr. 2004, doi: 10.1142/S0129065704001899.
[140] M. Pontil, "Leave-one-out error and stability of learning algorithms with applications," *NATO Sci. Ser. SUB Ser. III …*, Jan. 2003, Accessed: Jul. 18, 2022. [Online]. Available: https://www.academia.edu/1107488/Leave_one_out_error_and_stability_of_learning_algorithms_with_applications
[141] S. Gao *et al.*, "Discriminating bipolar disorder from major depression based on kernel SVM using functional independent components," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2017, pp. 1–6. doi: 10.1109/MLSP.2017.8168110.
[142] C. Flint *et al.*, "Systematic misestimation of machine learning performance in neuroimaging studies of depression," *Neuropsychopharmacology*, pp. 1–8, May 2021, doi: 10.1038/s41386-021-01020-7.

[143] A. Stolicyn *et al.*, "Automated classification of depression from structural brain measures across two independent community-based cohorts," *Hum. Brain Mapp.*, vol. 41, no. 14, pp. 3922–3937, Oct. 2020, doi: 10.1002/hbm.25095.
[144] T. Habota *et al.*, "Cohort profile for the STratifying Resilience and Depression Longitudinally (STRADL) study: A depression-focused investigation of Generation Scotland, using detailed clinical, cognitive, and neuroimaging assessments," Wellcome Open Research, 4:185, Jul. 2021. doi: 10.12688/wellcomeopenres.15538.2.
[145] L. K. M. Han *et al.*, "Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group," *Mol. Psychiatry*, vol. 26, no. 9, pp. 5124–5139, Sep. 2021, doi: 10.1038/s41380-020-0754-0.
[146] X. Yang *et al.*, "Sex differences in the clinical characteristics and brain gray matter volume alterations in unmedicated patients with major depressive disorder," *Sci. Rep.*, vol. 7, p. 2515, May 2017, doi: 10.1038/s41598-017-02828-4.
[147] L. Snoek, S. Miletić, and H. S. Scholte, "How to control for confounds in decoding analyses of neuroimaging data," *NeuroImage*, vol. 184, pp. 741–760, Jan. 2019, doi: 10.1016/j.neuroimage.2018.09.074.
[148] R. Dinga, L. Schmaal, B. W. J. H. Penninx, D. J. Veltman, and A. F. Marquand, "Controlling for effects of confounding variables on machine learning predictions," *bioRxiv*, p. 2020.08.17.255034, Aug. 2020, doi: 10.1101/2020.08.17.255034.
[149] W. H. L. Pinaya, A. Mechelli, and J. R. Sato, "Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study," *Hum. Brain Mapp.*, vol. 40, no. 3, pp. 944–954, Oct. 2018, doi: 10.1002/hbm.24423.
[150] A. Solanes *et al.*, "Biased accuracy in multisite machine-learning studies due to incomplete removal of the effects of the site," *Psychiatry Res. Neuroimaging*, vol. 314, p. 111313, Aug. 2021, doi: 10.1016/j.pscychresns.2021.111313.
[151] F. Orlhac *et al.*, "A guide to ComBat harmonization of imaging biomarkers in multicenter studies," *J. Nucl. Med.*, Sep. 2021, doi: 10.2967/jnumed.121.262464.
[152] M. S. Depping *et al.*, "Common and distinct patterns of abnormal cortical gyrification in major depression and borderline personality disorder," *Eur. Neuropsychopharmacol.*, vol. 28, no. 10, pp. 1115–1125, Oct. 2018, doi: 10.1016/j.euroneuro.2018.07.100.
[153] Y. Zhang, C. Yu, Y. Zhou, K. Li, C. Li, and T. Jiang, "Decreased gyrification in major depressive disorder," *Neuroreport*, vol. 20, no. 4, pp. 378–380, Mar. 2009, doi: 10.1097/WNR.0b013e3283249b34.
[154] D. C. Van Essen *et al.*, "The Human Connectome Project: a data acquisition perspective," *NeuroImage*, vol. 62, no. 4, pp. 2222–2231, Oct. 2012, doi: 10.1016/j.neuroimage.2012.02.018.
[155] K. Gao *et al.*, "Deep Transfer Learning for Cerebral Cortex Using Area-Preserving Geometry Mapping," *Cereb. Cortex*, Nov. 2021, doi: 10.1093/cercor/bhab394.
[156] A. Abrol *et al.*, "Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning," *Nat. Commun.*, vol. 12, no. 1, Art. no. 1, Jan. 2021, doi: 10.1038/s41467-020-20655-6.
[157] D. Szucs and J. P. Ioannidis, "Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990-2012) and of latest practices (2017-2018) in high-impact journals," *NeuroImage*, vol. 221, p. 117164, Nov. 2020, doi: 10.1016/j.neuroimage.2020.117164.
[158] A. T. Drysdale *et al.*, "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nat. Med.*, vol. 23, no. 1, Art. no. 1, 2017, doi: 10.1038/nm.4246.
[159] T. Nakano *et al.*, "Enhancing Multi-Center Generalization of Machine Learning-Based Depression Diagnosis From Resting-State fMRI," *Front. Psychiatry*, vol. 11, 2020,

Accessed: Jul. 28, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyt.2020.00400
[160] D. Rivière, J.-F. Mangin, D. Papadopoulos-Orfanos, J.-M. Martinez, V. Frouin, and J. Régis, "Automatic recognition of cortical sulci of the human brain using a congregation of neural networks," *Med. Image Anal.*, vol. 6, no. 2, pp. 77–92, Jun. 2002, doi: 10.1016/s1361-8415(02)00052-x.
[161] A. Yamashita *et al.*, "Generalizable brain network markers of major depressive disorder across multiple imaging sites," *PLOS Biol.*, vol. 18, no. 12, p. e3000966, Dec. 2020, doi: 10.1371/journal.pbio.3000966.
[162] C. M. O'Brien, "Statistical Learning with Sparsity: The Lasso and Generalizations," *Int. Stat. Rev.*, vol. 84, no. 1, pp. 156–157, 2016.
[163] A. Yamashita *et al.*, "Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias," *PLOS Biol.*, vol. 17, no. 4, p. e3000042, Apr. 2019, doi: 10.1371/journal.pbio.3000042.
[164] K. Qin *et al.*, "Using graph convolutional network to characterize individuals with major depressive disorder across multiple imaging sites," *eBioMedicine*, vol. 78, p. 103977, Apr. 2022, doi: 10.1016/j.ebiom.2022.103977.
[165] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: a comprehensive review," *Comput. Soc. Netw.*, vol. 6, no. 1, p. 11, Nov. 2019, doi: 10.1186/s40649-019-0069-y.
[166] J. D. Power *et al.*, "Functional network organization of the human brain," *Neuron*, vol. 72, no. 4, Art. no. 4, Nov. 2011, doi: 10.1016/j.neuron.2011.09.006.
[167] R. S. Desikan *et al.*, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, Jul. 2006, doi: 10.1016/j.neuroimage.2006.01.021.
[168] C. DESTRIEUX, B. FISCHL, A. DALE, and E. HALGREN, "Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature," *NeuroImage*, vol. 53, no. 1, pp. 1–15, Oct. 2010, doi: 10.1016/j.neuroimage.2010.06.010.
[169] F. M. Krienen, B. T. T. Yeo, and R. L. Buckner, "Reconfigurable task-dependent functional coupling modes cluster around a core functional architecture," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 369, no. 1653, p. 20130526, Oct. 2014, doi: 10.1098/rstb.2013.0526.
[170] E. M. Gordon, T. O. Laumann, B. Adeyemo, and S. E. Petersen, "Individual Variability of the System-Level Organization of the Human Brain," *Cereb. Cortex*, vol. 27, no. 1, pp. 386–399, Jan. 2017, doi: 10.1093/cercor/bhv239.
[171] M. Salehi, A. Karbasi, D. S. Barron, D. Scheinost, and R. T. Constable, "Individualized functional networks reconfigure with cognitive state," *NeuroImage*, vol. 206, no. September 2019, p. 116233, 2020, doi: 10.1016/j.neuroimage.2019.116233.
[172] M. Salehi, A. S. Greene, A. Karbasi, X. Shen, D. Scheinost, and R. T. Constable, "There is no single functional atlas even for a single individual: Functional parcel definitions change with task," *NeuroImage*, vol. 208, p. 116366, Mar. 2020, doi: 10.1016/j.neuroimage.2019.116366.
[173] "[1608.06993] Densely Connected Convolutional Networks." https://arxiv.org/abs/1608.06993 (accessed Aug. 15, 2022).
[174] G. S. Wig *et al.*, "Parcellating an individual subject's cortical and subcortical brain structures using snowball sampling of resting-state correlations," *Cereb. Cortex N. Y. N 1991*, vol. 24, no. 8, Art. no. 8, Aug. 2014, doi: 10.1093/cercor/bht056.
[175] D. Kim *et al.*, "Machine Learning Classification of First-Onset Drug-Naive MDD Using Structural MRI," *IEEE Access*, vol. 7, pp. 153977–153985, 2019, doi: 10.1109/ACCESS.2019.2949128.

[176] D. Zhu *et al.*, “Classification of Major Depressive Disorder via Multi-site Weighted LASSO Model,” Sep. 2017, pp. 159–167. doi: 10.1007/978-3-319-66179-7_19.
[177] H. Guo, M. Qin, J. Chen, Y. Xu, and J. Xiang, “Machine-Learning Classifier for Patients with Major Depressive Disorder: Multifeature Approach Based on a High-Order Minimum Spanning Tree Functional Brain Network,” *Computational and Mathematical Methods in Medicine*, Dec. 14, 2017. https://www.hindawi.com/journals/cmmm/2017/4820935/ (accessed Jan. 18, 2021).
[178] M. D. Sacchet, G. Prasad, L. C. Foland-Ross, P. M. Thompson, and I. H. Gotlib, “Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory,” *Front. Psychiatry*, vol. 6, no. FEB, pp. 1–10, 2015, doi: 10.3389/fpsyt.2015.00021.
[179] R. Pomponio, “Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan,” p. 15, 2020.
[180] A. A. Chen *et al.*, “Mitigating site effects in covariance for machine learning in neuroimaging data,” *Hum. Brain Mapp.*, vol. 43, no. 4, pp. 1179–1195, Mar. 2022, doi: 10.1002/hbm.25688.
[181] R. C. Kessler and E. J. Bromet, “The epidemiology of depression across cultures,” *Annu. Rev. Public Health*, vol. 34, pp. 119–138, 2013, doi: 10.1146/annurev-publhealth-031912-114409.
[182] Y. Cho *et al.*, “Factors associated with quality of life in patients with depression: A nationwide population-based study,” *PLOS ONE*, vol. 14, no. 7, p. e0219455, Jul. 2019, doi: 10.1371/journal.pone.0219455.
[183] A. F. S. W. C. D. K. M. B. M. L. P. A. J. Cleare, “A Multidimensional Tool to Quantify Treatment Resistance in Depression: The Maudsley Staging Method,” *J. Clin. Psychiatry*, vol. 70, no. 2, p. 12363, Jan. 2009, doi: 10.4088/JCP.08m04309.
[184] L. K. M. Han *et al.*, “Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group,” *Mol. Psychiatry*, pp. 1–16, May 2020, doi: 10.1038/s41380-020-0754-0.
[185] C. Kraus, B. Kadriu, R. Lanzenberger, C. A. Zarate Jr., and S. Kasper, “Prognosis and improved outcomes in major depression: a review,” *Transl. Psychiatry*, vol. 9, no. 1, Art. no. 1, Apr. 2019, doi: 10.1038/s41398-019-0460-3.
[186] J. M. Gorman, “Comorbid depression and anxiety spectrum disorders,” *Depress. Anxiety*, vol. 4, no. 4, pp. 160–168, 1997 1996, doi: 10.1002/(SICI)1520-6394(1996)4:4<160::AID-DA2>3.0.CO;2-J.
[187] A. Steffen, J. Nübel, F. Jacobi, J. Bätzing, and J. Holstiege, “Mental and somatic comorbidity of depression: a comprehensive cross-sectional analysis of 202 diagnosis groups using German nationwide ambulatory claims data,” *BMC Psychiatry*, vol. 20, no. 1, p. 142, Mar. 2020, doi: 10.1186/s12888-020-02546-8.
[188] D. Arnone, A. M. McIntosh, K. P. Ebmeier, M. R. Munafò, and I. M. Anderson, “Magnetic resonance imaging studies in unipolar depression: systematic review and meta-regression analyses,” *Eur. Neuropsychopharmacol. J. Eur. Coll. Neuropsychopharmacol.*, vol. 22, no. 1, pp. 1–16, Jan. 2012, doi: 10.1016/j.euroneuro.2011.05.003.
[189] P. M. Thompson *et al.*, “The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data,” *Brain Imaging Behav.*, vol. 8, no. 2, pp. 153–182, Jun. 2014, doi: 10.1007/s11682-013-9269-5.
[190] J. Kambeitz *et al.*, “Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies,” *Biol. Psychiatry*, vol. 82, no. 5, pp. 330–338, Sep. 2017, doi: 10.1016/j.biopsych.2016.10.028.
[191] J. Algermissen and D. Mehler, “May the power be with you: Are there highly powered studies in neuroscience, and how can we get more of them?,” *J. Neurophysiol.*, vol. 119, Feb. 2018, doi: 10.1152/jn.00765.2017.

[192] Y. Zhang-James, M. Hoogman, B. Franke, and S. V. Faraone, “Machine Learning And MRI-Based Diagnostic Models For ADHD: Are We There Yet?,” medRxiv, Oct. 2020. doi: 10.1101/2020.10.20.20216390.
[193] E. Duerden, M. Chakravarty, J. Lerch, and M. Taylor, “Sex-Based Differences in Cortical and Subcortical Development in 436 Individuals Aged 4-54 Years,” *Cereb. Cortex N. Y. N 1991*, vol. 30, Dec. 2019, doi: 10.1093/cercor/bhz279.
[194] E. D. Gennatas *et al.*, “Age-Related Effects and Sex Differences in Gray Matter Density, Volume, Mass, and Cortical Thickness from Childhood to Young Adulthood,” *J. Neurosci.*, vol. 37, no. 20, pp. 5065–5073, May 2017, doi: 10.1523/JNEUROSCI.3550-16.2017.
[195] L. Schmaal *et al.*, “ENIGMA MDD: seven years of global neuroimaging studies of major depression through worldwide data sharing,” *Transl. Psychiatry*, vol. 10, no. 1, Art. no. 1, May 2020, doi: 10.1038/s41398-020-0842-6.
[196] P. E. Shrout and J. L. Rodgers, “Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis,” *Annu. Rev. Psychol.*, vol. 69, no. 1, pp. 487–510, 2018, doi: 10.1146/annurev-psych-122216-011845.
[197] H. Takao, N. Hayashi, and K. Ohtomo, “Effect of scanner in longitudinal studies of brain volume changes,” *J. Magn. Reson. Imaging*, vol. 34, no. 2, pp. 438–444, 2011, doi: 10.1002/jmri.22636.
[198] E. C. Brown, D. L. Clark, S. Hassel, G. MacQueen, and R. Ramasubbu, “Intrinsic thalamocortical connectivity varies in the age of onset subtypes in major depressive disorder,” *Neuropsychiatr. Dis. Treat.*, vol. 15, pp. 75–82, Dec. 2018, doi: 10.2147/NDT.S184425.
[199] K. Z. LeWinn, M. A. Sheridan, K. M. Keyes, A. Hamilton, and K. A. McLaughlin, “Sample composition alters associations between age and brain structure,” *Nat. Commun.*, vol. 8, no. 1, Art. no. 1, Oct. 2017, doi: 10.1038/s41467-017-00908-7.
[200] W. E. Johnson, C. Li, and A. Rabinovic, “Adjusting batch effects in microarray expression data using empirical Bayes methods,” *Biostatistics*, vol. 8, no. 1, pp. 118–127, Jan. 2007, doi: 10.1093/biostatistics/kxj037.
[201] J.-P. Fortin *et al.*, “Harmonization of cortical thickness measurements across scanners and sites,” *NeuroImage*, vol. 167, pp. 104–120, Feb. 2018, doi: 10.1016/j.neuroimage.2017.11.024.
[202] J.-P. Fortin *et al.*, “Harmonization of multi-site diffusion tensor imaging data,” *NeuroImage*, vol. 161, pp. 149–170, Nov. 2017, doi: 10.1016/j.neuroimage.2017.08.047.
[203] J. Radua *et al.*, “Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA,” *NeuroImage*, vol. 218, no. May, 2020, doi: 10.1016/j.neuroimage.2020.116956.
[204] A. Abraham *et al.*, “Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example,” *NeuroImage*, vol. 147, pp. 736–745, Feb. 2017, doi: 10.1016/j.neuroimage.2016.10.045.
[205] R. Pomponio *et al.*, “Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan,” *NeuroImage*, vol. 208, p. 116450, Mar. 2020, doi: 10.1016/j.neuroimage.2019.116450.
[206] A. A. Chen *et al.*, “Removal of Scanner Effects in Covariance Improves Multivariate Pattern Analysis in Neuroimaging Data,” *bioRxiv*, p. 858415, Dec. 2020, doi: 10.1101/858415.
[207] G. Mårtensson *et al.*, “The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study,” *Med. Image Anal.*, vol. 66, p. 101714, Dec. 2020, doi: 10.1016/j.media.2020.101714.
[208] M. Rozycki *et al.*, “Multisite Machine Learning Analysis Provides a Robust Structural Imaging Signature of Schizophrenia Detectable Across Diverse Patient Populations and

Within Individuals," *Schizophr. Bull.*, vol. 44, no. 5, pp. 1035–1044, Aug. 2018, doi: 10.1093/schbul/sbx137.
[209] T. Zindler, H. Frieling, A. Neyazi, S. Bleich, and E. Friedel, "Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies," *BMC Bioinformatics*, vol. 21, Jun. 2020, doi: 10.1186/s12859-020-03559-6.
[210] R. Dinga, L. Schmaal, B. W. J. H. Penninx, D. J. Veltman, and A. F. Marquand, "Controlling for effects of confounding variables on machine learning predictions," Bioinformatics, preprint, Aug. 2020. doi: 10.1101/2020.08.17.255034.
[211] H. T. N. Tran *et al.*, "A benchmark of batch-effect correction methods for single-cell RNA sequencing data," *Genome Biol.*, vol. 21, no. 1, p. 12, Jan. 2020, doi: 10.1186/s13059-019-1850-9.
[212] K. Dadi *et al.*, "Benchmarking functional connectome-based predictive models for resting-state fMRI," *NeuroImage*, vol. 192, pp. 115–134, May 2019, doi: 10.1016/j.neuroimage.2019.02.062.
[213] J.-H. Jung *et al.*, "Penalized logistic regression using functional connectivity as covariates with an application to mild cognitive impairment," *Commun. Stat. Appl. Methods*, vol. 27, no. 6, pp. 603–624, Nov. 2020, doi: 10.29220/CSAM.2020.27.6.603.
[214] A. Caprihan, G. D. Pearlson, and V. D. Calhoun, "Application of principal component analysis to distinguish patients with schizophrenia from healthy controls based on fractional anisotropy measurements," *NeuroImage*, vol. 42, no. 2, pp. 675–682, Aug. 2008, doi: 10.1016/j.neuroimage.2008.04.255.
[215] Q. Ma *et al.*, "Classification of multi-site MR images in the presence of heterogeneity using multi-task learning," *NeuroImage Clin.*, vol. 19, pp. 476–486, Jan. 2018, doi: 10.1016/j.nicl.2018.04.037.
[216] W. Hopkins, X. Li, T. Crow, and N. Roberts, "Vertex- and atlas-based comparisons in measures of cortical thickness, gyrification and white matter volume between humans and chimpanzees," *Brain Struct. Funct.*, vol. 222, Jan. 2017, doi: 10.1007/s00429-016-1213-1.
[217] I. Petrusic, D. Marko, K. Kacar, and J. Zidverc-Trajkovic, "Migraine with Aura: Surface-Based Analysis of the Cerebral Cortex with Magnetic Resonance Imaging," *Korean J. Radiol.*, vol. 19, p. 767, Jul. 2018, doi: 10.3348/kjr.2018.19.4.767.
[218] D. Xu, G. Xu, Z. Zhao, M. E. Sublette, J. M. Miller, and J. J. Mann, "Diffusion tensor imaging brain structural clustering patterns in major depressive disorder," *Hum. Brain Mapp.*, vol. 42, no. 15, pp. 5023–5036, Jul. 2021, doi: 10.1002/hbm.25597.
[219] M. Ramezani *et al.*, "Temporal-lobe morphology differs between healthy adolescents and those with early-onset of depression," *NeuroImage Clin.*, vol. 6, pp. 145–155, Jan. 2014, doi: 10.1016/j.nicl.2014.08.007.
[220] P.-C. Tu, L.-F. Chen, J.-C. Hsieh, Y.-M. Bai, C.-T. Li, and T.-P. Su, "Regional cortical thinning in patients with major depressive disorder: A surface-based morphometry study," *Psychiatry Res. Neuroimaging*, vol. 202, no. 3, pp. 206–213, Jun. 2012, doi: 10.1016/j.pscychresns.2011.07.011.
[221] M. Lener *et al.*, "Cortical Abnormalities and Association with Symptom Dimensions Across the Depressive Spectrum," *J. Affect. Disord.*, vol. 190, pp. 529–536, Nov. 2015, doi: 10.1016/j.jad.2015.10.027.
[222] G. Fung *et al.*, "Distinguishing bipolar and major depressive disorders by brain structural morphometry: A pilot study," *BMC Psychiatry*, vol. 15, Nov. 2015, doi: 10.1186/s12888-015-0685-5.
[223] Z. Iscan *et al.*, "Test–retest reliability of freesurfer measurements within and between sites: Effects of visual approval process," *Hum. Brain Mapp.*, vol. 36, no. 9, pp. 3472–3485, 2015, doi: 10.1002/hbm.22856.

[224] L. Qiu *et al.*, “Characterization of major depressive disorder using a multiparametric classification approach based on high resolution structural images,” *J. Psychiatry Neurosci.*, vol. 39, no. 2, pp. 78–86, Mar. 2014, doi: 10.1503/jpn.130034.
[225] J. Li *et al.*, “White-matter functional topology: a neuromarker for classification and prediction in unmedicated depression,” *Transl. Psychiatry*, vol. 10, no. 1, Art. no. 1, Oct. 2020, doi: 10.1038/s41398-020-01053-4.
[226] S. Liang *et al.*, “White Matter Abnormalities in Major Depression Biotypes Identified by Diffusion Tensor Imaging,” *Neurosci. Bull.*, vol. 35, no. 5, pp. 867–876, Oct. 2019, doi: 10.1007/s12264-019-00381-w.
[227] R. Goya-Maldonado, K. Brodmann, M. Keil, S. Trost, P. Dechent, and O. Gruber, “Differentiating unipolar and bipolar depression by alterations in large-scale brain networks,” *Hum. Brain Mapp.*, vol. 37, no. 2, pp. 808–818, 2016, doi: 10.1002/hbm.23070.
[228] H. C. Whalley *et al.*, “Prediction of Depression in Individuals at High Familial Risk of Mood Disorders Using Functional Magnetic Resonance Imaging,” *PLOS ONE*, vol. 8, no. 3, p. e57357, Mar. 2013, doi: 10.1371/journal.pone.0057357.
[229] “Missing Data: Five Practical Guidelines - Daniel A. Newman, 2014.” https://journals.sagepub.com/doi/full/10.1177/1094428114548590 (accessed Oct. 10, 2022).
[230] X. Han *et al.*, “Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer,” *NeuroImage*, vol. 32, no. 1, pp. 180–194, Aug. 2006, doi: 10.1016/j.neuroimage.2006.02.051.
[231] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl, “Within-subject template estimation for unbiased longitudinal image analysis,” *NeuroImage*, vol. 61, no. 4, pp. 1402–1418, Jul. 2012, doi: 10.1016/j.neuroimage.2012.02.084.
[232] J. Cramer, “The Origins of Logistic Regression,” Tinbergen Institute, Tinbergen Institute Discussion Paper 02-119/4, Dec. 2002. Accessed: Dec. 09, 2022. [Online]. Available: https://econpapers.repec.org/paper/tinwpaper/20020119.htm
[233] J. Wang, Q. Chen, and Y. Chen, “RBF Kernel Based Support Vector Machine with Universal Approximation and Its Application,” in *Advances in Neural Networks – ISNN 2004*, Berlin, Heidelberg, 2004, pp. 512–517. doi: 10.1007/978-3-540-28647-9_85.
[234] K. Fawagreh, M. M. Gaber, and E. Elyan, “Random forests: from early developments to recent advancements,” *Syst. Sci. Control Eng.*, vol. 2, no. 1, pp. 602–609, Dec. 2014, doi: 10.1080/21642583.2014.956265.
[235] A. V. Lebedev *et al.*, “Random Forest ensembles for detection and prediction of Alzheimer’s disease with a good between-cohort robustness,” *NeuroImage Clin.*, vol. 6, pp. 115–125, 2014, doi: 10.1016/j.nicl.2014.08.023.
[236] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, “Linear discriminant analysis: A detailed tutorial,” *AI Commun.*, vol. 30, no. 2, pp. 169–190, Jan. 2017, doi: 10.3233/AIC-170729.
[237] M.-A. Schulz *et al.*, “Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets,” *Nat. Commun.*, vol. 11, no. 1, Art. no. 1, Aug. 2020, doi: 10.1038/s41467-020-18037-z.
[238] L. Kohoutová *et al.*, “Toward a unified framework for interpreting machine-learning models in neuroimaging,” *Nat. Protoc.*, vol. 15, no. 4, Art. no. 4, 2020, doi: 10.1038/s41596-019-0289-5.
[239] R. R. Wilcox, *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York, NY, US: Springer-Verlag Publishing, 2001, pp. xiii, 258. doi: 10.1007/978-1-4757-3522-2.

[240] W. H. L. Pinaya *et al.*, "Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multi-cohort study," *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Aug. 2021, doi: 10.1038/s41598-021-95098-0.
[241] Y. Zhang, D. F. Jenkins, S. Manimaran, and W. E. Johnson, "Alternative empirical Bayes models for adjusting for batch effects in genomic studies," *BMC Bioinformatics*, vol. 19, no. 1, p. 262, Jul. 2018, doi: 10.1186/s12859-018-2263-6.
[242] R. Garcia-Dias *et al.*, "Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners," *NeuroImage*, vol. 220, p. 117127, Oct. 2020, doi: 10.1016/j.neuroimage.2020.117127.
[243] M. J. Friedrich, "Depression Is the Leading Cause of Disability Around the World," *JAMA*, vol. 317, no. 15, p. 1517, Apr. 2017, doi: 10.1001/jama.2017.3826.
[244] M. Machado, M. Iskedjian, I. Ruiz, and T. R. Einarson, "Remission, dropouts, and adverse drug reaction rates in major depressive disorder: a meta-analysis of head-to-head trials," *Curr. Med. Res. Opin.*, vol. 22, no. 9, pp. 1825–1837, Sep. 2006, doi: 10.1185/030079906X132415.
[245] A. J. Rush *et al.*, "Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report," *Am. J. Psychiatry*, vol. 163, no. 11, pp. 1905–1917, Nov. 2006, doi: 10.1176/ajp.2006.163.11.1905.
[246] V. Belov *et al.*, "Multi-site benchmark classification of major depressive disorder using machine learning on cortical and subcortical measures," Jun. 2022, Accessed: Aug. 25, 2022. [Online]. Available: https://arxiv-export1.library.cornell.edu/abs/2206.08122
[247] W. S. Kremen *et al.*, "Genetic and environmental influences on the size of specific brain regions in midlife: the VETSA MRI study," *NeuroImage*, vol. 49, no. 2, pp. 1213–1223, Jan. 2010, doi: 10.1016/j.neuroimage.2009.09.043.
[248] T. White, N. C. Andreasen, and P. Nopoulos, "Brain volumes and surface morphology in monozygotic twins," *Cereb. Cortex N. Y. N 1991*, vol. 12, no. 5, pp. 486–493, May 2002, doi: 10.1093/cercor/12.5.486.
[249] J. Li *et al.*, "Cortical structural differences in major depressive disorder correlate with cell type-specific transcriptional signatures," *Nat. Commun.*, vol. 12, no. 1, Art. no. 1, Mar. 2021, doi: 10.1038/s41467-021-21943-5.
[250] R. S. Cruz, L. Lebrat, P. Bourgeat, C. Fookes, J. Fripp, and O. Salvado, "DeepCSR: A 3D Deep Learning Approach for Cortical Surface Reconstruction," presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 806–815. Accessed: Sep. 12, 2022. [Online]. Available: https://openaccess.thecvf.com/content/WACV2021/html/Santa_Cruz_DeepCSR_A_3D_Deep_Learning_Approach_for_Cortical_Surface_Reconstruction_WACV_2021_paper.html
[251] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "FastSurfer - A fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, vol. 219, p. 117012, Oct. 2020, doi: 10.1016/j.neuroimage.2020.117012.
[252] L. Z. J. Williams, A. Fawaz, M. F. Glasser, A. D. Edwards, and E. C. Robinson, "Geometric Deep Learning of the Human Connectome Project Multimodal Cortical Parcellation," in *Machine Learning in Clinical Neuroimaging*, Cham, 2021, pp. 103–112. doi: 10.1007/978-3-030-87586-2_11.
[253] W. Yan *et al.*, "Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site FMRI data," *EBioMedicine*, vol. 47, pp. 543–552, Sep. 2019, doi: 10.1016/j.ebiom.2019.08.023.
[254] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[255] S.-B. Seong, C. Pae, and H.-J. Park, “Geometric Convolutional Neural Network for Analyzing Surface-Based Neuroimaging Data,” *Front. Neuroinformatics*, vol. 12, 2018, doi: 10.3389/fninf.2018.00042.
[256] Z. Su *et al.*, “Optimal Mass Transport for Shape Matching and Comparison,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2246–2259, Nov. 2015, doi: 10.1109/TPAMI.2015.2408346.
[257] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.
[258] X. Lu *et al.*, “Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images,” *Medicine (Baltimore)*, vol. 95, no. 30, p. e3973, Jul. 2016, doi: 10.1097/MD.0000000000003973.
[259] V. Wottschel *et al.*, “SVM recursive feature elimination analyses of structural brain MRI predicts near-term relapses in patients with clinically isolated syndromes suggestive of multiple sclerosis,” *NeuroImage Clin.*, vol. 24, p. 102011, 2019, doi: 10.1016/j.nicl.2019.102011.
[260] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[261] R. Caruana, “Multitask Learning,” *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997, doi: 10.1023/A:1007379606734.
[262] N. Dinsdale, M. Jenkinson, and A. Namburete, “Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal,” *NeuroImage*, vol. 228, p. 117689, Dec. 2020, doi: 10.1016/j.neuroimage.2020.117689.
[263] N. Demirci and M. A. Holland, “Cortical thickness systematically varies with curvature and depth in healthy human brains,” *Hum. Brain Mapp.*, vol. 43, no. 6, pp. 2064–2084, 2022, doi: 10.1002/hbm.25776.
[264] T. C. Ho *et al.*, “Sex differences in myelin content of white matter tracts in adolescents with depression,” *Neuropsychopharmacology*, vol. 46, no. 13, Art. no. 13, Dec. 2021, doi: 10.1038/s41386-021-01078-3.
[265] M. D. Sacchet and I. H. Gotlib, “Myelination of the brain in Major Depressive Disorder: An in vivo quantitative magnetic resonance imaging study,” *Sci. Rep.*, vol. 7, no. 1, Art. no. 1, May 2017, doi: 10.1038/s41598-017-02062-y.
[266] L. van Velzen *et al.*, “White matter disturbances in major depressive disorder: a coordinated analysis across 20 international cohorts in the ENIGMA MDD working group,” *Mol. Psychiatry*, vol. 25, Jul. 2020, doi: 10.1038/s41380-019-0477-2.
[267] M. D. Hettwer *et al.*, “Coordinated cortical thickness alterations across six neurodevelopmental and psychiatric disorders,” *Nat. Commun.*, vol. 13, no. 1, Art. no. 1, Nov. 2022, doi: 10.1038/s41467-022-34367-6.
[268] S. Ayyash *et al.*, “Exploring brain connectivity changes in major depressive disorder using functional-structural data fusion: A CAN-BIND-1 study,” *Hum. Brain Mapp.*, vol. 42, no. 15, pp. 4940–4957, 2021, doi: 10.1002/hbm.25590.
[269] Y.-D. Zhang *et al.*, “Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation,” *Inf. Fusion*, vol. 64, pp. 149–187, Dec. 2020, doi: 10.1016/j.inffus.2020.07.006.
[270] A. Fornito, A. Zalesky, and M. Breakspear, “The connectomics of brain disorders,” *Nat. Rev. Neurosci.*, vol. 16, no. 3, Art. no. 3, Mar. 2015, doi: 10.1038/nrn3901.
[271] G. Wig *et al.*, “Parcellating an Individual Subject’s Cortical and Subcortical Brain Structures Using Snowball Sampling of Resting-State Correlations,” *Cereb. Cortex N. Y. N 1991*, vol. 24, Mar. 2013, doi: 10.1093/cercor/bht056.

[272] J. Zhou *et al.*, “Graph Neural Networks: A Review of Methods and Applications,” *ArXiv181208434 Cs Stat*, Jul. 2019, Accessed: Jan. 07, 2021. [Online]. Available: http://arxiv.org/abs/1812.08434
[273] J. M. M. Bayer *et al.*, “Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses,” *Front. Neurol.*, vol. 13, 2022, Accessed: Dec. 21, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fneur.2022.923988
[274] A. Solanes *et al.*, “Removing the effects of the site in brain imaging machine-learning – Measurement and extendable benchmark,” *NeuroImage*, p. 119800, Dec. 2022, doi: 10.1016/j.neuroimage.2022.119800.
[275] C. Hsu, C. Chang, and C.-J. Lin, “A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin,” Nov. 2003.
[276] M. S. George *et al.*, “Prefrontal repetitive transcranial magnetic stimulation (rTMS) changes relative perfusion locally and remotely,” *Hum. Psychopharmacol. Clin. Exp.*, vol. 14, no. 3, pp. 161–170, 1999, doi: https://doi.org/10.1002/(SICI)1099-1077(199904)14:3<161::AID-HUP73>3.0.CO;2-2.
[277] L. Cocchi, M. V. Sale, A. Lord, A. Zalesky, M. Breakspear, and J. B. Mattingley, “Dissociable effects of local inhibitory and excitatory theta-burst stimulation on large-scale brain dynamics,” *J. Neurophysiol.*, vol. 113, no. 9, Art. no. 9, May 2015, doi: 10.1152/jn.00850.2014.
[278] M. C. Eldaief, M. A. Halko, R. L. Buckner, and A. Pascual-Leone, “Transcranial magnetic stimulation modulates the brain’s intrinsic activity in a frequency-dependent manner,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 52, Art. no. 52, Dec. 2011, doi: 10.1073/pnas.1113103109.
[279] M. V. Sale, J. B. Mattingley, A. Zalesky, and L. Cocchi, “Imaging human brain networks to improve the clinical efficacy of non-invasive brain stimulation,” *Neurosci. Biobehav. Rev.*, vol. 57, pp. 187–198, Oct. 2015, doi: 10.1016/j.neubiorev.2015.09.010.
[280] A. Singh *et al.*, “Default mode network alterations after intermittent theta burst stimulation in healthy subjects,” *Transl. Psychiatry*, vol. 10, no. 1, Art. no. 1, Feb. 2020, doi: 10.1038/s41398-020-0754-5.
[281] N. Valchev *et al.*, “cTBS delivered to the left somatosensory cortex changes its functional connectivity during rest,” *NeuroImage*, vol. 114, pp. 386–397, Jul. 2015, doi: 10.1016/j.neuroimage.2015.04.017.
[282] T. Watanabe *et al.*, “Bidirectional effects on interhemispheric resting-state functional connectivity induced by excitatory and inhibitory repetitive transcranial magnetic stimulation,” *Hum. Brain Mapp.*, vol. 35, no. 5, Art. no. 5, May 2014, doi: 10.1002/hbm.22300.
[283] A. L. Cohen *et al.*, “Defining functional areas in individual human brains using resting functional connectivity MRI,” *NeuroImage*, vol. 41, no. 1, Art. no. 1, May 2008, doi: 10.1016/j.neuroimage.2008.01.066.
[284] G. S. Wig *et al.*, “Parcellating an individual subject’s cortical and subcortical brain structures using snowball sampling of resting-state correlations,” *Cereb. Cortex N. Y. N 1991*, vol. 24, no. 8, pp. 2036–2054, Aug. 2014, doi: 10.1093/cercor/bht056.
[285] J. D. Power *et al.*, “Functional network organization of the human brain,” *Neuron*, vol. 72, no. 4, pp. 665–678, Nov. 2011, doi: 10.1016/j.neuron.2011.09.006.
[286] S. M. Smith *et al.*, “Functional connectomics from resting-state fMRI,” *Trends Cogn. Sci.*, vol. 17, no. 12, pp. 666–682, Dec. 2013, doi: 10.1016/j.tics.2013.09.016.
[287] B. T. T. Yeo *et al.*, “The organization of the human cerebral cortex estimated by intrinsic functional connectivity,” *J. Neurophysiol.*, vol. 106, no. 3, pp. 1125–1165, Sep. 2011, doi: 10.1152/jn.00338.2011.

[288] R. Goya-Maldonado, K. Brodmann, M. Keil, S. Trost, P. Dechent, and O. Gruber, "Differentiating unipolar and bipolar depression by alterations in large-scale brain networks," *Hum. Brain Mapp.*, vol. 37, no. 2, pp. 808–818, 2016, doi: https://doi.org/10.1002/hbm.23070.
[289] Y. Wei *et al.*, "Local functional connectivity alterations in schizophrenia, bipolar disorder, and major depressive disorder," *J. Affect. Disord.*, vol. 236, pp. 266–273, Aug. 2018, doi: 10.1016/j.jad.2018.04.069.
[290] M. L. Stanley, M. N. Moussa, B. Paolini, R. G. Lyday, J. H. Burdette, and P. J. Laurienti, "Defining nodes in complex brain networks," *Front. Comput. Neurosci.*, vol. 7, 2013, doi: 10.3389/fncom.2013.00169.
[291] E. M. Gordon, T. O. Laumann, B. Adeyemo, J. F. Huckins, W. M. Kelley, and S. E. Petersen, "Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations," *Cereb. Cortex N. Y. N 1991*, vol. 26, no. 1, pp. 288–303, Jan. 2016, doi: 10.1093/cercor/bhu239.
[292] J. Tourville, A. Nieto-Castañón, M. Heyne, and F. Guenther, "Functional Parcellation of the Speech Production Cortex.," *J. Speech Lang. Hear. Res. JSLHR*, 2019, doi: 10.1044/2019_JSLHR-S-CSMC7-18-0442.
[293] M. P. van den Heuvel and O. Sporns, "A cross-disorder connectome landscape of brain dysconnectivity," *Nat. Rev. Neurosci.*, vol. 20, no. 7, Art. no. 7, Jul. 2019, doi: 10.1038/s41583-019-0177-6.
[294] V. D. Calhoun, J. Liu, and T. Adalı, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *NeuroImage*, vol. 45, no. 1, Supplement 1, pp. S163–S172, Mar. 2009, doi: 10.1016/j.neuroimage.2008.10.057.
[295] F. Agosta, M. Pievani, C. Geroldi, M. Copetti, G. B. Frisoni, and M. Filippi, "Resting state fMRI in Alzheimer's disease: beyond the default mode network," *Neurobiol. Aging*, vol. 33, no. 8, pp. 1564–1578, Aug. 2012, doi: 10.1016/j.neurobiolaging.2011.06.007.
[296] L. R. Peraza *et al.*, "fMRI resting state networks and their association with cognitive fluctuations in dementia with Lewy bodies," *NeuroImage Clin.*, vol. 4, pp. 558–565, 2014, doi: 10.1016/j.nicl.2014.03.013.
[297] S. B. Eickhoff, B. T. T. Yeo, and S. Genon, "Imaging-based parcellations of the human brain," *Nat. Rev. Neurosci.*, vol. 19, no. 11, Art. no. 11, 2018, doi: 10.1038/s41583-018-0071-7.
[298] M. F. Glasser *et al.*, "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, 11 2016, doi: 10.1038/nature18933.
[299] R. Kong *et al.*, "Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion," *Cereb. Cortex N. Y. N 1991*, vol. 29, no. 6, pp. 2533–2551, 01 2019, doi: 10.1093/cercor/bhy123.
[300] R. Saxe, J. M. Moran, J. Scholz, and J. Gabrieli, "Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects," *Soc. Cogn. Affect. Neurosci.*, vol. 1, no. 3, pp. 229–234, Dec. 2006, doi: 10.1093/scan/nsl034.
[301] E. M. Gordon *et al.*, "Precision Functional Mapping of Individual Human Brains," *Neuron*, vol. 95, no. 4, pp. 791-807.e7, Aug. 2017, doi: 10.1016/j.neuron.2017.07.011.
[302] T. O. Laumann *et al.*, "Functional System and Areal Organization of a Highly Sampled Individual Human Brain," *Neuron*, vol. 87, no. 3, pp. 657–670, Aug. 2015, doi: 10.1016/j.neuron.2015.06.037.
[303] I. Tavor, O. Parker Jones, R. B. Mars, S. M. Smith, T. E. Behrens, and S. Jbabdi, "Task-free MRI predicts individual differences in brain activity during task performance," *Science*, vol. 352, no. 6282, pp. 216–220, Apr. 2016, doi: 10.1126/science.aad8127.
[304] D. B. Archer *et al.*, "Development and Validation of the Automated Imaging Differentiation in Parkinsonism (AID-P): A Multi-Site Machine Learning Study," *Lancet*

Digit. Health, vol. 1, no. 5, pp. e222–e231, Sep. 2019, doi: 10.1016/s2589-7500(19)30105-0.
[305] B. Cao *et al.*, "Treatment response prediction and individualized identification of first-episode drug-naïve schizophrenia using brain functional connectivity," *Mol. Psychiatry*, vol. 25, no. 4, Art. no. 4, Apr. 2020, doi: 10.1038/s41380-018-0106-5.
[306] T. Evgeniou and M. Pontil, "Support Vector Machines: Theory and Applications," Jan. 2001, vol. 2049, pp. 249–257. doi: 10.1007/3-540-44673-7_12.
[307] S. Chaplot, L. M. Patnaik, and N. R. Jagannathan, "Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network," *Biomed. Signal Process. Control*, vol. 1, no. 1, pp. 86–92, Jan. 2006, doi: 10.1016/j.bspc.2006.05.002.
[308] J. Miao and L. Niu, "A Survey on Feature Selection," *Procedia Comput. Sci.*, vol. 91, pp. 919–926, Dec. 2016, doi: 10.1016/j.procs.2016.07.111.
[309] B. Jin *et al.*, "Feature selection for fMRI-based deception detection," *BMC Bioinformatics*, vol. 10, no. 9, p. S15, Sep. 2009, doi: 10.1186/1471-2105-10-S9-S15.
[310] J. Sui *et al.*, "Combination of FMRI-SMRI-EEG data improves discrimination of schizophrenia patients by ensemble feature selection," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2014, pp. 3889–3892. doi: 10.1109/EMBC.2014.6944473.
[311] L. L. Gollo, J. A. Roberts, and L. Cocchi, "Mapping how local perturbations influence systems-level brain dynamics," *NeuroImage*, vol. 160, pp. 97–112, 15 2017, doi: 10.1016/j.neuroimage.2017.01.057.
[312] V. Kozyrev, R. Staadt, U. T. Eysel, and D. Jancke, "TMS-induced neuronal plasticity enables targeted remodeling of visual cortical maps," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 25, pp. 6476–6481, 19 2018, doi: 10.1073/pnas.1802798115.
[313] V. Kozyrev, U. T. Eysel, and D. Jancke, "Voltage-sensitive dye imaging of transcranial magnetic stimulation-induced intracortical dynamics," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 37, pp. 13553–13558, Sep. 2014, doi: 10.1073/pnas.1405508111.
[314] A. Singh, T. Erwin-Grabner, G. Sutcliffe, A. Antal, W. Paulus, and R. Goya-Maldonado, "Personalized repetitive transcranial magnetic stimulation temporarily alters default mode network in healthy subjects," *Sci. Rep.*, vol. 9, no. 1, p. 5631, Apr. 2019, doi: 10.1038/s41598-019-42067-3.
[315] Y. Chao-Gan and Z. Yu-Feng, "DPARSF: A MATLAB Toolbox for 'Pipeline' Data Analysis of Resting-State fMRI," *Front. Syst. Neurosci.*, vol. 4, May 2010, doi: 10.3389/fnsys.2010.00013.
[316] T. D. Satterthwaite *et al.*, "An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data," *NeuroImage*, vol. 64, pp. 240–256, Jan. 2013, doi: 10.1016/j.neuroimage.2012.08.052.
[317] C.-G. Yan *et al.*, "A Comprehensive Assessment of Regional Variation in the Impact of Head Micromovements on Functional Connectomics," *NeuroImage*, vol. 76, pp. 183–201, Aug. 2013, doi: 10.1016/j.neuroimage.2013.03.004.
[318] H. Togo *et al.*, "Effects of Field-Map Distortion Correction on Resting State Functional Connectivity MRI," *Front. Neurosci.*, vol. 11, Dec. 2017, doi: 10.3389/fnins.2017.00656.
[319] J. D. Power, M. Plitt, T. O. Laumann, and A. Martin, "Sources and implications of whole-brain fMRI signals in humans," *NeuroImage*, vol. 146, pp. 609–625, Feb. 2017, doi: 10.1016/j.neuroimage.2016.09.038.
[320] M. Khosla, K. Jamison, G. H. Ngo, A. Kuceyeski, and M. R. Sabuncu, "Machine learning in resting-state fMRI analysis," *Magn. Reson. Imaging*, vol. 64, pp. 101–121, 2019, doi: 10.1016/j.mri.2019.05.031.

[321] O. Kolade, A. A. Olayinka, and U. Ovie, “Fingerprint Database Optimization Using Watershed Transformation Algorithm,” *Open J. Optim.*, vol. 3, no. 4, Art. no. 4, Nov. 2014, doi: 10.4236/ojop.2014.34006.
[322] V. Wottschel *et al.*, “SVM recursive feature elimination analyses of structural brain MRI predicts near-term relapses in patients with clinically isolated syndromes suggestive of multiple sclerosis,” *NeuroImage Clin.*, vol. 24, p. 102011, 2019, doi: 10.1016/j.nicl.2019.102011.
[323] K. R. A. Van Dijk, M. R. Sabuncu, and R. L. Buckner, “The influence of head motion on intrinsic functional connectivity MRI,” *NeuroImage*, vol. 59, no. 1, pp. 431–438, Jan. 2012, doi: 10.1016/j.neuroimage.2011.07.044.
[324] N. A. Crossley *et al.*, “Cognitive relevance of the community structure of the human brain functional coactivation network,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 28, pp. 11583–11588, Jul. 2013, doi: 10.1073/pnas.1220826110.
[325] M. P. van den Heuvel and O. Sporns, “Network hubs in the human brain,” *Trends Cogn. Sci.*, vol. 17, no. 12, pp. 683–696, Dec. 2013, doi: 10.1016/j.tics.2013.09.012.
[326] S. Hirose *et al.*, “Local Signal Time-Series during Rest Used for Areal Boundary Mapping in Individual Human Brains,” *PLOS ONE*, vol. 7, no. 5, p. e36496, May 2012, doi: 10.1371/journal.pone.0036496.
[327] J. M. Huntenburg, P.-L. Bazin, and D. S. Margulies, “Large-Scale Gradients in Human Cortical Organization,” *Trends Cogn. Sci.*, vol. 22, no. 1, pp. 21–31, Jan. 2018, doi: 10.1016/j.tics.2017.11.002.
[328] S. Jbabdi, S. N. Sotiropoulos, and T. E. Behrens, “The topographic connectome,” *Curr. Opin. Neurobiol.*, vol. 23, no. 2, pp. 207–215, Apr. 2013, doi: 10.1016/j.conb.2012.12.004.
[329] B. A. Vogt, L. Vogt, and S. Laureys, “Cytology and functionally correlated circuits of human posterior cingulate areas,” *NeuroImage*, vol. 29, no. 2, pp. 452–466, Jan. 2006, doi: 10.1016/j.neuroimage.2005.07.048.
[330] F. Radicchi and A. Arenas, “Abrupt transition in the structural formation of interconnected networks,” *Nat. Phys.*, vol. 9, no. 11, Art. no. 11, Nov. 2013, doi: 10.1038/nphys2761.
[331] S. L. Florence, N. Jain, and J. H. Kaas, “Plasticity of Somatosensory Cortex in Primates,” *Semin. Neurosci.*, vol. 9, no. 1, pp. 3–12, Jan. 1997, doi: 10.1006/smns.1997.0101.
[332] A. Holtmaat and K. Svoboda, “Experience-dependent structural synaptic plasticity in the mammalian brain,” *Nat. Rev. Neurosci.*, vol. 10, no. 9, pp. 647–658, Sep. 2009, doi: 10.1038/nrn2699.
[333] M. Lenz *et al.*, “Repetitive magnetic stimulation induces plasticity of inhibitory synapses,” *Nat. Commun.*, vol. 7, p. 10020, Jan. 2016, doi: 10.1038/ncomms10020.
[334] N. Eshel *et al.*, “Global connectivity and local excitability changes underlie antidepressant effects of repetitive transcranial magnetic stimulation,” *Neuropsychopharmacology*, vol. 45, no. 6, Art. no. 6, May 2020, doi: 10.1038/s41386-020-0633-z.
[335] G. Castrillon, N. Sollmann, K. Kurcyus (Bieńkowska), A. Razi, S. Krieg, and V. Riedl, “The physiological effects of noninvasive brain stimulation fundamentally differ across the human cortex,” *Sci. Adv.*, vol. 6, p. eaay2739, Jan. 2020, doi: 10.1126/sciadv.aay2739.
[336] S. W. Davis, B. Luber, D. L. K. Murphy, S. H. Lisanby, and R. Cabeza, “Frequency-specific neuromodulation of local and distant connectivity in aging and episodic memory function,” *Hum. Brain Mapp.*, vol. 38, no. 12, pp. 5987–6004, 2017, doi: https://doi.org/10.1002/hbm.23803.

[337] A. L. W. Bokde *et al.*, "Functional connectivity of the fusiform gyrus during a face-matching task in subjects with mild cognitive impairment," *Brain*, vol. 129, no. 5, pp. 1113–1124, May 2006, doi: 10.1093/brain/awl051.
[338] C. Chang and G. H. Glover, "Effects of model-based physiological noise correction on default mode network anti-correlations and correlations," *NeuroImage*, vol. 47, no. 4, pp. 1448–1459, Oct. 2009, doi: 10.1016/j.neuroimage.2009.05.012.
[339] J. Jiang, J. Beck, K. Heller, and T. Egner, "An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands," *Nat. Commun.*, vol. 6, no. 1, Art. no. 1, Sep. 2015, doi: 10.1038/ncomms9165.
[340] A. A. Taren *et al.*, "Mindfulness Meditation Training and Executive Control Network Resting State Functional Connectivity: A Randomized Controlled Trial," *Psychosom. Med.*, vol. 79, no. 6, pp. 674–683, 2017, doi: 10.1097/PSY.0000000000000466.
[341] R. F. Becker, "Essay on the cerebral cortex. By Gerhardt von Bonin. Charles C Thomas, Springfield, Ill. 1950. 150 pp," *Am. J. Phys. Anthropol.*, vol. 11, no. 3, pp. 441–444, 1953, doi: https://doi.org/10.1002/ajpa.1330110317.
[342] M. F. Glasser and D. C. V. Essen, "Mapping Human Cortical Areas In Vivo Based on Myelin Content as Revealed by T1- and T2-Weighted MRI," *J. Neurosci.*, vol. 31, no. 32, pp. 11597–11616, Aug. 2011, doi: 10.1523/JNEUROSCI.2180-11.2011.
[343] V. Wottschel *et al.*, "Predicting outcome in clinically isolated syndrome using machine learning," *NeuroImage Clin.*, vol. 7, pp. 281–287, 2015, doi: 10.1016/j.nicl.2014.11.021.
[344] A. Choudhary, L. Tong, Y. Zhu, and M. D. Wang, "Advancing Medical Imaging Informatics by Deep Learning-Based Domain Adaptation," *Yearb. Med. Inform.*, vol. 29, no. 1, pp. 129–138, Aug. 2020, doi: 10.1055/s-0040-1702009.
[345] E. S. Finn *et al.*, "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nat. Neurosci.*, vol. 18, no. 11, Art. no. 11, Nov. 2015, doi: 10.1038/nn.4135.
[346] X. Zhu *et al.*, "Evidence of a dissociation pattern in resting-state default mode network connectivity in first-episode, treatment-naive major depression patients," *Biol. Psychiatry*, vol. 71, no. 7, Art. no. 7, Apr. 2012, doi: 10.1016/j.biopsych.2011.10.035.
[347] L. Luo *et al.*, "Abnormal large-scale resting-state functional networks in drug-free major depressive disorder," *Brain Imaging Behav.*, vol. 15, no. 1, pp. 96–106, Feb. 2021, doi: 10.1007/s11682-019-00236-y.
[348] J. Zhang *et al.*, "Dynamic changes of large-scale resting-state functional networks in major depressive disorder," *Prog. Neuropsychopharmacol. Biol. Psychiatry*, vol. 111, p. 110369, Dec. 2021, doi: 10.1016/j.pnpbp.2021.110369.
[349] Z. Fang, W. Li, J. Zou, and Q. Du, "Using CNN-based high-level features for remote sensing scene classification," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul. 2016, pp. 2610–2613. doi: 10.1109/IGARSS.2016.7729674.
[350] S. Targ, D. Almeida, and K. Lyman, "Resnet in Resnet: Generalizing Residual Architectures," Feb. 2016, Accessed: Aug. 31, 2022. [Online]. Available: https://openreview.net/forum?id=lx9l4r36gU2OVPy8Cv9g
[351] R. Dinga *et al.*, "Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017)," *NeuroImage Clin.*, vol. 22, no. March, p. 101796, 2019, doi: 10.1016/j.nicl.2019.101796.
[352] T. Tokuda *et al.*, "Identification of depression subtypes and relevant brain regions using a data-driven approach," *Sci. Rep.*, vol. 8, no. 1, Art. no. 1, Sep. 2018, doi: 10.1038/s41598-018-32521-z.
[353] S. Liang *et al.*, "Biotypes of major depressive disorder: Neuroimaging evidence from resting-state default mode network patterns," *NeuroImage Clin.*, vol. 28, p. 102514, Jan. 2020, doi: 10.1016/j.nicl.2020.102514.

[354] N. U. F. Dosenbach *et al.*, “Prediction of Individual Brain Maturity Using fMRI,” *Science*, vol. 329, no. 5997, pp. 1358–1361, Sep. 2010, doi: 10.1126/science.1194144.