Aus dem Institut für Anatomie I
der Heinrich-Heine-Universität Düsseldorf
Direktorin : Univ.-Prof. Dr. med. Dr. rer. pol. Svenja Caspers

Diagnosis of Depressive Disorder based on Structural Imaging using Automated Machine Learning (AML)

Dissertation

zur Erlangung des Grades eines Doktors der Medizin
der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Mathieu Ehlinger Delcourt
2024

Als Inauguraldissertation gedruckt mit Genehmigung
der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

gez.:
Dekan: Prof. Dr. med. Nikolaj Klöcker
Erstgutachterin: Univ.-Prof. Dr. med. Dr. rer. pol. Svenja Caspers
Zweitgutachterin: Univ.-Prof. Dr. med. Eva Meisenzahl

"Il faut imaginer Sisyphe heureux"

Albert Camus, *Le Mythe de Sisyphe*

Zusammenfassung

Akkurate *machine learning* (ML) Modelle mit hoher Generalisierbarkeit für die Diagnose einer Depression könnten sich in mehrfacher Hinsicht als wertvoll erweisen. Sie könnten beim Screening auf depressive Störungen zum Einsatz kommen, bei der Differentialdiagnose von klinisch ähnlichen Entitäten helfen und Erkenntnisse über die Pathomechanismen der Krankheit liefern. Ob eine solche Klassifikation mit klassischen ML Methoden möglich ist, ist eine offene und derzeit umstrittene Frage. Hierzu gibt es derzeit widersprüchliche Berichte und Veröffentlichungen. Parallel dazu entwickelte sich in den letzten Jahren ein neues Teilgebiet von ML, nämlich *automated machine learning* (AML). Ziel dieser Arbeit war es in einem ersten Schritt, die Anwendbarkeit von AML auf die strukturelle Neurobildgebung anhand von zwei gängigen ML Aufgaben zu überprüfen (nämlich Alter- sowie Geschlechtsvorhersage). Im zweiten Teil nutzten wir es zur Klassifizierung einer Depression als Alternative zu Standard ML Ansätzen.

Zwei separate Stichproben wurden aus der 1000BRAINS Studie ($N_{1000BRAINS}$=1157) für eine AML Benchmark herangezogen. Die strukturelle Magnetresonanztomographie (MRT) der Probanden wurde für die regionsweise Extraktion von kortikalen Eigenschaften vorverarbeitet. Die erste Stichprobe wurde für die Herstellung der AML Modelle verwendet, die zweite für die Validierung der Leistungen auf neue Subjekte. Wir beendeten die AML Benchmark mit einem letzten Test an einer externen Stichprobe aus der BiDirect Studie ($N_{BiDirect}$=1102). Zur Klassifikation der Depression wurden zwei separate Stichproben aus BiDirect abgeleitet, um unsere AML Modelle zunächst zu entwickeln und dann zu evaluieren. Die AML Modelle erreichten bei der Altersvorhersage 5.69 *mean absolute error* (MAE) und bei der Geschlechtsvorhersage 85.8% *balanced accuracy* (BA) in der externen Validierung, ähnlich wie *state-of-the-art* ML Modelle. Hinsichtlich der Klassifikation einer Depression schnitten AML Modelle nur geringfügig besser ab als eine Zufallsvorhersage (55.1% BA).

Mit dieser Arbeit konnten wir zeigen, dass AML eine tragfähige und effiziente Alternative zu klassischen ML Methoden in *neuroimaging* basierten Aufgaben ist, mit guter Generalisierbarkeit. Wie klassische ML Modelle konnten AML Modelle depressive Patienten von *healthy controls* (HC) nicht unterscheiden. Dieses negative Ergebnis konvergiert mit aktuellen negativen Ergebnissen für diese ML Aufgabe.

Abstract

Accurate and well-generalizing machine learning (ML) models for depression diagnosis could prove to be tremendously valuable in multiple ways. They could find application in screening for depressive disorder, assist in differential diagnosis from clinically similar entities and provide insights about the illness's pathomechanisms. Whether such a classification is possible with classical ML methods is an open and currently debated question. Conflicting reports and publications exist on this matter. In parallel, the recent years saw the development of automated machine learning (AML), a relatively new subfield of ML.

This thesis aimed first at assessing the applicability of AML to structural neuroimaging in two commonly performed ML tasks (namely age and sex prediction). In the second part, we applied it to depression classification as an alternative to standard ML approaches.

Two separate samples were drawn from the 1000BRAINS study ($N_{1000BRAINS}$=1157) for an AML benchmark . Subjects' structural MRIs were preprocessed for region-wise extraction of cortical features, namely cortical thickness (CT), grey matter volume (GMV) and surface area (SA). The first sample was used for AML model design, the second for validation of performance on subjects not involved in pipeline design to test the generalizability of our models. We ended the AML benchmark with an external validation on a sample drawn from the BiDirect study ($N_{BiDirect}$=1102). Regarding depression classification: two separate samples were derived from BiDirect for designing then validating our AML models. AML performed similar to state of the art ML models for age (5.69 mean absolute error (MAE)) and sex prediction (85.8% balanced accuracy (BA)). For depression classification, AML models only performed slightly better than random (55.1% BA).

With this work, we showed that AML is a viable and efficient alternative to standard ML methods in neuroimaging based tasks, with good generalization power. Similarly to classical ML methods, AML models could not differentiate depressive patients from healthy controls (HC). This negative result converges with recent negative findings for this task.

Acronyms

ACC: anterior cingulate cortex

AML: automated machine learning

BA: balanced accuracy

CT: cortical thickness

DALY: disability adjusted life years

DEAP: Distributed Evolutionary Algorithms in Python

DL: deep learning

DLPFC: dorsolateral prefrontal cortex

DSM-5: *Diagnostic and statistical manual of mental disorders* (5th ed.; American Psychiatric Association, 2013)

DT: decision tree

DTI: diffusion tensor imaging

DW-MRI: diffusion-weighted magnetic resonance imaging

ECT: electroconvulsive therapy

eTIV: estimated total intracranial volume

FA: flip angle

fMRI: functional MRI

FoV: field of view

GMV: grey matter volume

GP: Gaussian process

gwMRF: gradient-weighted Markov Random Field

HAM-A-14: Hamilton Anxiety Rating Scale
HAM-D-17: Hamilton Depression Rating Scale
HC: healthy controls
HF-rTMS: high-frequency repetitive transcranial magnetic stimulation
HNR: Heinz Nixdorf Recall

ICD-10: *International statistical classification of diseases and related health problems* (10th ed.; World Health Organization, 1993)

LOOCV: leave-one-out-cross-validation

MAE: mean absolute error
MDD: major depressive disorder
MINI: Mini-International Neuropsychiatric Interview
ML: machine learning
MLP: multi-layer perceptron
MP-RAGE: Magnetization Prepared - RApid Gradient Echo
MRI: magnetic resonance imaging
MSE: mean squared error

OFC: orbitofrontal cortex

PAC: Predictive Analytics Competition
PFC: prefrontal cortex

QC: quality control

RBF: radial basis function
rs-fMRI: resting state fMRI

SA: surface area

SD: standard deviation

SGD: stochastic gradient descent

sMRI: structural MRI

SVM: support vector machines

SVR: support vector regression

TE: echo time

TI: inversion time

TPOT: Tree-Based Pipeline Optimization Tool

TR: repetition time

VR: voxel resolution

Contents

CHAPTER 1

Introduction

Depression undoubtedly counts to the major illnesses of our time, with a lifetime prevalence of 15.7% in Germany for diagnosed depressive disorders (Boudes et al., 2014). The 12 months prevalence of major depressive disorder (MDD) is at 4.2% for men and 9.9% for women in Germany, numbers that have been on the rise (Maske et al., 2016; Steffen et al., 2020). Depression is thought to be one of the most common cause for disability adjusted life years (DALY) worldwide (Vos et al., 2020; Rehm & Shield, 2019). Not only does it have an important impact on the quality of life of the patients themselves, but on their relatives and close ones as well.

There is as such a strong incentive to better understand this disease. The pathomechanisms associated with MDD however remain complex, multivariate and to a vast extent unclear. Part of the effort to widen our knowledge focused on its correlates with the structure of the brain.

Identifying significant differences in terms of brain structure between depressive subjects and healthy controls (HC) lead so far to only partly consistent results (Schmaal et al., 2017; Li et al., 2020; Gray et al., 2020). The widespread problems of cohort and research team specific findings played a role in the creation of conflicting reports (Marek et al., 2022; Botvinik-Nezer et al., 2020). Such reported group-wise differences tend to contradict themselves, making them usable for depression classification only to a limited extent. Even significant group-wise statistical differences would not imply guaranteed feasibility of ML classifiers for the task (Arbabshirani et al., 2017).

The idea of classifying depressive subjects and HC using machine learning (ML) based on brain structure still lead to the realization of a plethora of studies (Gao et al., 2018,

review; Patel et al., 2016, *review*). A high proportion of those studies reported high classifier performances (Gao et al., 2018, *review*; Patel et al., 2016, *review*). Those studies however suffered from relatively low sample sizes as well as limited testing for generalizability. Later attempts on larger cohorts for the same task lead to negative results (Flint et al., 2021; Stolicyn et al., 2020; Schulz et al., 2022), so that the realizability of exportable, well performing classifiers yet has to be demonstrated.

Parallel to those developments, the ML branch saw the emergence of a new promising subfield in recent years: automated machine learning (AML) (Waring et al., 2020, *review*). AML algorithmically automatizes part of the ML pipeline design process. This allows for a selection of pipeline components basing on high-speed evaluation of their performance amidst the infinite domain of definition of possible ML structures. As an alternative to already attempted conventional ML approaches, the application of AML may provide useful insights on the problem and allow an improvement of performances. AML however saw until now relatively little usage for neuroimaging related ML tasks (Dafflon et al., 2020; Waring et al., 2020, *review*; Musigmann et al., 2022).

This leads us to the two research aims of our project: First, we tested the applicability of AML to neuromaging tasks with two common and realizable ML tasks, sex classification and age regression. We then attempted to differentiate depressive subjects from HC based on structural neuroimaging techniques and data derived from them using AML.

CHAPTER 2

Literature review

2.1 Depressive disorder

Depressive disorder is a mental disorder defined in the *International statistical classification of diseases and related health problems* (10th ed.; World Health Organization, 1993) (ICD-10) by the following 3 main criteria: persistent sadness or low mood, loss of pleasure in normally enjoyable situations and of interests, and persistent fatigue or loss of energy. According to the *Diagnostic and statistical manual of mental disorders* (5th ed.; American Psychiatric Association, 2013) (DSM-5) and the ICD-10, further symptoms of depressive disorder are: loss of appetite, sleep disorders, diminished concentration, diminished sense of self-worth, unreasonable feelings of self-reproach or inappropriate feeling of guilt, hopelessness and suicidal ideation (World Health Organization, 1993; American Psychiatric Association, 2013). The presence of symptoms can lead to the diagnosis of MDD if they persist over a duration of at least 2 weeks (American Psychiatric Association, 2013; DGPPN et al., 2017). The presence of the 3 main symptoms and of at least 4 side symptoms combined with the time factor are required for this according to German guidelines (DGPPN et al., 2017).

A diagnosis of depressive disorder should be paired with an examination of additional symptoms as well as the patient's medical history (DGPPN et al., 2017; World Health Organization, 1993; American Psychiatric Association, 2013). A depressive disorder that was preceded by maniac or hypomaniac phase(s) leads to the diagnosis of a bipolar disorder (DGPPN et al., 2017; World Health Organization, 1993; American Psychiatric Association, 2013). If psychotic phases or symptoms are known, the diagnosis should be changed in depressive disorder with psychotic symptoms (DGPPN et al., 2017; World Health Organization, 1993;

American Psychiatric Association, 2013). Those two differential diagnoses have considerably different treatment in comparison with the treatment of unipolar depressive disorder without psychotic symptoms and should therefore be considered as different entities (DGPPN et al., 2017; World Health Organization, 1993; American Psychiatric Association, 2013).

The prevalence of depressive disorders among adults is estimated at 5.0% worldwide (Vos et al., 2020). In Germany, it has an estimated prevalence of 10.1% (Bretschneider et al., 2017). The impact of depressive disorders on individual, societal and economic levels tends to be underrated and is on the rise (Vos et al., 2020; McLaughlin, 2011). From 1990 to 2020, the DALY caused by depressive disorders augmented by 60%. In 2019, it was ranked 13th in the leading causes of disability across all ages worldwide (Vos et al., 2020). It ranked as high as 4th leading cause of disability for younger generations globally in that year (10-24 years, Vos et al., 2020). Additionally, depression is also recognized as a major cause for suicide, with an estimated 5% to 8 % of patients with depression dying by suicide (Brådvik, 2018). High prevalence, impact on populations as well as perceived suffering for the patients have motivated research on depression for decades (Holtzheimer 3rd & Nemeroff, 2006).

The origins for depressive disorder are described by the widely accepted biopsychosocial model as a complex interaction between elements of the patient's environment, intrinsic mental processes and biological processes (Engel, 1977). The pathophysiology for depressive disorder is often described at cellular scale using the monoamine hypothesis (Hirschfeld, 2000). This hypothesis explains depressive symptoms through a perturbation in neuronal pathways caused by an imbalance in the presence of neurotransmitters. More specifically, serotonin, dopamine, norepinephrine and epinephrine pathways are associated with this hypothesis, although others have been the subject of studies (e.g. glutamate; Sanacora et al., 2012; Onaolapo & Onaolapo, 2021). It became widely popular due to the fact it explained the efficiency of a broad range of antidepressant medications and is a common part of psychoeducative programs (Rabovsky & Stoppe, 2008). It also has the merit of establishing a bridge between the clinical picture of depression and neurophysiological knowledge. Still, it encountered a lot of criticism and seems incomplete (Moncrieff et al., 2022, *review*; Hirschfeld, 2000; Delgado, 2000). Reasons for this is among others the insufficient amount of evidence

supporting the theory (Moncrieff et al., 2022, *review*). The specific idea of serotonin depletion as a cause for depressive disorder, a wide-spread idea, is also partly contradicted by the mechanism of action of Tianeptine, a serotonin reuptake enhancer, and when induced did not lead to depressive symptoms (McEwen et al., 2010; Delgado, 2000).

As another axis for depression research, anomalies were sought for at the scale of the central nervous system's anatomy (Trifu et al., 2020; Zhang et al., 2018). MDD has been linked to alterations of volume and cortical thickness in different regions. The prefrontal cortex (PFC) is a region located rostral to the primary motor area. It is involved in emotional regulations, social behaviour regulation and planning, among others. The PFC is the region most commonly associated with depression induced transformation (Trifu et al., 2020). Sections commonly affected are the dorsolateral prefrontal cortex (DLPFC), the orbitofrontal cortex (OFC), the middle PFC and the anterior cingulate cortex (ACC). Studies based on magnetic resonance imaging (MRI) have supported the association of reduced PFC volume in MDD, as well as a significant reduction in cortical thickness (CT) (Trifu et al., 2020; Zhang et al., 2018). Other arguments supporting its involvement is the success of high-frequency repetitive transcranial magnetic stimulation (HF-rTMS) used on the left DLPFC in the treatment of MDD (Berlim et al., 2014, *review*). Structural changes affecting the hippocampus in MDD were also reported (Videbech & Ravnkilde, 2004; Trifu et al., 2020). The hippocampus plays an important role in memory recall (Burgess et al., 2002). Meta-analysis have shown a smaller hippocampus in depressed patients in comparison with HCs. Explanation for this is given to the high concentration of glucocorticoid receptors present in the region (R. M. Sapolsky et al., 1984), giving it a sensitivity to elevated cortisol levels (R. Sapolsky, 1985). Cortisol levels have been reported as higher in stress situation and in depression (Dienes et al., 2013; Carroll et al., 2007), although these postulates are still debated (Nandam et al., 2020). Thus, it it hypothesized that cortisol levels reach a toxic level in depression, causing hippocampus volume reduction (Trifu et al., 2020). There is evidence that successful therapy with antidepressant treatment and electroconvulsive therapy (ECT) can revert this effect (Nordanskog et al., 2010). The thalamus is believed to play an important role in information processing forwarding between different areas (Fama & Sullivan, 2015). It is involved in emotion as well as in sleep and wakefulness regulation (Barson et al., 2020; Coulon et al., 2012). The grey matter volume

(GMV) of both sides of the thalamus was found to be reduced in MDD. Lastly, the striatum, the parietal lobe and the dorsolateral prefrontal circuit were all shown to be significantly different in their anatomy when compared with HC.

Studies hinting at structural differences in depressive subjects compared to healthy controls brought the following question: Could those differences be analogous and striking enough to allow for a diagnosis of depressive disorder using solely the data of an sMRI ? There exist no report of such a classification being successfully performed by a trained clinician without techniques derived from the field of ML.

ML is a field of mathematics centered around the design and usage of self-learning algorithms (Rebala et al., 2019, Jung, 2022). It lies at the crossroad between statistics, algorithmics and computer science. ML can be used to create predictive models (that will be called ML or AML models in this work) using a set of input variables called features as input to predict a variable called target (Jung, 2022). Different mathematical objects can be used to design an ML model's architecture (e.g. statistical trees, forest of trees, neural networks). The choice of ML model algorithm, subcomponents and hyperparameters affect its predicting performances and generalization power (Jung, 2022). The performances of an ML model are dependent on the task given to it as well the quality and size of the samples used to train it (Jung, 2022; Marek et al., 2022). ML models are at risk of a phenomenon called overfitting (Jung, 2022; Mutasa et al., 2020; Graham et al., 2019; Ying, 2019). An overfitted ML model predicts accurately on its training sample yet fails to reproduce those performances on new, unknown subjects. ML was used with structural MRI (sMRI) derived features in order to attempt sMRI based depression prediction.

2.2 AML/ML based depression diagnosis

Previous studies have shown that MDD patients could be successfully distinguished from HC using structural imaging data with ML methods. In Patel et al., 2016, 15 of such studies were reviewed, with accuracies ranging between 67.6% - 94.3%. The classifiers used in those studies all originated from conventional ML approaches (i.e., neither deep learning

(DL) nor AML were used). The most used type of algorithms was support vector machines (SVM), with different kernels and normalization methods having been employed. The datasets were all balanced in classes, although relatively small (74 samples in the most important dataset). Performances were mainly measured using accuracy as a metric. The evaluation was performed predominantly with leave-one-out-cross-validation (LOOCV). An other review supported similar findings in terms of accuracy (Gao et al., 2018, *review*). Again, the main metric used was to a large extent accuracy, and the method for evaluation LOOCV. Overall, the results reported in Patel et al. (2016) and Gao et al. (2018) were positive and hinting at the realizability of ML based depression diagnosis using sMRI data.

Those studies however suffered from a number of limitations. First, most of them used relatively small sample sizes (Patel et al., 2016, *review*; Gao et al., 2018, *review*). The majority of them additionally did not make use of separate train and validation datasets. In ML this is a choice that can lead to undetected overfitting (Graham et al., 2019; Flint et al., 2021). Both of these restrictions come with a considerable risk of reporting overoptimistic results, and should therefore be taken with caution.

The reported positive results of such studies was questioned in Flint et al. (2021). A competition called Predictive Analytics Competition (PAC) took place in 2018. Approximately 170 ML experts split across 49 ML teams from around the world competed to train the best ML based predictor for depressive disorders, based on sMRI. First, every team was provided with the same training dataset, that included the sMRI as well as diagnosis and confounders for each subject. Each ML team then constructed the best ML model they could using this dataset. In a second phase, each team received a new dataset to evaluate performance, that we will call test dataset. The test dataset included again sMRI and confounders for each subjects. This time though, the diagnosis were not provided. The ML models produced previously were used to predict whether those new subjects were HC or patients. It is on their performance for the prediction on those unknown subjects that each ML team was ranked. In contrast with previously mentioned studies, we here had a strict separation between the dataset used to find the optimal model and the validation dataset.

Results from studies such as those mentioned above (Patel et al., 2016, *review*; Gao et al., 2018, *review*) had led to a conservative expectation for results reaching at least the 80% accuracy threshold. The accuracies of the winning classifiers were much lower, ranging between 60% and 65%. The important discrepancy between the competition's results and prior findings highlighted the drastic limitations of the latter.

The explanations for the divergence between the previous, small-sample studies and the results of the PCA 2018 are multiple (Patel et al., 2016, *review*; Gao et al., 2018, *review*; Flint et al., 2021). First, the inverted proportionality between accuracy and number of samples recently identified for the task is very likely to have played a role in it (Schulz et al., 2022; Flint et al., 2021). The tendency of studies to show overly positive results was thus attributed to models being able to adapt to patterns specific to the samples of the dataset when the population was small enough (Flint et al., 2021; Graham et al., 2019). Aside from this, the risk of misestimation of the accuracy was linked with the use of LOOCV as evaluation method (Flint et al., 2021). LOOCV was shown to lead to non-generalizable estimations of performances with high variance (Flint et al., 2021; Varoquaux et al., 2017). The absence of separation between train and dataset in most publications probably further contributed to overoptimistic reports (Iniesta et al., 2016). Overall, the ML predictors built previously were deemed as being probably unable to export to unknown samples. The negative result of the PAC 2018 on a large validation sample for depression classification with sMRI data was reproduced on a different sample in Stolicyn et al. (2020).

Thus, it still remains unknown if ML can successfully be used to perform prediction of depressive disorder on unknown subjects using sMRI based data.

2.3 Rationale for the use of AML

AML is an emerging subfield of ML were different ML tasks such as selecting the data preprocessors, feature preprocessors, optimal predictive models and architecture as well as optimising the models' hyperparameters is solved algorithmically using different possible meta-heuristics. Although first attempts at automating the process of ML pipeline design

already existed in the 1990's, the field gained more attention in the last decade (Zöller & Huber, 2021). This is due partly to vast improvements in available computing power as well as in the existing AML implementations (Zöller & Huber, 2021). A likely explanation for this development is certainly the increased demand for ML experts and shortages in that regard.

AML aims at improving the quality of ML pipelines designed for common predictive tasks. This is achieved by using different automatized metaheuristics in order to efficiently select an ML architecture amidst the infinite possibilities available. Where human operators often select among a limited subset of possible architectures due to human-time limitations (Zöller & Huber, 2021), AML can rapidly evaluate thousands of different combinations of ML algorithms. AML also vastly accelerates parts of the ML development by automatizing time-consuming tasks (e.g. hyperparameter optimization), thus allowing ML operators to reallocate their productivity in a more optimized ways. A rapid integration of AML in common ML workflows is to be expected in the current context due to the great demand in ML experts (Musigmann et al., 2022).

Automation is inherent to the ML field. The extent to which the ML pipeline creation and optimization process is automated is to be understood as a continuum, and not as a binary attribute (Zöller & Huber, 2021). Methods designed to automatize part of the ML optimization process exists. Grid-search or random-search can for example be used to automatize hyperparameter optimization (Liashchynskyi & Liashchynskyi, 2019), while recursive feature elimination is an example of automatic feature selection method (Escanilla et al., 2018). AML brings the logic of automatizing the ML design process to a further extent (Zöller & Huber, 2021). Modern AML libraries can, with limited intervention from a human operator, create a prediction-ready AML model by simply analyzing a dataset given as input (Feurer et al., 2019; Olson & Moore, 2016). Even though partial automations of the ML model design process are common, the use of fully automated ML design processes understood under AML is relatively new in the medical field (Waring et al., 2020, *review*; Dafflon et al., 2020; Musigmann et al., 2022).

Although AML has progressively gotten more attention in the field of medical research (Waring et al., 2020, *review*; Dafflon et al., 2020; Musigmann et al., 2022), there is still little

literature on the usage of AML in neuroimaging (Dafflon et al., 2020; Musigmann et al., 2022). Encouraging results were already reached for the tasks of age predictions based on sMRI, with performances nearing the ones of conventional ML methods (Dafflon et al., 2020). Furthermore, AML performed convincingly in a small-sample study for the prediction of the achievability of resection of meningioma (N=138; Musigmann et al., 2022). Additionally, as previously described (2.2), conventional ML models were already tested extensively for the task of diagnosing depressive disorder. We thus decided to use AML for the task at hand as a new and potentially advantageous option.

2.4 Objectives of the study

First objective: AML benchmark

The applicability of AML to neuroscientific problems acquired some attention (Dafflon et al., 2020; Musigmann et al., 2022), yet still deserves further exploration. Therefore, the first part of our study consists in testing AML models for the prediction of two targets where ML based prediction was already proven to work, namely on age and sex prediction (Chekroud et al., 2016; Jiang et al., 2020). For this aim, we used sMRI from a pool of subjects from the 1000BRAINS study (Caspers et al., 2014). Testing on sex and age allows us to test AML for both regression and classification tasks.

During this part of the study, the effect of the dataset (train, validation or test) chosen for metrics evaluation on reported performance will furthermore be analyzed. This will be achieved by comparing the evaluation performances on each dataset.

Second objective: Depression classification

Whether depression can be diagnosed using ML models based on sMRI derived features or not remains an open question. To our knowledge, little to no studies were conducted using AML instead of classical ML methods for this task. After the establishment of a baseline in the first part of our project, we turned to depression diagnosis using data from the BiDirect study.

The generalization power of AML models will be assessed in this step. We will furthermore perform an analysis of the effects of confounders on prediction accuracy for depression diagnosis.

CHAPTER 3

Material and Methods

This chapter describes the protocols that were followed in order to answer our two aims, from the raw data acquisition to performance evaluation of our AML models. The material we had access to and its preprocessing is initially described. We then detailed the protocol followed for our first objective of performing an AML benchmark along with its rationale. The steps followed for our second objective, the depression classification, AML tasks are finally reported.

3.1 Material

As detailed in Falk et al., 2013, cohort-specific, non-generalisable findings are an ubiquitous problem in neuroscience. In order to minimize these, populations from different studies were acquired for our experiments. Here, we first provide context and descriptions on those: The 1000BRAINS study's sample (Caspers et al., 2014) was used for our AML Benchmark; the BiDirect study's sample (Teismann et al., 2014) for depression classification. The way from raw MRIs to the AML ready features is further detailed; along with the steps for subject inclusion/exclusion.

3.1.1 1000BRAINS

1000BRAINS (Caspers et al., 2014) is a longitudinal study that aims at winning insights in the effects of aging on the function and the anatomy of the human brain. For this purpose, subjects were recruited from the cohorts of two different sources (Caspers et al., 2014; Erbel et al., 2012; Erbel et al., 2010). Subjects from 10-year follow-up cohort of the German Heinz

Nixdorf Recall (HNR) Study as well as from the HNR MultiGeneration Study (Erbel et al., 2012; Erbel et al., 2010). Those studies' main focus lies on the estimation of cardiologic pathologies and risk factors affiliated to those (Erbel et al., 2012; Erbel et al., 2010). Exclusion criteria were related to MRI and strong magnetic field exposure (stents, pacemaker, surgical implants or protheses, claustrophobia, tatoo, history of neurosurgery). Since 1000BRAINS is a population-based study, the only exclusion criteria were based on the eligibility of the MR (Caspers et al., 2014). An extensive protocol including sMRI, functional MRI (fMRI) and diffusion tensor imaging (DTI) was followed by the subjects. Various psychological variables were also collected using different questionnaires. All subjects provided written informed consent prior to inclusion (Caspers et al., 2014). The study protocol was approved by the Ethics Committee of the University of Essen, Germany (Caspers et al., 2014).

The process of subject selection is here described. For a visual representation, see 1. We initially had access to the data of 1314 subjects. We selected those with available sMRI, proceeded to preprocess their data using Freesurfer's pipeline (Fischl, 2012) and structured the output in the dataset we used later on. 124 subjects were excluded because of missing MRI scan, failed MRI preprocessing or missing demographic data (here age and sex). A quality control (QC) (Shewhart & Deming, 1939) was then performed, resulting in 33 exclusions. A final sample of 1157 subjects was present in our main dataset for the 1000BRAINS study (mean_{age} = 61.01 years, SD_{age} = 12.84 years, min_{age} = 18.50 years, max_{age} = 85.40 years, $\text{percent}_{\text{female}}$ = 46%).

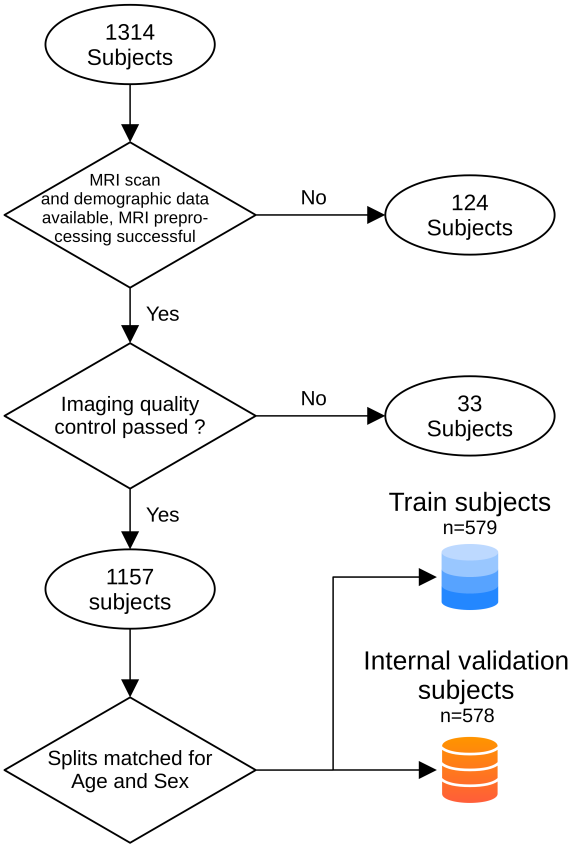

Fig. 1: Subject selection for the 1000BRAINS sample

3.1.2 BiDirect

The BiDirect Study (Teismann et al., 2014) is a prospective observational study of 3 cohorts designed with the aim of exploring the bidirectional effects of depression and (subclinical) arteriosclerosis. The study was approved by the ethics committee of the University of Münster and the Westphalian Chamber of Physicians in Münster, North-Rhine-Westphalia, Germany (Teismann et al., 2014). Written informed consent for participation in the study was obtained from all participants (Teismann et al., 2014).

The data of 2 from the 3 main cohorts of the study were employed for our investigation. The first cohort we used is the MDD cohort. It has a population size of 999 patients. These patients suffered an episode of depression at the time of recruitment. Additional inclusion criteria were: 1. Age between 35 years and 65 years (included) 2. Ongoing treatment of

acute depression. Exclusion criteria were: 1. Additional diagnosis of dementia 2. Additional diagnosis of drug abuse 3. Compulsory admission. Diagnosis of depression was conducted at recruitment in one of the 6 centers that are part of the study and limited to those who had been hospitalized due to a depressive episode at least once during the last year prior to recruitment. Potential participants for the study were evaluated by certified psychologists, after which eligible patients were invited to participate in BiDirect-Baseline. The second cohort we used is the control cohort. It is composed of 912 subjects of age ranging between 35 and 65 years (included).

Participants followed an extensive data collection procedure that further assessed the presence or absence of depressive symptoms. They were required to fill the modules A, A', B, D, and O of the Mini-International Neuropsychiatric Interview (MINI). The Hamilton Depression Rating Scale (HAM-D-17) and the 14 items version of the Hamilton Anxiety Rating Scale (HAM-A-14) were also used for that purpose (Hamilton, 1960; Hamilton, 1959).

The acquisition of the BiDirect sample was again followed by exclusions of certain subjects in parallel to the preparation of the data. The subjects were excluded when their sMRI was not available. This resulted in 1428 subjects' sMRIs being preprocessed using Freesurfer's pipeline (Fischl, 2012). For each ML task, the same sets of features were calculated as for the benchmark. A QC was performed which lead to 34 exclusions (Shewhart & Deming, 1939). The availability of age, sex and the depression/HC variable of each subject was controlled for missing values, leading to 12 subjects further exclusions. A supplementary step was taken in this task to ensure that demographic data had no statistical effects on the diagnosis using propensity score matching, resulting in 280 subjects sorted out (Austin, 2011). In total, 1102 subjects were present in our datasets for the BiDirect study (mean_{age} = 51.08 years, SD_{age} = 7.51 years, min_{age} = 35.30 years, max_{age} = 70.05 years, $\text{percent}_{\text{female}}$ = 58%).

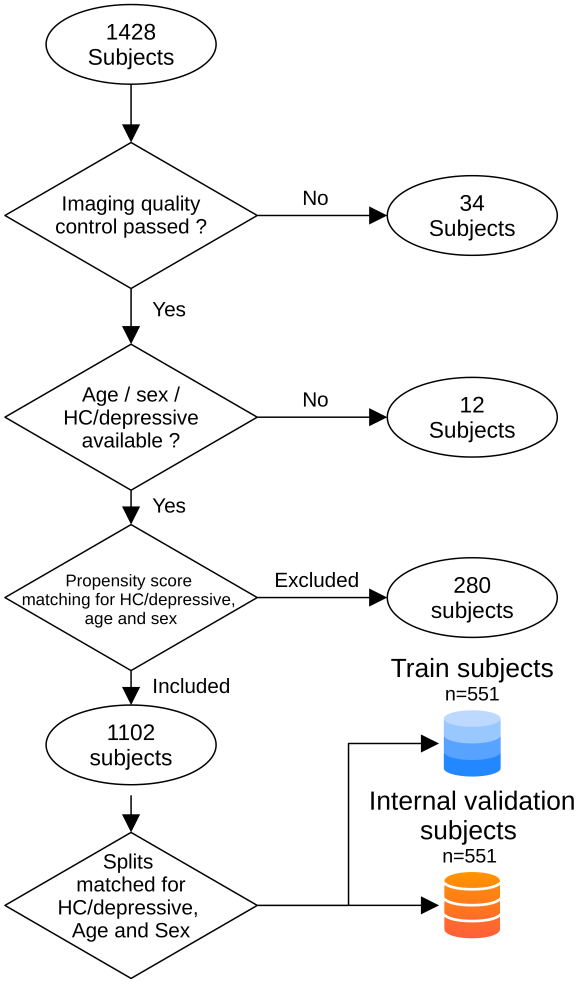

Fig. 2: Subject selection for the BiDirect sample

3.1.3 Image acquisition and structural image preprocessing

The acquisition of the structural MRIs used in this work from 1000BRAINS was performed on a 3 Tesla MR scanner (Caspers et al., 2014). The structural data used in our study is derived from the anatomical 3D T1-weighted Magnetization Prepared - RApid Gradient Echo (MP-RAGE) sequence (176 slices, repetition time (TR) = 2.25 s, echo time (TE) = 3.03 ms, inversion time (TI) = 900 ms, field of view (FoV) = 256 × 256 mm^2, flip angle (FA) = 9°, voxel resolution (VR) = 1 × 1 × 1mm^3; Caspers et al., 2014). From the BiDirect study, we used the T1-weighted 3D anatomical images in the next preprocessing steps (TR = 7.26 ms,

TE= 3.56 ms, FA = 9°, 256 × 256 mm^2, VR = 1 × 1 × 1mm^3; 2 signal averages; Opel et al., 2019; Teuber et al., 2017).

ML and AML models' performances are sensible to the peaking phenomenon which requires restricting the number of features used in prediction (Sima & Dougherty, 2008). The peaking phenomenon describes the following tendency: With an always growing number of features, and with a constant number of subjects in the training sample, the performance of a predictive algorithm tends to first grow, but then decreases for every additional feature of the set (Sima & Dougherty, 2008). Large number of features may also result in a loss of generalization power (Dougherty et al., 2009). Voxel-wise analysis of the sMRI would result in features sets with a magnitude exceeding the million features, submitting our AML models to the adverse consequences of the peaking phenomenon. Moreover, AML models using voxels as input and trained with one MRI acquisition configuration could not export well to another one due to varying matrix configurations. We responded to those technical difficulties by parcellating the sMRI, thus reducing the number of features and improving the generalizability of our AML models to other sMRI acquisition configurations. We used a resting state fMRI (rs-fMRI) based parcellation, because they may allow to establish a connection between structural anomalies and anomalies in the function of cognitive networks known in depressive disorder (Dai et al., 2019, *review*).

Anatomical preprocessing of the data was accomplished using Freesurfer (Fischl, 2012) v7.1.0. on the T1-weighted 3D anatomical images. The original pipeline was adapted to also include the 400-node Schaefer parcellation (Schaefer et al., 2018), which is based on cortical surface models calculated from rs-fMRI measurements of 1489 participants using a gradient-weighted Markov Random Field (gwMRF) approach (Kindermann & Snell, 1980). First, the parcellation was transformed to individual space using FreeSurfer's mris_ca_label tool (Fischl, 2012). Then, morphology values were gathered for every transformed node using FreeSurfer's mris_anatomical_stats tool (Fischl, 2012). For each node, the following features were determined: the surface area in mm^2, the gray matter volume in mm^3 and the average cortical thickness in mm.

3.1.4 Quality control and propensity score matching

A quality control was performed for the sample of each study after the features were computed (CT, GMV and SA; Shewhart & Deming, 1939). The average GMV, average CT and average surface area (SA) were calculated for each participant and used for the QC. A deviation from 2.68 standard deviation (SD) for one of those values lead to the exclusion of the subject.

AML models may capture the effect of confounders to draw inference on the target, an effect we wanted to restrict. We theorized that the effects of age and sex on the depression/HC variable may be used by our depression predictors in such a way. A last step was thus undertaken in order to limit the effect of the sex and age variable on the depression/HC variable for the BiDirect dataset, using propensity score matching (Austin, 2011). This resulted in 280 subjects being excluded from our BiDirect dataset.

3.2 First objective: AML benchmark protocol

Whether depressive states can be predicted or diagnosed using structural MRI is still up to debate (Flint et al., 2021). The efficiency of AML in the field of neuroscience, although beginning to receive attention in the last years, also requires further investigations (Dafflon et al., 2020; Musigmann et al., 2022). Therefore, a failure to predict depressive states with AML and structural MRIs could either be inherent to the task, due to the use of AML, or both. A benchmark was performed with the targets sex for a classification ML task and age for a regression ML task using Auto-sklearn as our AML library (Feurer et al., 2019). Those two targets are predictable to a certain extent using region-wise GMV, CT and SA with standard ML models (Chekroud et al., 2016; Jiang et al., 2020). They were therefore used to test the efficiency of our AML models before application to depression prediction.

In this section, we describe the protocol that we followed for the AML benchmark once the data were preprocessed for the creation and performance evaluation of our AML models. For a visual representation, see fig. 3. First, the included subjects for 1000BRAINS were split in a train- and an internal validation dataset. We then created multiple datasets, using the same

subjects' train/validation partitioning but different feature sets. Afterwards, the creation of our AML models was performed for each feature set using the subjects of the 1000BRAINS train dataset. Lastly, the performance of the so designed AML models were tested separately on the train dataset, the internal validation dataset and the BiDirect dataset that served as the external validation dataset.

3.2.1 Preparation of the train and internal validation datasets

Building a successful AML model requires adequate preparation of the raw data in datasets for designing, training and testing purposes (Jung, 2022). The split in train and validation is a splitting that occurs between subjects and does not involve any form of feature selection (Jung, 2022). This measure serves the purpose of limiting the risk of undetected overfitting and overoptimistic positive results (Arbabshirani et al., 2017). An AML model is said to be overfitting when it performs well on the dataset used to design it, however is unable to reproduce such performances on new, unknown samples (Mutasa et al., 2020; Jung, 2022). Overfitting is due to an AML model having adapted to characteristics and patterns specific to the train dataset without having found an actual abstract and generalizable rule to solve the problem (Mutasa et al., 2020; Jung, 2022, Ying, 2019). It is an ubiquitous problem in ML (Ying, 2019).

The 1000BRAINS subjects were thus split into a train and an internal validation dataset (Jung, 2022). The train and the validation datasets had strictly different samples from each other and were designed so as to be demographically homogeneous (Jung, 2022). Dimensions for a train/test split usually range from a 50% / 50% of the main dataset for respectively the train dataset and the test dataset to 90% / 10% of the main dataset (Rácz et al., 2021). There exist no strict recommendation on the size of splits to choose (Jung, 2022). We opted for 50% / 50% splits in order to maximize the size of the internal validation datasets and thus limit the variance of reported results (Wickenberg-Bolin et al., 2006). A test dataset may potentially also be extracted from the main dataset, distinct from the train and validation dataset. This was not done for this aim, since we had access to the sample of the BiDirect study for external validation (Teismann et al., 2014).

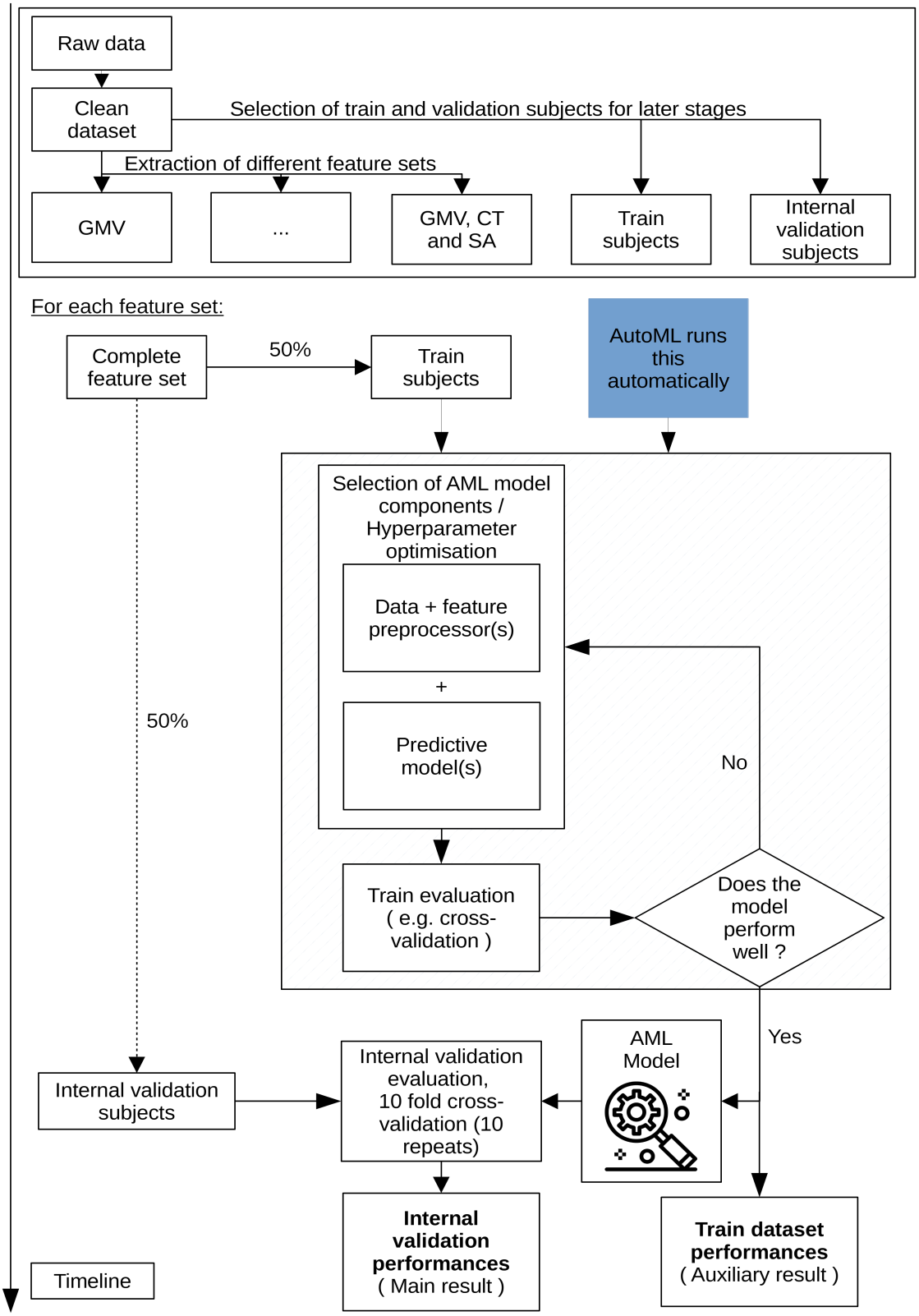

Fig. 3: Path from raw data to a prediction-ready AML model for each AML task. The hashed space represents the part of ML that is run automatically by most AML libraries. This workflow was followed for the sex classification, the age regression and the depression classification tasks separately.

The practice of creating separate train, internal validation and external validation datasets serves the purpose of evaluating performances on future unknown samples in the last steps of the process as realistically as possible (Jung, 2022; Cabitza et al., 2021; Bleeker et al., 2003). Previous studies attempting the diagnosis of depressive disorders based on sMRI have mostly trained and tested on the same dataset (Patel et al., 2016, *review*; Gao et al., 2018, *review*). While this is more time and ressource efficient, designing an AML model and evaluating its performances on the same dataset comes with a high risk of overfitting, thus reporting overoptimistically high and non generalizable performances (Flint et al., 2021). Splitting the experience dataset in an AML design (or train) dataset and an internal validation dataset represents a necessary measure against such a risk (Jung, 2022). Ideally, this evaluation should be followed by an external validation with datasets from another source (Bleeker et al., 2003; Cabitza et al., 2021; Rose, 2018).

Our splitting of the initial dataset (A) in a train (B) and an internal validation dataset (C) of equal size (+/- 1 subject) followed the here detailed procedure. A random pairwise matching algorithm was implemented using a parameter list of matching-relevant features, composed of either continuous or binary features (e.g.: age, then sex). The order played a role, as the matching algorithm prioritized selecting similar values for the feature provided first. A random seed allowed to repeat the matching or to perform it in a different order. The algorithm ran in a recursive fashion with the starting dataset A being reduced by two subjects in each iteration (with the exception of the last were 3 subjects could be substracted) while datasets B and C gained one subject per iteration (with again the exception of the last iteration). For each subject in the dataset A, a subset Y of matching yet non-matched subjects was selected for the first of the feature to match for. From this subset, another subset of matching subjects was then selected for the next matching-relevant feature, and so on. This procedure stopped when every feature in the list of matching-relevant features was used to filter the successively produced subset. A subject was then selected randomly from the final subset. The original and the selected subjects were then stored, one in the datasets B, the other in C, and subtracted out of A. This procedure was repeated until there are no subjects left in the dataset A.

Using this procedure with the arguments sex followed by age to split the 1000BRAINS dataset resulted in homogeneous demographics in the train and internal validation datasets, presented in table 1.

	Train dataset	Internal validation dataset	Complete dataset
Number of subjects	579	578	1157
Age (in years)	61.07 ± 12.94	60.95 ± 12.74	61.01 ± 12.84
Sex (percent of female)	46%	46%	46%

Table 1: Demographics of the 1000BRAINS complete, train and internal validation datasets

3.2.2 1000BRAINS feature sets preparation

The subject-wise splitting of the initial 1000BRAINS dataset was followed by the feature-wise splitting in multiple datasets. Different feature sets were prepared as input to the AML Benchmark: (i) GMV for each node (400 features), (ii) CT for each node (400 features), (iii) SA for each node (400 features), (iv) GMV, CT and SA for each node (1200 features).

The estimated total intracranial volume (eTIV) is a variable known to interfere with the prediction of various variables. As shown in 2, eTIV and the structural parameters we used tend to be intercorrelated. In Hilger et al., 2020, the authors therefore chose to regress the factor eTIV out of the variables and used the residuals as input to perform the ML tasks with both the raw data and residuals in different feature sets. We chose this approach too.The same four feature sets combinations with eTIV regressed out were computed and added as feature sets. As such, we derived a total of eight feature sets to be tested in our AML benchmark, that both were used for sex classification and age regression.

Pearson corr. coef.	eTIV	Average volume	Average thickness	Average surface
eTIV	-	0.61***	0.11***	0.62***
Average volume	0.61***	-	0.44***	0.82***
Average thickness	0.11***	0.44***	-	-0.13***
Average surface	0.62***	0.82***	-0.13***	-

Table 2: Pearson correlation coefficient between eTIV, average volume, average thickness, average surface in the 1000BRAINS dataset. Significant correlation between eTIV and each of those features could be shown.

3.2.3 AML model design

A predictive AML model was designed with optimized hyperparameters for each feature set using solely the subjects of the train dataset. We here provide a description of the algorithms used for AML model design along with their search space. The rationale behind AML library choice, search space, search setting and evaluation metrics is here detailed. The ML algorithms included in the AML search space are of importance to report, because they provide an overview of the options that were evaluated algorithmically for future research projects. At the end of this phase, one AML model was available for each of the 8 feature sets of the AML benchmark.

3.2.3.1 AML library selection

The rationale behind the selection of AML library is described here. Different AML libraries were designed and made available, with no guideline available on which one to use in which situation. Auto-sklearn (Feurer et al., 2019), TPOT (Olson & Moore, 2016), Auto-Weka (Thornton et al., 2012) as well as H2O (LeDell & Poirier, 2020) are just a few examples among the libraries developed for this purpose.

Two important criteria for the selection of an AML library are: performances reported in the literature, and interpretability of the final AML models for a human operator. For the diagnosis task, the priority consists in finding a model that performs better than a random predictor. This is why we prioritized performance over interpretability of the outputted AML models in this work. In the event of accurate AML models intended for clinical use, the explainability of the decisional algorithm should be granted, so that priorities should be changed (Vilone & Longo, 2021).

In Zöller & Huber, 2021, AML libraries were benchmarked on 114 publicly available real-world datasets. It appeared that the performances reached by the different competing algorithms were on average very similar, with a maximum performance difference of only 2.2%. For most datasets, the performance differences were not significantly different. However, this benchmark did not include Auto-sklearn 2.0 (Feurer et al., 2021), that seemed to outperform

other AML libraries for classification tasks. The possible superiority of Auto-sklearn in classification tasks is also supported in Balaji & Allen, 2018. The study however shows that TPOT outperforms other AML libraries when it comes to regression tasks. Based on this literature, we selected Auto-sklearn 2.0 as our main AML library. We furthermore decided to compare the results of TPOT and Auto-sklearn 2.0 for the diagnostic task.

Auto-sklearn

We here provide a short description of Auto-sklearn (Feurer et al., 2019), that we used to perform the majority of AML model search. Auto-sklearn is an AML library focused on optimizing ML ensembles, creating ensembles of up to 50 ML pipelines including data and feature preprocessors in its default configuration. It takes advantage of a large variety of preprocessors and classifiers from the scikit-learn library (Pedregosa et al., 2011). Auto-sklearn has multiple implementations. For regression tasks, we used a recent version of Auto-sklearn 1.0, the version 0.13.0. An other version of Auto-sklearn specifically optimized for classification, Auto-sklearn 2.0 (from the version 0.13.0), is used in our classification tasks.

For each run, we used the default search spaces provided in Auto-sklearn. The search-space of Auto-sklearn 2.0 is described in Feurer et al., 2021. The following classifiers are included in Auto-sklearn 2.0: Extra Trees, Gradient Boosting, Multilayer perceptron, Passive aggressive-classifier, Random Forest and Stochastic gradient descent. Following preprocessors are available: categorical encoding, category coalescence, imputation of missing values, rescaling, quantile transformer, robust scaling.

TPOT: Tree-based Pipeline Optimization Tool

In order to offer comparison for our Auto-sklearn AML models, AML models for depression classification were additionally searched for using Tree-Based Pipeline Optimization Tool (TPOT) (Olson & Moore, 2016). TPOT also bases on preprocessors and classifiers from the scikit-learn library (Pedregosa et al., 2011), as well as custom models proper to TPOT only. Searching for a well-performing pipeline is here done using a genetic-algorithm inspired

heuristic as implemented in the Python package Distributed Evolutionary Algorithms in Python (DEAP). We allowed TPOT to perform its generational optimization process for a maximum of 100 generations.

The following models are included in the search space we used for TPOT: decision trees-, random forests-, eXtreme Gradient Boosting- and k-Nearest Neighbor classifier as well as logistic regression. Available preprocessors are: standard -, robust -, min-max -, and MaxAbs scalers, as well as randomized PCA, binarizing, and polynomial features.

3.2.3.2 Evaluation metrics

Both Auto-sklearn and TPOT require a main evaluation metrics for the AML design process. The balanced accuracy (BA) was used as the main performance indicator in order to monitor the efficiency of our classifiers in the training-, validation- and test-phases. The BA score is a variation of the accuracy score that takes in account the number of occurrence of a class in the dataset. For a binary classification problem, it is calculated as follows:

$$BAS = \frac{TPR + TNR}{2}$$

with:

BAS = BA score
TPR = true positive rate
TNR = true negative rate

An important advantage of BA is its robustness to class imbalance. Imbalanced classes are a problem that arises when some of the classes (e.g. the HC) are more represented than others (e.g. the depressive subjects / subjects screened positively) in available datasets. Due to this problem, the use of accuracy as a main or unique performance evaluation in neuroscience ML tasks has been described as problematic (Alberg et al., 2004). Accuracy and BA should be similar or equal in values for our datasets, since the classes to be predicted were balanced. Generally speaking though, we see BA as a more representative and reliable measurement

and for this reason decided to use it as our main metric. In the case of binary classifications, BA equals the arithmetic mean of sensitivity and specificity.

Our regressors were built using mean squared error (MSE), which is defined as followed:

$$MSE = \frac{1}{n}\sum_{i=1}^{D}(X_i - Y_i)^2$$

with:

n = Number of samples
X = Predicted values
Y = True targets

A common problem in age regression is that the limits of the targets' domain of definition are generally poorly predicted (Smith et al., 2019). Subjects with true target at the lower end of the domain tend to be overestimated, subjects at the higher end underestimated. MSE punishes misestimation of outliers more severely than functions such as mean absolute error (MAE). This is why we selected MSE for the AML model building phase. For comparisons with results from previous studies, we also calculated the MAE.

Once the subjects' splitting, features sets' preparation and AML library's configuration was completed, the AML design process was started. For the Auto-sklearn based model search, each pipeline-building algorithm was given a maximum run time of one day with a maximum CPU-usage of 40 logical CPUs. The maximum size of models in memory was capped to 12GB. The runs lasted 8 days for each of the 2 tasks. The regression tasks ran with Auto-sklearn (0.13.0). In this version of Auto-sklearn, the evaluation processes in the inward loop of the search process can be either set manually or use a default configuration. We opted for a 10 fold cross-validation. The classification task ran with Auto-sklearn 2.0 (v 0.13.0). Auto-sklearn v2.0 uses meta-learning and selects the evaluation methods for the inner-loop automatically, based on knowledge of runs of real-world datasets.

3.2.4 Performance evaluation of the Benchmark's AML models

The AML design phase was followed by a thorough evaluation of performances that took place in multiple phases. The evaluation on the train and internal validation datasets consisted in a 10 fold cross-validation (10 repeats) on the train dataset first, then on the internal validation dataset. The external validation consisted in training the AML models on the complete dataset of 1000BRAINS and evaluating its performance on the BiDirect sample. Comparing the results of the internal and the external validation allowed us to get an insight on the magnitude of overfitting to the train population. It also provided an insight on how well AML models would export to samples of other studies, acquired with different scanner modalities and with some differences in demographic data. The results of the successive evaluations are presented in section 4.1.

3.3 Second objective: Depression classification protocol

3.3.1 Preparation of the train and validation datasets

The subjects were split in order to form the train- and the validation datasets using the algorithm described in subsection3.2.1. The splits were performed with a 50% to 50% distribution, matching primarily for HC/depressive patient followed by sex, followed by age. As we can see in 3, this results in homogeneous demographics.

	Train dataset	Validation dataset	Complete dataset
Number of subjects	551	551	1102
Healthy controls / total	50%	50%	50%
Age (in years)	51.14 ± 7.67	51.01 ± 7.35	51.08 ± 7.51
Sex (percent of female)	58%	58%	58%

Table 3: Demographics of the BiDirect datasets

The absence of influence of age on the diagnosis variable was inspected. When analysed with an independent t-test, there was not a significant difference in the age distributions for healthy controls (M = 51.14, SD = 7.79) and depressed patients (M = 51.01, SD = 7.22); t(1100) = 0.30, p = 0.76 in the BiDirect main dataset. Furthermore, there was not a significant difference in the age distributions for healthy controls (M = 51.21, SD = 7.92) and depressed patients (M = 51.08, SD = 7.43); t(549) = 0.20, p = .84 in the BiDirect train dataset and no significant difference in the age distributions for healthy controls (M = 51.07, SD = 7.68) and depressed patients (M = 50.93, SD = 7.01); t(549) = 0.22, p = .82 in the BiDirect validation dataset.

The absence of influence of sex on the diagnosis variable was also examined. A chi-square test of independence showed that there was no significant association between sex and depression in the BiDirect main dataset, X^2 (1, N = 1102) = 0.54, p = .46. Furthermore, a chi-square test of independence showed that there was no significant association between sex and depression in the BiDirect train dataset, X^2 (1, N = 551) = 0.23, p = .63 and a last chi-square test of

independence showed that there was no significant association between sex and depression in the BiDirect validation dataset, X^2 (1, N = 551) = 0.22, p = .64.

Base structural and confounder datasets

A total of 9 feature sets was first prepared for AML model design, train and validation. The 8 first contained the same features as the 8 datasets of the AML benchmark, calculated for the BiDirect population this time. One additional feature set was prepared where only the demographic data, i.e. age and sex, was used. Rationale for this is that in the PAC 2018 (Flint et al., 2021), one of the leading 5 teams reported using solely demographic data together with total GMV to train their ML models and make predictions. This feature set serves as a further, necessary measure to test the influence of confounders on AML model performances.

Feature sets based on functional networks

Studies reported functional anomalies in MDD patients in comparison with HC (Castanheira et al., 2019). Those changes may be concomitant with regional structural changes. In order to investigate whether changes of the individual functional networks were accompanied by significant changes in brain structure, we used the structural data of each network as feature sets for depression classification. Thus, we created 14 additional feature sets that are based on the regions of the 7 functional networks as defined in Schaefer et al., 2018. For each network, one feature set with the raw structural data and one with the residuals after regressing the factor eTIV out was used.

Feature sets based on anatomical priors

We decided to additionally perform depression classification with a restricted amount of features based on anatomical priors. Restricting feature with the help of domain-specific knowledge may improve AML model performances (Remeseiro & Bolon-Canedo, 2019, *review*). Here, we describe the method and sources used in this intent.

We used the meta-analysis described in Schmaal et al., 2017. Regions with statistically relevant differences among adults in CT bewteen HC and MDD patients (FDR P-value < .05) were selected. This resulted in 13 regions from the Desikan atlas (Desikan et al., 2006) being included in the feature set. Our previous preprocessing protocol outputted features according to the Schaefer parcellation. The conversion from one atlas to the other was realized with Freesurfer (Fischl, 2012). Every Schaefer-node being included at 50% or more of its surface in one of the relevant Desikan regions was included (Schaefer et al., 2018). This resulted in 54 Schaefer-nodes being incorporated in this AML task for each of the 8 base feature sets, hence an additional 8 feature sets for our analysis.

3.3.2 AML model design

Since depression diagnosis is a classification task, we used the same AML settings as during the sex classification task of the AML benchmark, as described in 3.2.3. Again, each AML algorithm was given a maximum run time of one day with a maximum CPU-usage of 40 logical CPUs per feature set and ran only on subjects of the train dataset. The maximum size of models in memory was capped to 12GB. The depression classification model design was performed with Auto-sklearn 2.0 (v 0.13.0). At the end of this process, for each of the thirty-one feature sets, one AML model was ready for the evaluation round.

In order to compare our AML models with the ones of another library, we additionally generated a TPOT AML model for the 8 base feature sets as well as the age + sex dataset. For the TPOT-based model search, the population per generation was set to 200 along with a maximum of 100 generations, for a maximum of 20200 possible pipelines being evaluated. An early-stop policy was also set, with runs being terminated if there were no improvement in performance after 10 generations. A mutation rate of 0.9 with a crossover rate of 0.1 was used. Performance of the pipelines were measured in the inner loop with a 5 fold cross-validation process.

3.3.3 AML performance evaluation

The evaluation was performed in 2 rounds: once in 10 folds cross-validation (10 repeats) on the train dataset, then again on the validation dataset. The performances of the TPOT AML models were evaluated using the same scheme. At the end of the validation round, the results did not require further confirmation because of negative results, so that no external validation was performed for our second objective. The results for this part of the study are presented in section 4.2.

CHAPTER 4

Results

4.1 First objective: AML benchmark results

The sex classification and age regression AML model design phases ran for 8 days each, with a capped maximum of 30 CPU made available. We thus allowed Auto-sklearn to search for model very extensively in comparison with commonly used configurations (Balaji & Allen, 2018; Feurer et al., 2021). At the end of this procedure, for each dataset, a functioning AML model was created. The algorithmic composition as well as the performances of these are described in this section.

4.1.1 AML models' validation performances

The performance of the AML models in the cross-validation evaluation on the validation sample are presented in figure 4 (for additional scoring information, see table 5, table 6 and table 7). The best results were achieved with a combination of the 3 modalities across targets and input data. A BA of 87±4% was reached by the top AML model for sex prediction with raw data. The top model for age prediction reached a mean absolute error of 5.90±0.62 years. The results obtained by the best models for both tasks outmatch by a large margin those of a random prediction (87% vs. 50% BA for sex classification; 5.9 years vs 9.73 years MAE for age regression). Raw data and residuals led to similar AML performance for age prediction. This, however, was not the case for the sex classification. Here, raw data provided better ML results for all modalities except CT (Raw: 57±7% BA, Residuals: 72±6.7% BA). Overall, classifiers as well as regressors displayed satisfying capabilities in the internal validation.

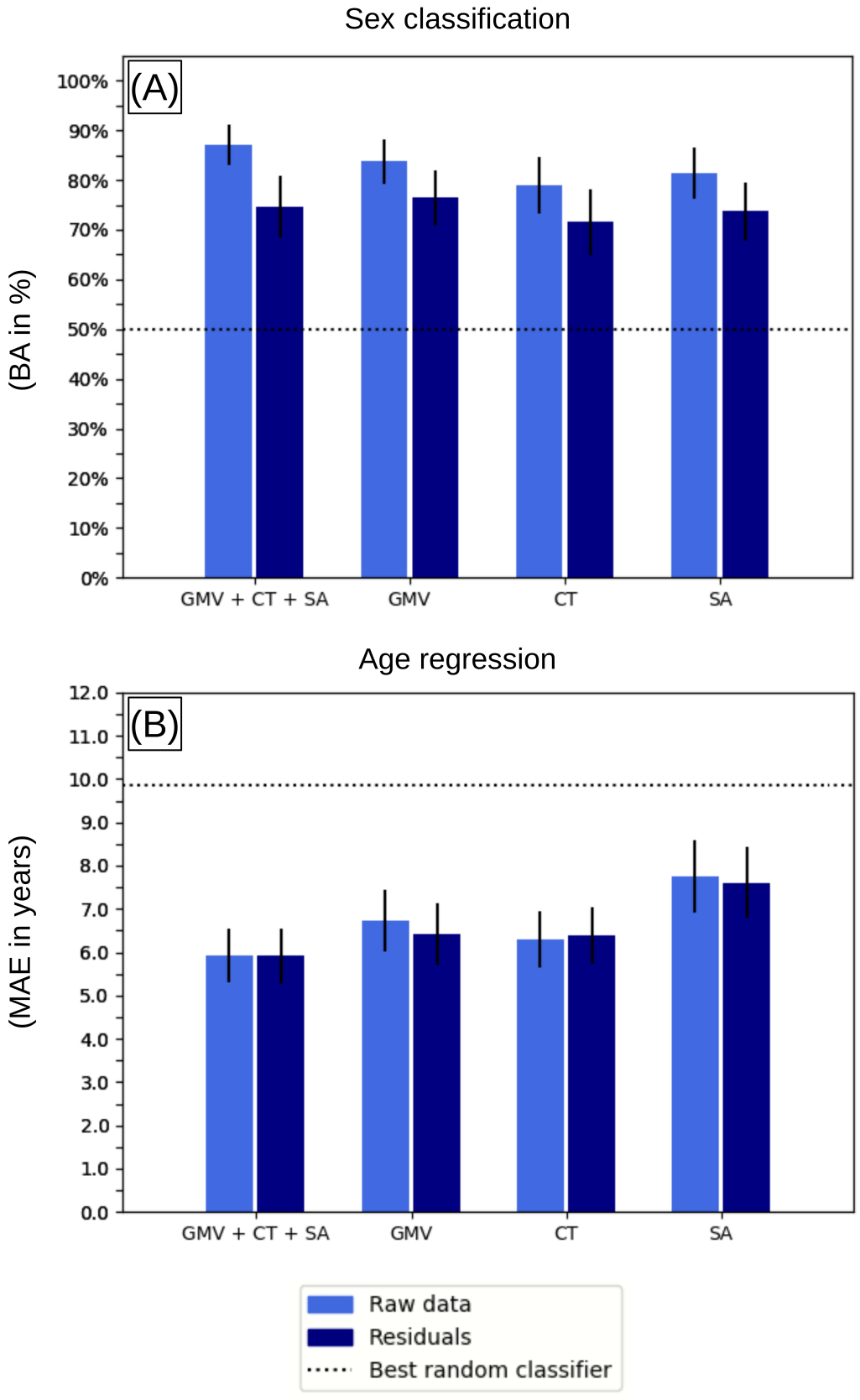

Fig. 4: AML models' internal validation performances for (A) sex classification and (B) age regression. AML performance measured as BA in %. Error bars represent standard deviation.

4.1.2 Sensitivity analysis

We assessed the performance of our AML models as a function of the input subjects, as rendered in figure 5. AML related overfitting can be analysed with 3 different measurements (see Fabris & Freitas, 2019): training-validation overfitting, training-testing overfitting and validation-testing overfitting. In this sub-section, validation refers to internal validation and testing to external validation.

The sex classification models based on raw data with SA, GMV, as well as GMV+CT+SA display similar behavior. The GMV dataset showed some training-testing and validation-testing overfitting, while the CT dataset has marked training-validation, validation-testing as well as training-testing overfitting. When based on residuals, sex classification models seemed to all display a slight training-testing and validation-testing overfit.

In the regression task, the performances displayed by the AML models on their respective modality seemed consistent on the train, internal validation and external validation datasets. The performance of our AML models in one test phase seemed to be a good estimator of its performance in other performance estimation phases, except for a marked underperformance on the residuals CT dataset and overperformance on the residuals SA dataset.

AML models designed during the benchmark tended to keep similar and convincing performances at all stages of the evaluation. This rule applied to classification as well as regression tasks. The fact AML models showed solid performances in the external validation qualifies them as a valid method for prediction on unknown subjects.

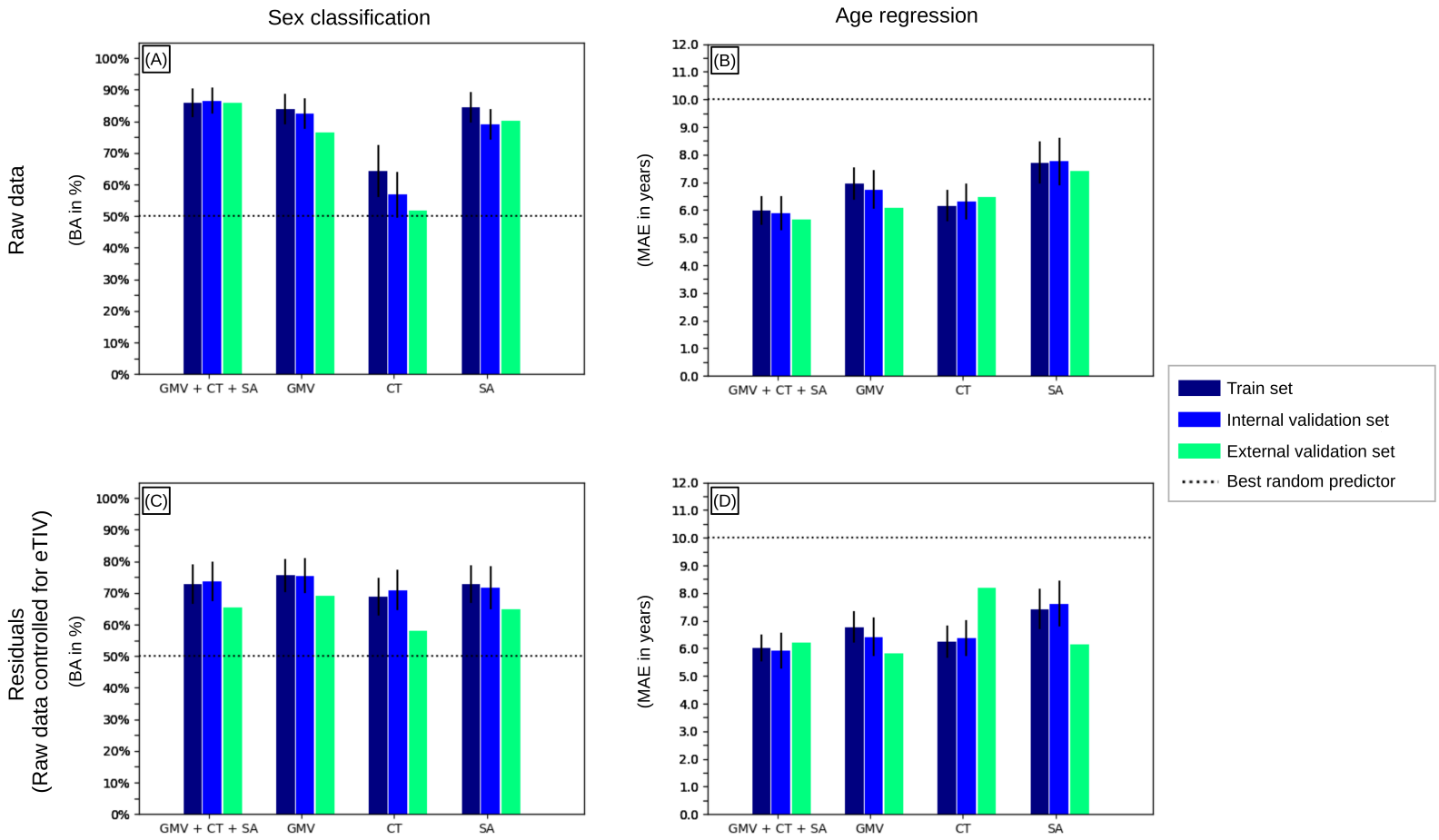

Fig. 5: Sensitivity analysis of the benchmark's AML models. AML models' evaluation performances on the train dataset (dark blue), in the internal validation (light blue) and the external validation (light green). The represented tasks are (A) sex classification with raw data, (B) age regression with raw data, (C) sex classification on residuals with eTIV regressed out, (D) age regression on residuals with eTIV regressed out. Performance is measured with MAE in years for (A) and (C), BA in % for (B) and (D). Error bars represent standard deviation. The evaluation on the train sample and test sample was performed with 10 fold cross-validation (10 repeats). The external validation was performed on the BiDirect sample set after training on the 1000BRAINS sample. Since this evaluation scheme implies measuring performance once, there are no error bars.

4.1.3 AML models' compositions

The type of model architecture chosen by AML may also be further examined and allows insight into the internal model building procedure. For each dataset, an ensemble was outputted consisting of up to 50 submodels. The composition of the ensembles as outputted by Auto-sklearn is presented in 4. The importance of a submodel's predictions in the results outputted by the ensemble is determined by the weight assigned to it. In order to reflect the impact of a certain type of pipeline in the ensemble, we calculated the relative importance of the pipeline's model type in the decisions.

For the sex classification task, Auto-sklearn built ensembles with an average size of 20.88 sub models. Auto-sklearn overwhelmingly decided to use stochastic gradient descent (SGD). It was the only type of classifier used for the raw data based tasks and had more than 50% of the weights of models basing on residuals. Only the ensemble basing on the GMV dataset with residuals did not use them. Instead of this, this ensemble used gradient boosting. Multilayer perceptrons, random forest, passive-aggressive and extra-trees classifiers all found use to a smaller extent, mostly in the ensemble of the SA dataset with residuals. The composition of age regression AML models relied on smaller ensembles, with an average ensemble size of 9.3 submodels. Those ensembles also heavily relied on SGDs, with more than half of the total weights attributed to them. multi-layer perceptron (MLP), Gaussian process (GP) and support vector regression (SVR) were also part of some ensembles.

We further analyzed the underlying ML algorithms used together with SGD pipelines, since SGD is strictly speaking an optimization heuristics that can be used on different ML architectures. They are not fully independent ML models. Scikit-learn implements SGD on regularized linear models. The loss function and other hyperparameters used define the decision mechanism of the pipeline. For the classification AML task, SGDs loss functions used were: squared hinge (30% of weights), modified huber (25%), perceptron (14%), hinge (5%) and log(25%). Using hinge based loss functions means fitting an SVM-like submodel, while a logarithmic loss function generates logistic regression like classifiers. For the regression AML tasks, SGDs based as a loss function on squared epsilon incentive for 68%, squared loss for 14%, epsilon incentive for 11% and huber for 7% of total weights.

Sex classification							
Data type	Ensemble size	Stochastic gradient descent	Random forest	Multilayer perceptron	Passive aggressive	Gradient boosting	Extra trees
Raw data							
CT	32	100%	-	-	-	-	-
GMV	15	100%	-	-	-	-	-
SA	20	100%	-	-	-	-	-
GMV+CT+SA	25	100%	-	-	-	-	-
Residuals for eTIV							
CT	14	100%	-	-	-	-	-
GMV	15	-	-	-	-	100%	-
SA	20	23%	10%	15%	-	25%	27%
GMV+CT+SA	26	100%	-	-	-	-	-

Age regression							
Data type	Ensemble size	Stochastic gradient descent	Multilayer perceptron	Gaussian process	Libsvm SVR	Liblinear SVR	-
Raw data							
CT	5	44%	-	-	56%	-	-
GMV	3	100%	-	-	-	-	-
SA	6	100%	-	-	-	-	-
GMV+CT+SA	2	50%	-	-	-	50%	-
Residuals for eTIV							
CT	6	100%	-	-	-	-	-
GMV	3	46%	-	-	54%	-	-
SA	8	14%	26%	14%	46%	-	-
GMV+CT+SA	4	38%	54%	-	-	8%	-

Table 4: Composition of Auto-sklearn ensembles - Benchmark. Composition of the AML models' ensembles for the sex classification and the age regression tasks. The ensemble size was set to a maximum of 50 pipelines for each task. For each model type, the percent represents the relative amount of weights assigned to a model type and thus how impactful the model type is in the decisions of the AML model.

Sex classification						
Raw data	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.646	0.670	0.832	0.362	0.458	0.646
GMV	0.840	0.841	0.835	0.817	0.823	0.840
SA	0.843	0.842	0.818	0.845	0.829	0.843
GMV+CT+SA	0.858	0.859	0.849	0.841	0.842	0.858
Residuals	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.689	0.693	0.672	0.640	0.652	0.689
GMV	0.755	0.758	0.741	0.726	0.729	0.755
SA	0.729	0.731	0.712	0.693	0.698	0.729
GMV+CT+SA	0.747	0.751	0.744	0.701	0.715	0.747

Age regression			
Raw data	MAE	MSE	R2 score
CT	6.152	59.719	0.623
GMV	6.955	76.097	0.526
SA	7.723	99.950	0.385
GMV+CT+SA	5.962	57.477	0.640
Residuals	MAE	MSE	R2 score
CT	6.237	61.266	0.615
GMV	6.760	72.358	0.548
SA	7.427	92.982	0.426
GMV+CT+SA	5.995	57.697	0.638

Table 5: Performance of benchmark AML models on the train dataset. Evaluation scheme: 10 fold cross-validation (10 repeats)

Sex classification						
Raw data	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.573	0.609	0.776	0.162	0.247	0.573
GMV	0.823	0.825	0.816	0.794	0.801	0.823
SA	0.792	0.793	0.775	0.766	0.767	0.792
GMV+CT+SA	0.867	0.868	0.863	0.842	0.850	0.867
Residuals	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.709	0.713	0.696	0.66	0.673	0.709
GMV	0.754	0.754	0.734	0.725	0.724	0.754
SA	0.717	0.719	0.704	0.666	0.680	0.717
GMV+CT+SA	0.749	0.752	0.753	0.687	0.712	0.749

Age regression			
Raw data	MAE	MSE	R2 score
CT	6.300	69.082	0.555
GMV	6.742	75.715	0.514
SA	7.750	96.537	0.382
GMV+CT+SA	5.901	58.724	0.621
Residuals	MAE	MSE	R2 score
CT	6.378	69.674	0.552
GMV	7.608	94.371	0.395
SA	6.403	70.248	0.548
GMV+CT+SA	5.924	58.553	0.621

Table 6: Performance of benchmark AML models on the internal validation dataset. Evaluation scheme: 10 fold cross-validation (10 repeats)

Sex classification						
Raw data	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.490	0.556	0.574	0.905	0.702	0.490
GMV	0.795	0.819	0.788	0.941	0.857	0.795
SA	0.821	0.828	0.816	0.908	0.860	0.821
GMV+CT+SA	0.859	0.864	0.877	0.890	0.884	0.859
Residuals	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.583	0.542	0.741	0.322	0.449	0.583
GMV	0.692	0.697	0.746	0.723	0.734	0.692
SA	0.653	0.669	0.700	0.750	0.724	0.653
GMV+CT+SA	0.654	0.638	0.759	0.551	0.638	0.654

Age regression			
Raw data	MAE	MSE	R2 score
CT	6.467	65.105	-0.155
GMV	6.067	58.352	-0.036
SA	7.436	87.955	-0.561
GMV+CT+SA	5.640	50.813	0.098
Residuals	MAE	MSE	R2 score
CT	6.214	59.533	-0.057
GMV	6.155	60.719	-0.078
SA	8.192	105.032	-0.864
GMV+CT+SA	5.833	53.621	0.048

Table 7: Performance of benchmark AML models on the external validation dataset. Evaluation scheme: Training on 1000BRAINS, testing on BiDirect

4.2 Second objective: Depression classification results

For the second part of our study, 31 feature sets were used in total to design 31 AML models, with each AML ensemble taking 1 day of calculations to build. The 8 starting feature sets consisted in the GMV, CT, SA and GMV+CT+SA feature sets, used with their raw value once and once with their residual values with eTIV regressed out. 14 further feature sets were formed with only the CT of the nodes of each of the 7 network according to the Schaefer parcellation (Schaefer et al., 2018), again once the raw value and once the residuals. 8 further feature sets were created, this time with anatomical priors (3.3.1). One feature set consisted only of age and sex of each sample to test the effect of confounders. It is the only feature set containing those 2 variables.

4.2.1 AML models' validation performances

Base feature sets

The 8 base AML models, based on the raw or residual data for GMV, CT, SA, and GMV+CT+SA, yielded validation performances that were at best slightly above those of a random predictor. The BA ranged between 45% and 55%. The top scoring AML model based on the raw value used the SA feature set and reached a BA of 55±6.2%. The best AML model using residuals used the GMV+CT+SA feature set, also with 55±6.2% BA. The performances of the AML models generated for our depression classification task are presented in figure 6 (for additional scoring information, see table 8 and table 9).

Anatomical priors

Restricting the number of features according to the results of a meta-analysis (Schmaal et al., 2017) did not seem to improve performances in any significant way. The validation results ranged from 51% BA (SA, residuals) to 55% BA (CT, raw data).

Network based datasets

The best validation performances were reached by separating the nodes respectively to their functional network and using their CT as features. The usage of raw data provided performances between 51% and 59% BA, outperforming the usage of residuals with those performances ranging from 50% BA to 53% BA. In this context, best AML performance was reached using raw data from nodes of the ventral attention network (BA: 59±6.0%), followed by the dorsal attention network (BA: 55±6.0%) and the control network (BA: 55±6.2%). Using residuals with networks lead at best to a 53±5.8% BA based on the dorsal attention network.

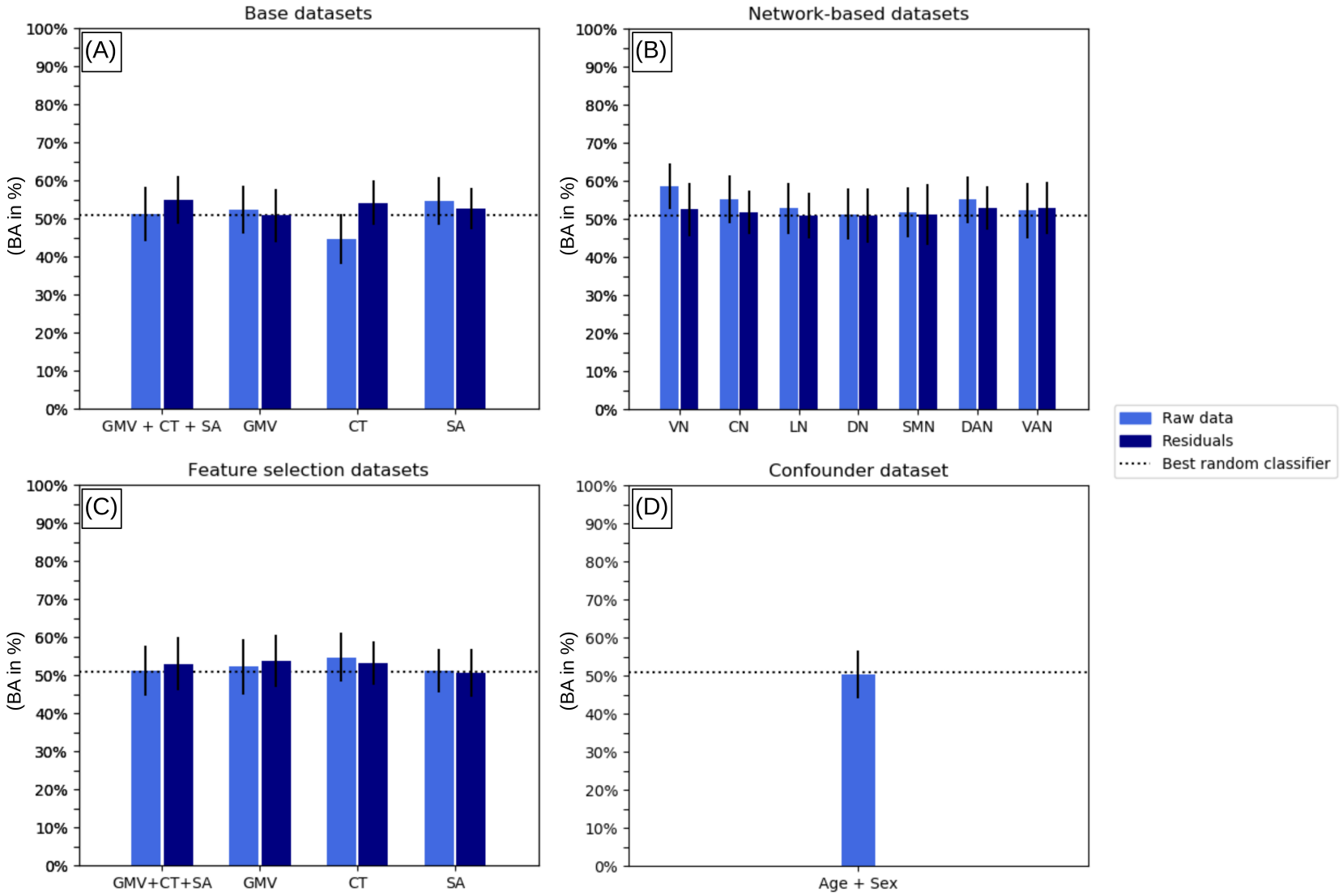

Fig. 6: Depression classification results. Depression classification results displayed across different input features/datasets. AML performance measured as BA in %. Error bars represent standard deviation. Models were built based on different feature sets: (A) data from all 400 nodes for different metrics , (B) data from 7 networks, (C) using anatomical priors for feature restriction, (D) restricting features to sex and age. VN = Visual network, CN = Control Network, LN = Limbic Network, DN = Default Network, SMN = Somatomotor Network, DAN = Dorsal Attention Network, VAN = Ventral Attention Network.

Raw data	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.465	0.467	0.466	0.470	0.464	0.465
GMV	0.538	0.535	0.539	0.512	0.518	0.538
SA	0.547	0.544	0.551	0.535	0.536	0.547
GMV+CT+SA	0.541	0.539	0.545	0.493	0.512	0.541

Residuals	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.540	0.524	0.543	0.527	0.511	0.540
GMV	0.544	0.543	0.542	0.548	0.539	0.544
SA	0.517	0.517	0.515	0.529	0.517	0.517
GMV+CT+SA	0.541	0.526	0.546	0.517	0.510	0.541

With anatomical priors - raw data	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.536	0.529	0.537	0.516	0.516	0.536
GMV	0.535	0.535	0.535	0.556	0.541	0.535
SA	0.531	0.523	0.528	0.504	0.504	0.531
GMV+CT+SA	0.533	0.533	0.533	0.521	0.522	0.533

With anatomical priors - residuals	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.562	0.561	0.561	0.562	0.557	0.562
GMV	0.557	0.557	0.556	0.559	0.554	0.557
SA	0.528	0.526	0.531	0.442	0.476	0.528
GMV+CT+SA	0.551	0.544	0.553	0.536	0.535	0.551

Network-based - raw data	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
Control Network	0.582	0.579	0.575	0.622	0.591	0.582
Default Network	0.563	0.561	0.564	0.562	0.557	0.563
Dorsal Attention	0.573	0.571	0.583	0.528	0.546	0.573
Limbic Network	0.548	0.540	0.552	0.477	0.498	0.548
Somatomotor Network	0.520	0.520	0.523	0.539	0.510	0.520
Ventral Attention Network	0.511	0.511	0.511	0.505	0.503	0.511
Visual Network	0.570	0.566	0.571	0.562	0.560	0.570

Network-based - residuals	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
Control Network	0.550	0.555	0.554	0.653	0.574	0.550
Default Network	0.524	0.522	0.522	0.529	0.520	0.524
Dorsal Attention	0.579	0.579	0.609	0.452	0.511	0.579
Limbic Network	0.553	0.547	0.549	0.605	0.561	0.553
Somatomotor Network	0.519	0.518	0.519	0.498	0.501	0.519
Ventral Attention Network	0.533	0.526	0.542	0.474	0.494	0.533
Visual Network	0.540	0.539	0.539	0.548	0.539	0.540

Table 8: Performance on train dataset of depression classifiers. Evaluation scheme: 10 fold cross-validation (10 repeats).

Raw data	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.447	0.446	0.445	0.437	0.437	0.447
GMV	0.533	0.530	0.536	0.504	0.514	0.533
SA	0.547	0.545	0.549	0.543	0.540	0.547
GMV+CT+SA	0.510	0.508	0.513	0.493	0.498	0.510

Residuals	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.544	0.527	0.549	0.555	0.529	0.544
GMV	0.511	0.511	0.510	0.500	0.500	0.511
SA	0.527	0.526	0.529	0.504	0.511	0.527
GMV+CT+SA	0.547	0.537	0.550	0.543	0.534	0.547

With anatomical priors - raw data	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.547	0.539	0.548	0.558	0.543	0.547
GMV	0.524	0.523	0.523	0.553	0.533	0.524
SA	0.515	0.508	0.518	0.530	0.512	0.515
GMV+CT+SA	0.522	0.522	0.523	0.502	0.508	0.522

With anatomical priors - residuals	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
CT	0.535	0.534	0.537	0.541	0.534	0.535
GMV	0.530	0.530	0.531	0.521	0.523	0.530
SA	0.507	0.506	0.508	0.439	0.465	0.507
GMV+CT+SA	0.525	0.519	0.526	0.524	0.516	0.525

Network-based - raw data	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
Control Network	0.552	0.546	0.553	0.572	0.552	0.552
Default Network	0.513	0.511	0.513	0.538	0.52	0.513
DorsalAttention	0.551	0.549	0.556	0.525	0.531	0.551
Limbic Network	0.528	0.516	0.53	0.528	0.511	0.528
Somatomotor Network	0.518	0.518	0.521	0.508	0.503	0.518
Ventral Attention Network	0.531	0.531	0.528	0.554	0.535	0.531
Visual Network	0.593	0.590	0.598	0.570	0.578	0.593

Network-based - residuals	Balanced accuracy	Accuracy	Precision	Recall	F1 score	ROC AUC
Control Network	0.518	0.513	0.513	0.706	0.572	0.518
Default Network	0.506	0.505	0.506	0.512	0.504	0.506
DorsalAttention	0.53	0.527	0.538	0.448	0.479	0.53
Limbic Network	0.51	0.505	0.506	0.651	0.557	0.51
Somatomotor Network	0.512	0.509	0.514	0.498	0.491	0.512
Ventral Attention Network	0.534	0.528	0.537	0.535	0.527	0.534
Visual Network	0.526	0.526	0.528	0.524	0.521	0.526

Table 9: Performance on validation dataset of depression classifiers. Evaluation scheme: 10 fold cross-validation (10 repeats).

Auto-sklearn vs TPOT

Our choice of using Auto-sklearn as our main AML library was driven by previous benchmarks and reports on estimated performances (3.2.3.1). We performed a retest with TPOT on the base feature sets for depression classification to evaluate whether this would result in significant discrepancies in performances.The results are presented in table 10.

Overall, when considering the mean BA of the 8 base models averaged for TPOT and Auto-sklearn, TPOT outperformed Auto-sklearn by a very thin margin (0.8%). The result of this retest in concordant with AML benchmarks reporting no clear superiority of one AML library over others across all datasets (Gijsbers et al., 2019; Zöller & Huber, 2021).

	Auto-sklearn Raw data	TPOT Raw data	Auto-sklearn residuals (eTIV)	TPOT residuals (eTIV)
CT	44.7±6%	54.7±6%	54.2±5%	50.6±7%
GMV	52.3±6%	50.7±6%	50.9±7%	53.7±6%
SA	54.6±6%	49.2±6%	52.7±5%	53.8±7%
GMV+CT+SA	51.3±7%	51.9±5%	54.5±6%	53.9±6%
Age + Sex	48.2±6%	49.6±3%	-	-

Table 10: Autosklearn vs. TPOT based models for depression classification. Comparison of BA in 10 fold cross-validation (10 repeats). Values represent the BA in %. The best values for each dataset are marked in bold font.

Effects of matching on performances

The effects of age and sex on structural neuroimaging derived data are multifaceted and complex (Jockwitz et al., 2021; Ruigrok et al., 2014, *review*). We used propensity score matching on the HC / MDD feature with age and sex while selecting samples (3.3.1). By doing this, we had the intention of negating the effect of those two confounders on prediction performances.

To verify the efficiency of the method for our problem, we ran a separate depression prediction task on the BiDirect sample, using solely age and sex as features. The main difference with our already presented age + sex (matched) dataset is that in this dataset (age + sex (no match)), the samples were not matched using propensity score matching. The results of this run are

presented in figure 7. With a BA of 60±5.1% in the absence of a confounder targeted matching strategy, the age + sex based classification of HC vs depressive subjects outperforms our results from the main analysis. With matching, the performance of classification based on those two features dropped to 48±6.0% BA.

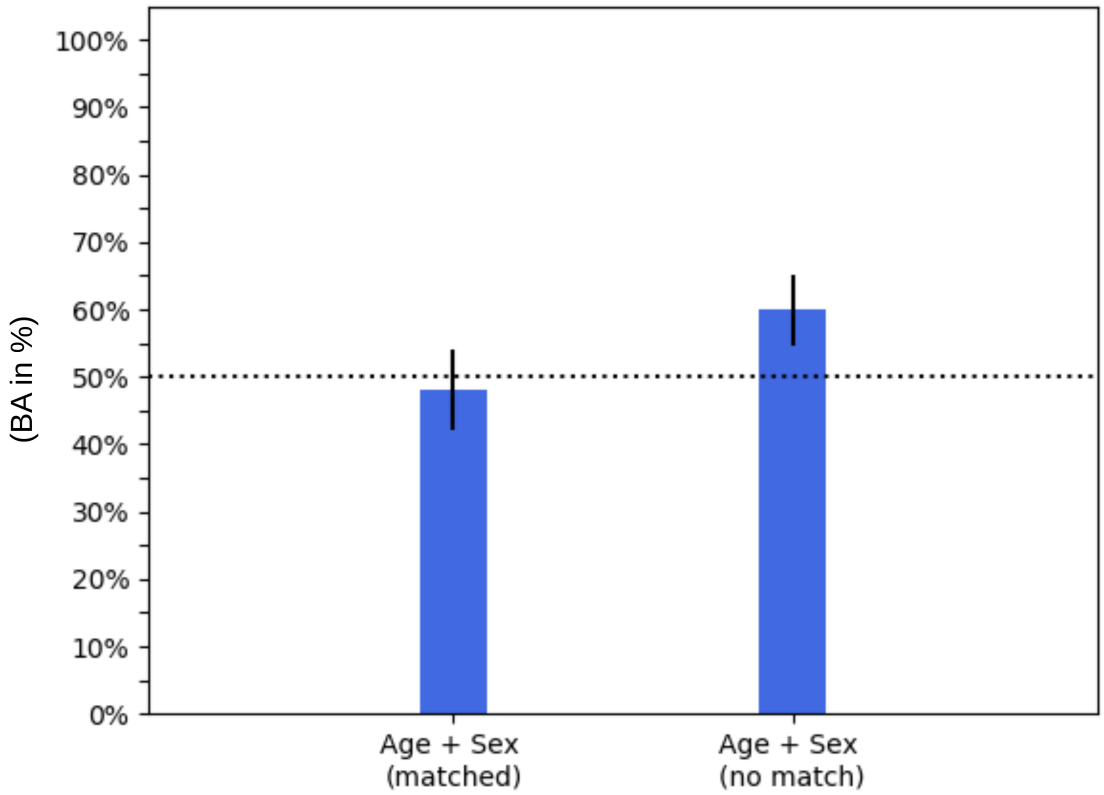

Fig. 7: Effect of confounders matching on depression classification

4.2.2 AML models' compositions

The model weights were calculated in the same manner as in 4.1.3, and are presented in table 11 and table 12.

For the AML task basing on the 8 base datasets, Auto-sklearn built ensembles with an average size of 22.4 submodels. Auto-sklearn split the submodels used in the ensembles more evenly than in the sex classification task. The most commonly used type of ML model architecture was random forest classifiers (30.25% of weights), followed by gradient boosting classifiers

(25.5%), then SGD (19.75%), extra-trees classifiers (19.5%), passive aggressive classifiers (4.1%) and finally MLP (0.75%).

With regard to the network based AML models, average ensemble size reached 18.6 submodels in average. The composition of the ensembles differs from the previously described ones. MLP was the most important ML model type, with 40% of weights, followed by gradient boosting (26%), random forest (13%), extra trees (13%) and SGD (2%). The AML models created with feature restriction based on anatomical priors had an average ensemble size of 19.8 and had similar compositions, basing on MLP, gradient boosting, random forest as well as extra trees.

Depression Diagnosis							
Data type	Ensemble size	Stochastic gradient descent	Random forest	Multilayer perceptron	Passive aggressive	Gradient boosting	Extra trees
Raw data							
CT	27	-	100%	-	-	-	-
GMV	8	48%	-	-	-	-	52%
SA	12	-	-	-	-	100%	-
GMV+CT+SA	32	100%	-	-	-	-	-
Residuals for eTIV							
CT	21	10%	52%	-	33%	4%	-
GMV	31	-	90%	6%	-	-	4%
SA	16	-	-	-	-	100%	-
GMV+CT+SA	32	-	-	-	-	-	100%

Depression Diagnosis – Network based							
Data type	Ensemble size	Stochastic gradient descent	Random forest	Multilayer perceptron	Passive aggressive	Gradient boosting	Extra trees
Raw data							
CN	16	-	-	100%	-	-	-
DAN	7	-	81%	19%	-	-	-
DN	22	-	-	100%	-	-	-
LN	13	-	-	31%	-	-	69%
SMN	18	24%	-	76%	-	-	-
VAN	21	-	-	-	-	100%	-
VN	4	-	-	-	-	-	100%
Residuals for eTIV							
CN	18	-	-	36%	-	64%	-
DAN	22	-	-	100%	-	-	-
DN	12	-	100%	-	-	-	-
LN	28	-	-	100%	-	-	-
SMN	22	-	-	91%	-	-	9%
VAN	32	-	-	-	-	100%	-
VN	35	-	-	-	-	100%	-

Confounders							
Age + Sex	33	-	96%	4%	-	-	-

Table 11: Composition of Auto-sklearn ensembles - MDD/HC Classification. Composition of the AML models' ensembles for the depression classification task. The ensemble size was set to a maximum of 50 pipelines for each task. For each model type, the percent represents the relative amount of weights assigned to a model type and thus how impactful the model type is in the decisions of the AML model.
CN = Control Network, DAN = Dorsal Attention Network, DN = Default Network, LN = Limbic Network, SMN = Somatomotor Network, VAN = Ventral Attention Network, VN = Visual network

Depression Diagnosis - Feature restriction with anatomical priors					
Data type	Ensemble size	Random forest	Multilayer perceptron	Gradient boosting	Extra trees
Raw data					
CT	6	7%	-	23 %	70 %
GMV	24	-	71 %	-	29 %
SA	35	-	-	100 %	-
GMV + CT +SA	1	-	-	-	100 %
Residuals for eTIV					
CT	7	-	-	100 %	-
GMV	44	-	-	-	100 %
SA	32	-	100 %	-	-
GMV + CT +SA	9	-	-	-	100 %

Table 12: Composition of Auto-sklearn ensembles - MDD/HC Classification using feature restriction with anatomical priors Composition of the AML models' ensembles for the depression classification task using feature restriction with anatomical priors. The ensemble size was set to a maximum of 50 pipelines for each task. For each model type, the percent represents the relative amount of weights assigned to a model type and thus how impactful the model type is in the decisions of the AML model.

CHAPTER 5

Discussion

The global aim of this study was to determine whether AML could be used for ML prediction using sMRI derived features by testing it with an AML benchmark, then to use it as an alternative to classical ML approaches for depression classification. The results of the first part of the study showed that AML based classification was a viable option for sMRI derived data, and an alternative worth considering for depression classification. The usage of AML lead to well-generalizing AML models for single-subject sex classification as shown in the external validation on the BiDirect dataset (Teismann et al., 2014). In the regression task, results close to those reported in the literature were achieved, qualifying AML for the task of rapidly providing efficient predictive models (Jiang et al., 2020; Cole et al., 2018; Cole et al., 2017; Dafflon et al., 2020). The ensembles designed by the AML libraries used in this work were different and more complex than those described in the literature, thus differentiating AML models from ML models designed with classical ML methods. The depression classification task resulted in low BA when attempting the task using AML on sMRI derived features. This finding supports those of recent studies attempting depression classification based on sMRI data on datasets with large sample sizes (Stolicyn et al., 2020; Flint et al., 2021). Thus, results from the current dissertation support the general applicability of AML for neuroimaging questions and highlight the limited performances of depression classification based on sMRI information.

5.1 First objective: AML Benchmark

Determining whether the employment of AML is pertinent for prediction based on sMRI derived data was the first step of this work, an important prerequisite before attempting depression classification. Only limited prior literature is available on the usage of AML for structural imaging data. In this context, prior investigations have focused on the application of AML to particular tasks, e.g. age prediction (Dafflon et al., 2020) or resectability of meningeoma (Musigmann et al., 2022). The current thesis extended previous work by providing a more detailed insight into the applicability to a classification task as well as a regression task on two large cohorts, with one being used solely for external validation ($N_{BiDirect}$=1102; Teismann et al., 2014).

This work first showed that AML was able to outperform random sex classification based on structural neuroimaging data. The benchmark results also made this apparent for regression, thereby corroborating the findings from Dafflon et al. (2020). The baseline objective of generating predictions more precise than a dummy predictive algorithm was completed both for the classification and the regression tasks. The top sex classifier reached 87% BA, surpassing the 50% BA of the random baseline classifier. The top age regressor reached 5.9 years MAE, also clearly besting the 9.73 years MAE of the random baseline regressor.

Outmatching random prediction was a first important step, but to represent a viable alternative to ML for depression classification, AML performing at least similarly to it was a significant milestone. The key results of the benchmark are here compared with the state-of-the-art performances for sex classification and age regression.

5.1.1 Sex classification

Overall, the results in sex classification reached by this study's top AML model were on the level of state-of-the-art ML models. In Nieuwenhuis et al. (2017), the average accuracy reached ranged from 81% to 94% on the samples of multiple studies (N_{total} = 389) using SVM with nested cross-validation (Cahn et al., 2002; Mourao-Miranda et al., 2012; Schaufelberger

et al., 2007; Crespo-Facorro et al., 2009; Velakoulis et al., 2006). In Anderson et al. (2019), an accuracy of 93.8% on the validation hold out set was obtained ($N_{validation}$ = 370) using radial basis function (RBF) SVM and logistic regression as ML model types, also fitting in this range. In Joel et al., 2018, 3 datasets from different studies were used for external validation with results ranging from 71% to 86% accuracy with SVMs. With 87% BA in the internal validation and 86% BA in the external validation, the AML sex classifier using all feature types as raw data scored on a level similar to these results. This study thus highlighted the efficiency of AML for sex classification based on morphological brain features. It also supports the general applicability of AML to neuroimaging classification tasks (Musigmann et al., 2022).

This study also showed that AML library provided an efficient way of selecting relevant ML model types for the task of sex classification, while still designing original pipeline structures. Without prior knowledge on optimal model type for the problem, Auto-sklearn 2.0 selected ML model types known in the literature (Feurer et al., 2021). Sex classification studies relied on logistic regression (Chekroud et al., 2016; Anderson et al., 2019) and SVMs (Anderson et al., 2019). The AML ensembles showed a more diverse composition (4.1.3), but the choice in the ensembles' SGDs of squared-hinge and hinge loss functions (making the models similar to SVMs) and logistic regression was predominant. AML, more specifically Auto-sklearn, came to similar conclusion regarding optimal model-choice than three other studies (Chekroud et al., 2016; Anderson et al., 2019; Joel et al., 2018). Those results indicate that AML is an efficient way of discovering efficient ML model types for neuroimaging classification problems.

5.1.2 Age regression

The performances of this work's AML age regressors were close to those known in the current literature. Previous studies, all covering the same large age range (18-90 years), have reported MAEs ranging from 4.16 to 5.55 years based on structural features (Jiang et al., 2020; Cole et al., 2018; Cole et al., 2017). A further example of age regression ML task on another sample yielded similar results (N = 2640, MAE on test set: 4.54 ± 0.06 - 5.82 ± 0.09 years,

female ratio = 53.1%, mean age = 35.87 ± 16.20 years, age range : 17-90 years; Treder et al., 2021). This study's AML models basing on raw data with GMV + CT + SA performed closely to those models with 5.90 years MAE in the internal validation evaluation (age range: 18-85 years).

Dafflon et al. (2020) also tested the efficiency of AML models for regression on age prediction with a large sample (N = 10307, mean age = 59.40 years, age range : 18-89 years). TPOT was the AML library used and reached a performance of 4.6 years MAE, performing better than the best age-regressor in the benchmark (5.90 years MAE). An explanation for this discordance in results could reside in the choice of atlas. The Desikan-Killiany atlas (Desikan et al., 2006) and ASEG Freesurfer atlas (Fischl et al., 2002) were used in Dafflon et al. (2020), contrasting with the choice made in this work to use an atlas designed based on the characteristics of rs-fMRI (Schaefer et al., 2018). The use of a structural atlas, such as the Juelich brain atlas, may yield improved performances (Amunts et al., 2020). Further explanations may reside in the normal variance of performances of ML regressors, population-specific attributes and the choice of AML library (Dafflon et al., 2020, Balaji & Allen, 2018).

The AML based approach partly selected submodels types known in the literature, while still combining them in original ML model architectures. Diverse ML model types are described in the literature for age regression. Convolutional neural networks, relevance vector regression / machine, Gaussian process regression , support vector regression / machine and random forest are examples of models that were used successfully in the past (Jiang et al., 2020, table 5). The structure of the AML models outputted by Auto-sklearn, combining multiple approaches into complex ensembles, seemed to be relatively unique (Feurer et al., 2019). Some of the submodels' types used (principally SGD, then MLP, SVR and finally Gaussian process) were already known for age regression (Jiang et al., 2020, table 5). In this regard, AML offered an effective way to rapidly sort out amidst a vast range of possibilities which ML pipeline type to favor for the task. This attribute is especially valuable for future research on AML tasks with targets that received little or no attention yet as a first approach.

5.1.3 Implications of the AML benchmark

Positive results for the classification tasks were an important requirement to proceed to the second objective. Results from the AML benchmark have demonstrated that comparable results to the literature can be achieved for two problems of the combined fields of neuroimagining and ML. This works supports other pilot studies stating that using AML's fully automatized pipeline design procedures is a solid method for ML tasks in neuroimaging (Dafflon et al., 2020; Musigmann et al., 2022). Both parts of the AML benchmark revealed results that corroborated Dafflon et al. (2020) findings, confirming the utility of AML as a practical first-approach tool for new ML tasks in the field.

This study moreover demonstrated the convincing generalization power of AML models when predicting with unknown subjects. Their performances were consistent in the external validation for two structural neuroimaging tasks on the large sample of the BiDirect study, external validation being the gold standard for performance assessment (Cabitza et al., 2021; Bleeker et al., 2003, Teismann et al., 2014). The top sex classifier of the AML benchmark went from an 87% BA in the internal validation to an 86% BA in the external validation, while the top age regressor went from 5.90 years MAE to 5.64 years MAE. Well exporting ML models for age and sex prediction were recently achieved for those two tasks (Flint et al., 2020; Baecker et al., 2021). AML displayed the potential to predict accurately in unknown ground truth scenarios in two sMRI based tasks, similarly to ML.

5.2 Second objective: Depression classification

The second aim of this study was to determine whether AML models could be used in order to accurately differentiate HC from depressive subjects based on sMRI derived data. The results obtained on this AML task showed performances slightly above those of a random predictor, with a 55 ± 6% BA for the top base classifier (GMV + CT + SA with residuals) and a 59 ± 6% BA for the top depression classifier overall (using the CT of the ventral attention network with raw data).

5.2.1 Influence of sample size and confounders on performances

In contrast, numerous small-sample studies reported high accuracies for depression classification on sMRI (Patel et al., 2016, *review*; Gao et al., 2018, *review*). Patel et al. (2016) for example has reviewed 15 studies on depression classification with accuracies ranging between 67.6% - 94.3%. The evaluation methodology in this work greatly differs from those previous reports in multiple regards. These studies used small samples (none with more than 100 depressed patients in Patel et al. (2016) for example) for their validation set. As a mean of comparison, Flint et al. (2021) recommended having at least 150 samples for a validation set in depression classification. Moreover, the validation schemes of those studies consisted overwhelmingly in using LOOCV or cross-validation on the dataset already used for ML model design (Patel et al., 2016, *review*; Gao et al., 2018, *review*). Such a protocol was long thought a safe method to avoid overoptimistic performance reports, but more recently proved insufficient in neuroscience (Hosseini et al., 2020) as well in predictive medical ML tasks (Navarro et al., 2021, *review*). Training and testing on the same dataset drastically increases the risk of overoptimistic results (Hosseini et al., 2020; Navarro et al., 2021, *review*). The use of LOOCV comes with an increased risk of unstable and biased estimates (Flint et al., 2021; Varoquaux et al., 2017). The combination of small sample size and lenient validation scheme opens a path for sporadic positive findings (Flint et al., 2021).

This work's AML models' design and validation scheme followed a comparatively stricter procedure, which is the probable explanation for the discrepancies in performance evaluation. The AML model design, including hyperparameter optimization, was performed exclusively on the train dataset. The models were not subjected to change in their architecture after the initial design phase. The validation sample of 551 subjects stood well-above the suggested threshold of 150 samples for depression classification (Flint et al., 2021). It was used for a 10 fold cross-validation (10 repeats) after AML model design. The use of this stricter validation protocol results in validation performances that can be seen as closer to the real-world performances of the AML models.

The notable lack of negative results in two literature reviews may be revealing of a publication bias in the field of depression classification based on sMRI (Patel et al., 2016, *review*; Gao

et al., 2018, *review*). Such a bias may result in the overly frequent reporting of positive or significant results and dismissal of negative results, caused both on researcher level and on publisher level (Dickersin & Min, 1993; Rothstein et al., 2005; Easterbrook et al., 1991). With negative results being tendentially discarded and validation schemes allowing for sporadic positive results, this would result in literature reports giving an overly positive estimation on the feasibility of the task.

The depression classification results in this work however converge with the findings of depression classification studies with larger samples as well as strict train/validation split procedures (Stolicyn et al., 2020; Flint et al., 2021). A study by Stolicyn et al. (2020) has shown a similar discordance between performances on the train dataset and performance on the external validation dataset. The best ML models on the train set reached 75% accuracy using classical ML methods (decision tree (DT)) based on data related to brain morphometry (Stolicyn et al., 2020). The accuracy of the ML model on an external validation dataset derived from the UK Biobank sample with self-reported depression as a target yet was 53.63% using brain morphometric features (Sudlow et al., 2015; Stolicyn et al., 2020). The results of this study are also in line with those presented in Flint et al. (2021) on depression classification. There, training and testing an ML classifier on a large and balanced dataset (N = 1868) between HC and depressive patients yielded an accuracy of 61% BA with SVMs (Flint et al., 2021). The AML depression classifiers designed in this work performed similarly, corroborating these results. Repeated negative results for the classification of depression with large datasets based on structural brain features may hint at the absence of relevant information in sMRI for this task (Schulz et al., 2022).

Both Flint et al. (2021) and Stolicyn et al. (2020) reviewed the influence of low sample sizes on prediction accuracy. They showed that the chances of artificially good performances increase continuously with decreased sample size (Flint et al., 2021; Stolicyn et al., 2020). This effect is partly independent of the nature of the information withheld in the features. It could be reproduced with a dummy classifier predicting randomly based solely on the prevalence of classes in a train sample (Flint et al., 2021). The BA of SVMs trained to test the impact of test sample size for depression classification with sMRI data averaged around 60%

BA, with a minimum of 35%BA and a maximum of 81% BA for N=100 using LOOCV (Flint et al., 2021). This elevated variability in performance estimates may result in misestimation when using low evaluation sample sizes. This problem is common in neuroimaging ML tasks and recently gained significant attention (Arbabshirani et al., 2017; Varoquaux, 2018). It may be part of the explanation for the numerous positive findings in depression classification in small sample studies and repeated negative results in studies with larger samples.

As an additional finding, this study furthermore showed that the effects of age and sex on structural data could inadvertently lead to reporting artificially high performances for the present AML task. Random-sampling alone did not seem to sufficiently negate the influence of these confounders on the target (4.2.1). The sole use of random sampling resulted in BA of 60% for depression prediction when using sex and age as features. Utilizing propensity score matching as a further preprocessing step between confounders and target reduced the BA to 48% with the same features. Future research would profit from baseline predictors based solely on demographics as a way to evaluate confounders effect on depression prediction.

5.2.2 Methodological considerations, limitations and future outlook

This work's results revealed that the employment of AML yielded performances on the level of state-of-the-art, conventional ML models tested in large-scale studies for depression classification using sMRI data (Flint et al., 2021; Stolicyn et al., 2020). It hence showed a limited potential for this task when evaluated on a large sample size, displaying results at best slightly above a random classifier (Stolicyn et al., 2020; Flint et al., 2021). This is not caused by an inability of AML to perform classification using neuroimaging data, as shown with the sex classification task, but inherent to the task of depression prediction. In the apparent dichotomy between small samples/high accuracy and high sample/low accuracy studies existing in the literature, present findings support the idea that high accuracy single subject AML classification for depression on unknown subject has yet to be achieved. Successive absence of positive results for the task do not prove the absence of feasibility. They may either

be the sign of a necessary evolution of methodology for depression classification or a hint of the impossibility of the task due to biological reasons.

The method in this work involved a rigid preprocessing of the sMRI which might be the reason for the negative results. It based on the use of the 400-node Schaefer parcellation (Schaefer et al., 2018). Using an rs-fMRI parcellation may, in the event of a positive result, have allowed to establish a connection between structural and functional changes in depressive disorder (Dai et al., 2019, *review*). An other large dataset study using the Desikan-Killiany atlas did not achieve better results through this process (Stolicyn et al., 2020; Desikan et al., 2006). The use of further parcellations may nonetheless prove more adapted to the problem.

The concept of parcellation itself implies a reduction of the initial information per subjects (Eickhoff et al., 2018). This allows the favoring of certain types of information and offers a practical way to orient the decision making of ML processes while avoiding the 'curse of dimensionality' (Eickhoff et al., 2018; Bellman, 1966). It however comes at the cost of discarding potentially useful information. Renouncing the parcellation of the data as was done in Flint et al. (2021) may allow ML models to combine different type of information for decision making in the future. Increasing the number of features per subjects in order to improve performances would nevertheless require increasing drastically the size of the datasets used (Jain & Waller, 1978).

Independently of the preprocessing methods used, whether sMRI alone withholds the necessary information for accurate single subject prediction of depressive disorder is unclear (Schulz et al., 2022). Multiple studies revealed group-wise structural brain differences between depressive patients and HC (Trifu et al., 2020; Zhang et al., 2018; Schmaal et al., 2017). Statistically significant group-wise differences in neuroanatomical features however do not necessarily imply effects strong enough on the individual level for single subject ML depression prediction (Schulz et al., 2022). The areas determined as significantly different in depression according to Schmaal et al. (2017) were used for feature restriction as described in section 3.3.1. This restriction of features according to anatomical priors yielded no improvements in performances. This result reinforces the idea of insufficient effect of depression on the individual level for single-subject depression prediction (Schulz et al., 2022).

The task of imaging based depression prediction could profit from the use of different MRI modalities. The possibility offered by fMRI to capture the patterns of function of the brain as well as of diffusion-weighted magnetic resonance imaging (DW-MRI) to hint at anomalies in its connectivity could very well improve classifier performances to clinically relevant levels. Studies investigating the possibility of combining modalities for diagnostic of psychiatric disease exist, with promising results (Gao et al., 2018, *review*; Arbabshirani & Calhoun, 2011).

Future research on the topic of depression classification would profit from larger samples (Marek et al., 2022; Arbabshirani et al., 2017; Iniesta et al., 2016; Varoquaux, 2018; Button et al., 2013, *review*). This would allow for better inference on the generalization power of designed ML models and improved performances (Button et al., 2013, *review*). It would additionally permit the usage of specific type of ML models, whose training require large amount of data (e.g. Neural networks, complex ensembles; Arbabshirani et al., 2017). Those model types allowed for breakthrough in other fields but are heavily reliant on vast amounts of subjects for training (Jiřík et al., 2022). Moreover, larger samples allow for the possibility to match for increasing amount of confounders while keeping a healthy dataset size. The field of neuroscience thankfully already is adapting to the necessity for larger data repositories. The last years saw the establishment of different central repositories for raw data (Turner, 2014). The assembly of such a repository for depressive disorder would be of tremendous value.

Whether sMRI alone withhold the necessary information for accurate single subject prediction of depressive disorder is an open question (Schulz et al., 2022). The task of imaging based depression prediction could profit from the use of different MRI modalities. The possibility offered by fMRI to capture the patterns of function of the brain as well as of DW-MRI to hint at anomalies in its connectivity could very well improve classifier performances to clinically relevant levels (Gao et al., 2018). Studies investigating the possibility of combining modalities for diagnostic of psychiatric disease exist, with promising results (Gao et al., 2018, *review*; Arbabshirani & Calhoun, 2011).

5.3 Conclusion

The aim of this work was to first investigate the applicability of AML to sMRI-derived features, then employ it for HC vs depressive subjects classification. The first part of the study evaluated the efficiency of AML in a common classification (sex) and regression (age) setup. The second part of the study consisted in the creation and evaluation of AML models designed for depression classification based on the same features as in the first part of the study. The effect of feature restriction according to anatomical priors and functional networks on prediction was additionally tested.

AML was shown to be similar in performances to ML models described in the literature, both for the sex classification and the age regression tasks when evaluated on the large sample of the 1000BRAINS study, supporting previous work on this matter (Dafflon et al., 2020; Musigmann et al., 2022). The AML models outputted by Auto-sklearn displayed convincing performances in the external validation, generalizing well to the BiDirect cohort. This speaks for a satisfying generalization power, qualifying AML as a suitable option for hypothesis testing and optimal model search in further neuroimaging research. The convincing performances and generalizability of AML confer to the method a position of choice for approaching new neuroimagining problems in a time-efficient manner while developing predictive algorithms of high quality. Its main downside reside in the complexity of the generated models, leading to restricted model explainability (Vilone & Longo, 2021).

Regarding depression classification, AML models performed slightly above random at best, similarly to ML models that were trained on large samples in other studies. These results hint at an inability from AML and ML models to find a general abstract rule allowing to differentiate depressive patients from HC using morphological brain data. Multiple large sample studies converging toward this null result make the hypothesis of insufficient information about the target in sMRI data realistic. Further research could profit from improvements in resolution yielding new types of data, combination of multiple imaging modalities and increases in the size of available datasets. With standard ML methods as well as with AML methods, reliably differentiating HC from depressive patients based on sMRI data does not currently seem in reach.

Bibliography

Alberg, A. J., Park, J. W., Hager, B. W., Brock, M. V., & Diener-West, M. (2004). The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *Journal of General Internal Medicine*, *19*(5 Pt 1), 460–465.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.).

Amunts, K., Mohlberg, H., Bludau, S., & Zilles, K. (2020). Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture. *Science (New York, N.Y.)*, *369*(6506), 988–992.

Anderson, N. E., Harenski, K. A., Harenski, C. L., Koenigs, M. R., Decety, J., Calhoun, V. D., & Kiehl, K. A. (2019). Machine learning of brain gray matter differentiates sex in a large forensic sample. *Human Brain Mapping*, *40*(5), 1496–1506.

Arbabshirani, M. R., & Calhoun, V. D. (2011). Functional network connectivity during rest and task: Comparison of healthy controls and schizophrenic patients. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2011*, 4418–4421.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*, 137–165.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, *46*(3), 399–424.

Baecker, L., Dafflon, J., da Costa, P. F., Garcia-Dias, R., Vieira, S., Scarpazza, C., Calhoun, V. D., Sato, J. R., Mechelli, A., & Pinaya, W. H. L. (2021). Brain age

prediction: A comparison between machine learning models using region- and voxel-based morphometric data. *Human Brain Mapping*, *42*(8), 2332–2346.

Balaji, A., & Allen, A. (2018). Benchmarking Automatic Machine Learning Frameworks. *arXiv:1808.06492*.

Barson, J. R., Mack, N. R., & Gao, W.-J. (2020). The Paraventricular Nucleus of the Thalamus Is an Important Node in the Emotional Processing Network. *Frontiers in Behavioral Neuroscience*, *14*.

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Berlim, M. T., Eynde, F. v. d., Tovar-Perdomo, S., & Daskalakis, Z. J. (2014). Response, remission and drop-out rates following high-frequency repetitive transcranial magnetic stimulation (rTMS) for treating major depression: A systematic review and meta-analysis of randomized, double-blind and sham-controlled trials. *Psychological Medicine*, *44*(2), 225–239.

Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R. T., Derksen-Lubsen, G., Grobbee, D. E., & Moons, K. G. M. (2003). External validation is necessary in prediction research: A clinical example. *Journal of Clinical Epidemiology*, *56*(9), 826–832.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88.

Boudes, E., Gilbert, G., Leppert, I. R., Tan, X., Pike, G. B., Saint-Martin, C., & Wintermark, P. (2014). Measurement of brain perfusion in newborns: Pulsed arterial spin labeling (PASL) versus pseudo-continuous arterial spin labeling (pCASL). *NeuroImage. Clinical*, *6*, 126–133.

Brådvik, L. (2018). Suicide Risk and Mental Disorders. *International Journal of Environmental Research and Public Health*, *15*(9), 2028.

Bretschneider, J., Kuhnert, R., & Hapke, U. (2017). Depressive symptoms among adults in germany. In *Journal of health monitoring*. Robert Koch-Institut, Epidemiologie und Gesundheitsberichterstattung.

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The Human Hippocampus and Spatial and Episodic Memory. *Neuron*, *35*(4), 625–641.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Cabitza, F., Campagner, A., Soares, F., García de Guadiana-Romualdo, L., Challa, F., Sulejmani, A., Seghezzi, M., & Carobene, A. (2021). The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, *208*.

Cahn, W., Pol, H. E. H., Lems, E. B., van Haren, N. E., Schnack, H. G., van der Linden, J. A., Schothorst, P. F., van Engeland, H., & Kahn, R. S. (2002). Brain volume changes in first-episode schizophrenia: A 1-year follow-up study. *Archives of general psychiatry*, *59*(11), 1002–1010.

Carroll, B. J., Cassidy, F., Naftolowitz, D., Tatham, N. E., Wilson, W. H., Iranmanesh, A., Liu, P. Y., & Veldhuis, J. D. (2007). Pathophysiology of hypercortisolism in depression. *Acta Psychiatrica Scandinavica. Supplementum*, (433), 90–103.

Caspers, S., Moebus, S., Lux, S., Pundt, N., Schütz, H., Mühleisen, T. W., Gras, V., Eickhoff, S. B., Romanzetti, S., Stöcker, T., Stirnberg, R., Kirlangic, M. E., Minnerop, M., Pieperhoff, P., Mödder, U., Das, S., Evans, A. C., Jöckel, K.-H., Erbel, R., . . . Amunts, K. (2014). Studying variability in human brain aging in a population-based German cohort—rationale and design of 1000BRAINS. *Frontiers in Aging Neuroscience*, *6*.

Castanheira, L., Silva, C., Cheniaux, E., & Telles-Correia, D. (2019). Neuroimaging Correlates of Depression—Implications to Clinical Practice. *Frontiers in Psychiatry*, *10*.

Chekroud, A. M., Ward, E. J., Rosenberg, M. D., & Holmes, A. J. (2016). Patterns in the human brain mosaic discriminate males from females. *Proceedings of the National Academy of Sciences*, *113*(14).

Cole, J. H., Ritchie, S. J., Bastin, M. E., Valdés Hernández, M. C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S. E., Zhang, Q., Wray, N. R., Redmond, P., Marioni, R. E., Starr, J. M., Cox, S. R., Wardlaw, J. M., Sharp, D. J., & Deary, I. J. (2018). Brain age predicts mortality. *Molecular Psychiatry*, *23*(5), 1385–1392.

Cole, J. H., Poudel, R. P. K., Tsagkrasoulis, D., Caan, M. W. A., Steves, C., Spector, T. D., & Montana, G. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, *163*, 115–124.

Coulon, P., Budde, T., & Pape, H.-C. (2012). The sleep relay—the role of the thalamus in central and decentral sleep regulation. *Pflügers Archiv - European Journal of Physiology*, *463*(1), 53–71.

Crespo-Facorro, B., Roiz-Santiáñez, R., Pérez-Iglesias, R., Tordesillas-Gutiérrez, D., Mata, I., Rodríguez-Sánchez, J. M., de Lucas, E. M., & Vázquez-Barquero, J. L. (2009). Specific brain structural abnormalities in first-episode schizophrenia. A comparative study with patients with schizophreniform disorder, non-schizophrenic non-affective psychoses and healthy volunteers. *Schizophrenia Research*, *115*(2-3), 191–201.

Dafflon, J., Pinaya, W. H. L., Turkheimer, F., Cole, J. H., Leech, R., Harris, M. A., Cox, S. R., Whalley, H. C., McIntosh, A. M., & Hellyer, P. J. (2020). An automated machine learning approach to predict brain age from cortical anatomical measures. *Human Brain Mapping*, *41*(13), 3555–3566.

Dai, L., Zhou, H., Xu, X., & Zuo, Z. (2019). Brain structural and functional changes in patients with major depressive disorder: A literature review. *PeerJ*, *7*, e8170.

Delgado, P. L. (2000). Depression: The Case for a Monoamine Deficiency. *The Journal of Clinical Psychiatry*, *61*(suppl 6), 4165.

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

DGPPN, BÄK, KBV & AWMF. (2017). *S3-leitlinie/nationale versorgungsleitlinie unipolare depression – kurzfassung, 2. auflage. version 1.*

Dickersin, K., & Min, Y. (1993). NIH clinical trials and publication bias. *The Online journal of current clinical trials*, *Doc No 50*.

Dienes, K. A., Hazel, N. A., & Hammen, C. L. (2013). Cortisol Secretion in Depressed and At-Risk Adults. *Psychoneuroendocrinology*, *38*(6), 927–940.

Dougherty, E. R., Hua, J., & Sima, C. (2009). Performance of Feature Selection Methods. *Current Genomics*, *10*(6), 365–374.

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872.

Eickhoff, S. B., Yeo, B. T. T., & Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, *19*(11), 672–686.

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science (New York, N.Y.)*, *196*(4286), 129–136.

Erbel, R., Eisele, L., Moebus, S., Dragano, N., Möhlenkamp, S., Bauer, M., Kälsch, H., & Jöckel, K.-H. (2012). [The Heinz Nixdorf Recall study]. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, *55*(6-7), 809–815.

Erbel, R., Möhlenkamp, S., Moebus, S., Schmermund, A., Lehmann, N., Stang, A., Dragano, N., Grönemeyer, D., Seibel, R., Kälsch, H., Bröcker-Preuss, M., Mann, K., Siegrist, J., & Jöckel, K.-H. (2010). Coronary Risk Stratification, Discrimination, and Reclassification Improvement Based on Quantification of Subclinical Coronary Atherosclerosis: The Heinz Nixdorf Recall Study. *Journal of the American College of Cardiology*, *56*(17), 1397–1406.

Escanilla, N. S., Hellerstein, L., Kleiman, R., Kuang, Z., Shull, J. D., & Page, D. (2018). Recursive Feature Elimination by Sensitivity Testing. *Proceedings of the ...*

International Conference on Machine Learning and Applications. International Conference on Machine Learning and Applications, *2018*, 40–47.

Fabris, F., & Freitas, A. A. (2019). Analysing the Overfit of the Auto-sklearn Automated Machine Learning Tool [Series Title: Lecture Notes in Computer Science]. In G. Nicosia, P. Pardalos, R. Umeton, G. Giuffrida & V. Sciacca (Eds.), *Machine Learning, Optimization, and Data Science* (pp. 508–520). Springer International Publishing.

Falk, E. B., Hyde, L. W., Mitchell, C., Faul, J., Gonzalez, R., Heitzeg, M. M., Keating, D. P., Langa, K. M., Martz, M. E., Maslowsky, J., Morrison, F. J., Noll, D. C., Patrick, M. E., Pfeffer, F. T., Reuter-Lorenz, P. A., Thomason, M. E., Davis-Kean, P., Monk, C. S., & Schulenberg, J. (2013). What is a representative brain? Neuroscience meets population science. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(44), 17615–17622.

Fama, R., & Sullivan, E. V. (2015). Thalamic structures and associated cognitive functions: Relations with age and aging. *Neuroscience and biobehavioral reviews*, *54*, 29–37.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F. (2021). Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning. *arXiv:2007.04074*.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2019). Auto-sklearn: Efficient and robust automated machine learning. In F. Hutter, L. Kotthoff & J. Vanschoren (Eds.), *Automated machine learning: Methods, systems, challenges* (pp. 113–134). Springer International Publishing.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole Brain Segmentation. *Neuron*, *33*(3), 341–355.

Flint, C., Cearns, M., Opel, N., Redlich, R., Mehler, D. M. A., Emden, D., Winter, N. R., Leenings, R., Eickhoff, S. B., Kircher, T., Krug, A., Nenadic, I., Arolt, V., Clark, S., Baune, B. T., Jiang, X., Dannlowski, U., & Hahn, T. (2021). Systematic

misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology*, *46*(8), 1510–1517.

Flint, C., Förster, K., Koser, S. A., Konrad, C., Zwitserlood, P., Berger, K., Hermesdorf, M., Kircher, T., Nenadic, I., Krug, A., Baune, B. T., Dohm, K., Redlich, R., Opel, N., Arolt, V., Hahn, T., Jiang, X., Dannlowski, U., & Grotegerd, D. (2020). Biological sex classification with structural MRI data shows increased misclassification in transgender women. *Neuropsychopharmacology*, *45*(10), 1758–1765.

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, *24*(11), 1037–1052.

Gijsbers, P., LeDell, E., Thomas, J., Poirier, S., Bischl, B., & Vanschoren, J. (2019). An open source automl benchmark. *arXiv preprint, arXiv:1907.00909*.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Current psychiatry reports*, *21*(11), 116.

Gray, J. P., Müller, V. I., Eickhoff, S. B., & Fox, P. T. (2020). Multimodal Abnormalities of Brain Structure and Function in Major Depressive Disorder: A Meta-Analysis of Neuroimaging Studies. *American Journal of Psychiatry*, *177*(5), 422–434.

Hamilton, M. (1959). The Assessment of Anxiety States by Rating. *British Journal of Medical Psychology*, *32*(1), 50–55.

Hamilton, M. (1960). A RATING SCALE FOR DEPRESSION. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*(1), 56–62.

Hilger, K., Winter, N. R., Leenings, R., Sassenhagen, J., Hahn, T., Basten, U., & Fiebach, C. J. (2020). Predicting intelligence from brain gray matter volume. *Brain Structure and Function*, *225*(7), 2111–2129.

Hirschfeld, R. M. (2000). History and evolution of the monoamine hypothesis of depression. *The Journal of Clinical Psychiatry*, *61 Suppl 6*, 4–6.

Holtzheimer 3rd, P. E., & Nemeroff, C. B. (2006). Future prospects in depression research. *Dialogues in Clinical Neuroscience*, *8*(2), 175–189.

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*, *119*, 456–467.

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, *46*(12), 2455.

Jain, A. K., & Waller, W. G. (1978). On the optimal number of features in the classification of multivariate Gaussian data. *Pattern Recognition*, *10*(5), 365–374.

Jiang, H., Lu, N., Chen, K., Yao, L., Li, K., Zhang, J., & Guo, X. (2020). Predicting Brain Age of Healthy Adults Based on Structural MRI Parcellation Using Convolutional Neural Networks. *Frontiers in Neurology*, *10*, 1346.

Jiřík, M., Moulisová, V., Hlaváč, M., Železný, M., & Liška, V. (2022). Artificial neural networks and computer vision in medicine and surgery. *Rozhledy V Chirurgii: Mesicnik Ceskoslovenske Chirurgicke Spolecnosti*, *101*(12), 564–570.

Jockwitz, C., Mérillat, S., Liem, F., Oschwald, J., Amunts, K., Jäncke, L., & Caspers, S. (2021). Generalizing Longitudinal Age Effects on Brain Structure – A Two-Study Comparison Approach. *Frontiers in Human Neuroscience*, *15*, 635687.

Joel, D., Persico, A., Salhov, M., Berman, Z., Oligschläger, S., Meilijson, I., & Averbuch, A. (2018). Analysis of Human Brain Structure Reveals that the Brain "Types" Typical of Males Are Also Typical of Females, and Vice Versa. *Frontiers in Human Neuroscience*, *12*.

Jung, A. (2022). *Machine Learning: The Basics*. arXiv:1805.05052.

Kindermann, R., & Snell, L. (1980). *Markov random fields and their applications* (Vol. 1). American Mathematical Society.

LeDell, E., & Poirier, S. (2020). H2O AutoML: Scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*.

Li, Q., Zhao, Y., Chen, Z., Long, J., Dai, J., Huang, X., Lui, S., Radua, J., Vieta, E., Kemp, G. J., Sweeney, J. A., Li, F., & Gong, Q. (2020). Meta-analysis of cortical

thickness abnormalities in medication-free patients with major depressive disorder. *Neuropsychopharmacology*, *45*(4), 703–712.

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: A big comparison for nas. *arXiv:1912.06059*.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654–660.

Maske, U. E., Buttery, A. K., Beesdo-Baum, K., Riedel-Heller, S., Hapke, U., & Busch, M. A. (2016). Prevalence and correlates of DSM-IV-TR major depressive disorder, self-reported diagnosed depression and current depressive symptoms among adults in Germany. *Journal of Affective Disorders*, *190*, 167–177.

McEwen, B. S., Chattarji, S., Diamond, D. M., Jay, T. M., Reagan, L. P., Svenningsson, P., & Fuchs, E. (2010). The neurobiological properties of Tianeptine (Stablon): From monoamine hypothesis to glutamatergic modulation. *Molecular psychiatry*, *15*(3), 237–249.

McLaughlin, K. A. (2011). The Public Health Impact of Major Depression: A Call for Interdisciplinary Prevention Efforts. *Prevention science : the official journal of the Society for Prevention Research*, *12*(4), 361–371.

Moncrieff, J., Cooper, R. E., Stockmann, T., Amendola, S., Hengartner, M. P., & Horowitz, M. A. (2022). The serotonin theory of depression: A systematic umbrella review of the evidence. *Molecular Psychiatry*.

Mourao-Miranda, J., Reinders, A., Rocha-Rego, V., Lappin, J., Rondina, J., Morgan, C., Morgan, K. D., Fearon, P., Jones, P. B., Doody, G. A., et al. (2012). Individualized prediction of illness course at the first psychotic episode: A support vector machine mri study. *Psychological medicine*, *42*(5), 1037–1047.

Musigmann, M., Akkurt, B. H., Krähling, H., Nacul, N. G., Remonda, L., Sartoretti, T., Henssen, D., Brokinkel, B., Stummer, W., Heindel, W., & Mannil, M. (2022).

Testing the applicability and performance of Auto ML for potential applications in diagnostic neuroradiology. *Scientific Reports*, *12*(1), 13648.

Mutasa, S., Sun, S., & Ha, R. (2020). Understanding artificial intelligence based radiology studies: What is overfitting? *Clinical imaging*, *65*, 96–99.

Nandam, L. S., Brazel, M., Zhou, M., & Jhaveri, D. J. (2020). Cortisol and Major Depressive Disorder—Translating Findings From Humans to Animal Models and Back. *Frontiers in Psychiatry*, *10*, 974.

Navarro, C. L. A., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ*, *375*, n2281.

Nieuwenhuis, M., Schnack, H. G., van Haren, N. E., Lappin, J., Morgan, C., Reinders, A. A., Gutierrez-Tordesillas, D., Roiz-Santiañez, R., Schaufelberger, M. S., Rosa, P. G., Zanetti, M. V., Busatto, G. F., Crespo-Facorro, B., McGorry, P. D., Velakoulis, D., Pantelis, C., Wood, S. J., Kahn, R. S., Mourao-Miranda, J., & Dazzan, P. (2017). Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients. *Neuroimage*, *145*(Pt B), 246–253.

Nordanskog, P., Dahlstrand, U., Larsson, M. R., Larsson, E.-M., Knutsson, L., & Johanson, A. (2010). Increase in hippocampal volume after electroconvulsive therapy in patients with depression: A volumetric magnetic resonance imaging study. *The journal of ECT*, *26*(1), 62–67.

Olson, R. S., & Moore, J. H. (2016). Tpot: A tree-based pipeline optimization tool for automating machine learning. In F. Hutter, L. Kotthoff & J. Vanschoren (Eds.), *Proceedings of the workshop on automatic machine learning* (pp. 66–74). PMLR.

Onaolapo, A. Y., & Onaolapo, O. J. (2021). Glutamate and depression: Reflecting a deepening knowledge of the gut and brain effects of a ubiquitous molecule. *World Journal of Psychiatry*, *11*(7), 297–315.

Opel, N., Cearns, M., Clark, S., Toben, C., Grotegerd, D., Heindel, W., Kugel, H., Teuber, A., Minnerup, H., Berger, K., Dannlowski, U., & Baune, B. T. (2019). Large-scale evidence for an association between low-grade peripheral inflammation and brain structural alterations in major depression in the BiDirect study. *Journal of psychiatry & neuroscience: JPN*, *44*(6), 423–431.

Patel, M., Khalaf, A., & Aizenstein, H. (2016). Studying depression using imaging and machine learning methods. *NeuroImage: Clinical*, *10*.

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

Rabovsky, K., & Stoppe, G. (2008). *Diagnosenübergreifende und multimodale psychoedukation*.

Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules*, *26*(4), 1111.

Rebala, G., Ravi, A., & Churiwala, S. (2019). Machine Learning Definition and Basics. In G. Rebala, A. Ravi & S. Churiwala (Eds.), *An Introduction to Machine Learning* (pp. 1–17). Springer International Publishing.

Rehm, J., & Shield, K. D. (2019). Global Burden of Disease and the Impact of Mental and Addictive Disorders. *Current Psychiatry Reports*, *21*(2), 10.

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, *112*, 103375.

Rose, S. (2018). Machine Learning for Prediction in Electronic Health Data. *JAMA network open*, *1*(4), e181404.

Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis : Prevention, assessment and adjustments*.

Ruigrok, A. N., Salimi-Khorshidi, G., Lai, M.-C., Baron-Cohen, S., Lombardo, M. V., Tait, R. J., & Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience and Biobehavioral Reviews*, *39*(100), 34–50.

Sanacora, G., Treccani, G., & Popoli, M. (2012). Towards a glutamate hypothesis of depression. *Neuropharmacology*, *62*(1), 63–77.

Sapolsky, R. M., Krey, L. C., & McEwen, B. S. (1984). Glucocorticoid-sensitive hippocampal neurons are involved in terminating the adrenocortical stress response. *Proceedings of the National Academy of Sciences*, *81*(19), 6174–6177.

Sapolsky, R. (1985). A mechanism for glucocorticoid toxicity in the hippocampus: Increased neuronal vulnerability to metabolic insults. *The Journal of Neuroscience*, *5*(5), 1228–1232.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex (New York, NY)*, *28*(9), 3095–3114.

Schaufelberger, M. S., Duran, F. L. S., Lappin, J. M., Scazufca, M., Amaro, E., Leite, C. C., de Castro, C. C., Murray, R. M., McGuire, P. K., Menezes, P. R., & Busatto, G. F. (2007). Grey matter abnormalities in Brazilians with first-episode psychosis. *The British Journal of Psychiatry. Supplement*, *51*, 117–122.

Schmaal, L., Hibar, D. P., Sämann, P. G., Hall, G. B., Baune, B. T., Jahanshad, N., Cheung, J. W., van Erp, T. G. M., Bos, D., Ikram, M. A., Vernooij, M. W., Niessen, W. J., Tiemeier, H., Hofman, A., Wittfeld, K., Grabe, H. J., Janowitz, D., Bülow, R., Selonke, M., … Veltman, D. J. (2017). 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. *Molecular Psychiatry*, *22*(6), 900–909.

Schulz, M.-A., Bzdok, D., Haufe, S., Haynes, J.-D., & Ritter, K. (2022). *Performance reserves in brain-imaging-based phenotype prediction* (preprint). Neuroscience.

Shewhart, W. A., & Deming, W. E. (1939). *Statistical method from the viewpoint of quality control*. Washington, The Graduate School, The Department of Agriculture.

Sima, C., & Dougherty, E. (2008). The peaking phenomenon in the presence of feature-selection. *Pattern Recognition Letters*, *29*, 1667–1674.

Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., & Miller, K. L. (2019). Estimation of brain age delta from brain imaging. *NeuroImage*, *200*, 528–539.

Steffen, A., Thom, J., Jacobi, F., Holstiege, J., & Bätzing, J. (2020). Trends in prevalence of depression in Germany between 2009 and 2017 based on nationwide ambulatory claims data. *Journal of Affective Disorders*, *271*, 239–247.

Stolicyn, A., Harris, M. A., Shen, X., Barbu, M. C., Adams, M. J., Hawkins, E. L., de Nooij, L., Yeung, H. W., Murray, A. D., Lawrie, S. M., Steele, J. D., McIntosh, A. M., & Whalley, H. C. (2020). Automated classification of depression from structural brain measures across two independent community-based cohorts. *Human Brain Mapping*, *41*(14), 3922–3937.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779.

Teismann, H., Minnerup, H., Nagel, M., Arolt, V., Heindel, W., Baune, B., Wellmann, J., Hense, H., & Berger, K. (2014). Establishing the bidirectional relationship between depression and subclinical arteriosclerosis – rationale, design, and characteristics of the BiDirect Study. *BMC psychiatry*, *14*, 174.

Teuber, A., Sundermann, B., Kugel, H., Schwindt, W., Heindel, W., Minnerup, J., Dannlowski, U., Berger, K., & Wersching, H. (2017). MR imaging of the brain in large cohort studies: Feasibility report of the population- and patient-based BiDirect study. *European Radiology*, *27*(1), 231–238.

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2012). Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms.

Treder, M. S., Shock, J. P., Stein, D. J., du Plessis, S., Seedat, S., & Tsvetanov, K. A. (2021). Correlation Constraints for Regression Models: Controlling Bias in Brain Age Prediction. *Frontiers in Psychiatry*, *12*, 615754.

Trifu, S. C., Trifu, A. C., Aluaş, E., Tătaru, M. A., & Costea, R. V. (2020). Brain changes in depression. *Romanian Journal of Morphology and Embryology*, *61*(2), 361–370.

Turner, J. A. (2014). The rise of large-scale imaging studies in psychiatry. *GigaScience*, *3*, 29.

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77.

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

Velakoulis, D., Wood, S. J., Wong, M. T. H., McGorry, P. D., Yung, A., Phillips, L., Smith, D., Brewer, W., Proffitt, T., Desmond, P., & Pantelis, C. (2006). Hippocampal and amygdala volumes according to psychosis stage and diagnosis: A magnetic resonance imaging study of chronic schizophrenia, first-episode psychosis, and ultra-high-risk individuals. *Archives of General Psychiatry*, *63*(2), 139–149.

Videbech, P., & Ravnkilde, B. (2004). Hippocampal volume and depression: A meta-analysis of MRI studies. *The American Journal of Psychiatry*, *161*(11), 1957–1966.

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106.

Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., Abolhassani, H., Aboyans, V., Abrams, E. M., Abreu, L. G., Abrigo, M. R. M., Abu-Raddad, L. J., Abushouk, A. I., … Murray, C. J. L. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1204–1222.

Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, *104*, 101822.

Wickenberg-Bolin, U., Göransson, H., Fryknäs, M., Gustafsson, M. G., & Isaksson, A. (2006). Improved variance estimation of classification performance via reduction of bias caused by small sample size. *BMC bioinformatics*, *7*, 127.

World Health Organization, W. (1993). *The icd-10 classification of mental and behavioural disorders*.

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*, 022022.

Zhang, F.-F., Peng, W., Sweeney, J. A., Jia, Z.-Y., & Gong, Q.-Y. (2018). Brain structure alterations in depression: Psychoradiological evidence. *CNS Neuroscience & Therapeutics*, *24*(11), 994–1003.

Zöller, M.-A., & Huber, M. F. (2021). Benchmark and Survey of Automated Machine Learning Frameworks. *arXiv:1904.12054*.

Acknowledgements

I would like to express my sincere gratitude to Univ.-Prof. Dr. med. Dr. rer. pol. Svenja Caspers as well as Prof. Dr. med. Dipl.-Inform. Julian Caspers for their thorough supervision and support at every step of the project, from the concretization of the original idea down to the final lines. Without the trust they displayed from the start as well as the time and resources they generously gave, this project would not have come to completion.

My thanks furthermore go to Dr. med. Christian Rubbert and Dr. rer. medic. Christiane Jockwitz, who provided me with many salutatory insights and guidance throughout the exercise.

A special place is to be given to Ms. Camilla Kraemer, who provided continuous support with as much expertise as engagement in the later stages of this work and tremendously helped overcome critical situations.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 945539 (HBP SGA3; SC).