Reproducibility of electroencephalography alpha band biomarkers for diagnosis of major depressive disorder

Hollenbenders, Y.[a,b], Maier, C.[a,c], & Reichenbach, A.[a,b]

[a]Center for Machine Learning, Heilbronn University, Max-Planck-Str. 39, 74081 Heilbronn, Germany

[b]Medical Faculty Heidelberg, University of Heidelberg, Grabengasse 1, 69117 Heidelberg, Germany

[c]Medical Informatics, Heilbronn University, Max-Planck-Str. 39, 74081 Heilbronn, Germany

* Corresponding Author: alexandra.reichenbach@hs-heilbronn.de

Keywords: Major Depressive Disorder, Electroencephalography, Biomarker, Diagnosis, Decision-Support, Classification, Multiverse Analysis, Reproducibility, Replicability, Robustness

Abstract

Major depressive disorder (MDD) and other psychiatric diseases can greatly benefit from objective decision support in diagnosis and therapy. Machine learning approaches based on electroencephalography (EEG) have the potential to serve as low-cost decision support systems. Despite the successful demonstration of this approach, contradictory findings regarding the diagnostic value of those biomarkers hamper their deployment in a clinical setting. Therefore, the *reproducibility* and *robustness* of these biomarkers needs to be established first. We employ a multiverse analysis to systematically investigate variations in five data processing steps, which may be one source of contradictory findings. These steps are normalization, time-series segment length, biomarker from the alpha band, aggregation, and classification algorithm. For *replicability* of our results, we utilize two publicly available EEG data sets with eyes-closed resting-state data containing 16/19 MDD patients and 14/14 healthy control subjects. The diagnostic classifiers range from chance level up to 85%, dependent on dataset and combination of processing steps. We find a large influence of choice of processing steps and their combinations. However, only the biomarker has an overall significant effect on both datasets. We find one biomarker candidate that has shown a *robust* and *reproducible* high performance for MDD diagnostic support, the **relative centroid frequency**. Overall, the *replicability* of our findings with the two datasets is rather inconsistent. This study is a showcase for the advantages of employing a multiverse approach in EEG data analysis and advocates for larger, well-curated data sets to further neuroscience research that can be translated to clinical practice.

Introduction

Major depressive disorder (MDD) is the most common mental disorder with a worldwide lifetime prevalence of 10.6%[1]. However, healthcare research assumes that a substantial number of cases is unreported, which might be due to current diagnostic practices, stigma, or public health policies[2]. Current diagnostic procedures are based on semi-structured interviews and self-assessed questionnaires, making the diagnosis rather subjective and dependent on the clinician's experience, the current condition of the patient, and their ability for self-assessment[3]. Objective and reliable physiological biomarkers have the potential to support and accelerate diagnostic and therapeutic decisions[4]. A non-invasive and cost-efficient method to provide such biomarkers by recording brain function is electroencephalography (EEG). Studies have shown that biomarkers extracted from EEG signals can distinguish between MDD patients and healthy controls (HC)[5–7]. However, results contradict each other regarding discriminatory features extracted from these signals, possibly due to the variety in EEG data acquisition and processing[7,8]. Therefore, it is necessary to establish the *reproducibility* and *robustness* of these biomarkers before they can be utilized in a clinical setting[9].

One source of inter-study variability, which may confound reproducibility, is the selection of participants. In addition to general factors such as age, gender, and ethnicity, a multifaceted disorder such as MDD is heterogeneous in terms of severity and symptoms, progression, (drug) treatment (success), and the presence of comorbidities[10]. Furthermore, researchers assign the label MDD on different foundations, e.g. by clinical diagnosis based on the DSM-IV manual, or by questionnaires and scores such as the Patient Health Questionnaire-9 (PHQ-9)[11]. Sample sizes of 50 participants (MDD and HC) that are commonly used in classical hypothesis-driven studies using statistical analysis[8] are very likely too small to capture the variability observed in MDD. As a consequence, the effect sizes of studies in neuroscience are often inflated[9,12,13], leading to overestimation of the selectivity of the biomarkers studied. Comparability of studies may further be hampered by the lack of standardization in the recording of EEG data. Not just different registration settings like the time of the day, temperature of the room or resting-state conditions can introduce variability. Technical differences like the electrode placements, e.g. the quasi-standard of the EEG-10-20/10/5 system[14] or the Geodesic Sensor Net configuration[15] might be other sources of variability. Furthermore, even landmark-standardized placements might record different physiological sources between participants[16]. More variability in technical recording properties comes from the choice of the electrode type, recording reference, sampling frequency, and acquisition hardware. Finding reliable biomarkers across these variations requires data sharing, and collaborative efforts are sorely needed[9]. *Replication* of results with publicly available datasets or newly recorded data and the original processing pipeline is necessary to strengthen the findings[9].

Once the data is recorded, preprocessing steps such as filtering, artifact removal, normalization, segmenting, subsampling or augmentation, and data aggregation are usually applied to the EEG data. Different choices in these steps between studies introduce further variability[6]. When biomarkers specific to the "classical" EEG bands are calculated, the bands are not always defined in the same frequency ranges. E.g. the alpha band, while still one of the most consistently defined frequency bands, is typically defined from 8 to 13 Hz, but can range within the borders of 6 to 14 Hz[8]. Biomarkers sharing the same name may be calculated in different ways such as the frequently reported alpha asymmetry[17], hampering *analytical reproducibility* and *replicability*[9]. The next substantial difference across studies is the actual analytical method for differentiation between patients and HCs. These differences can be analyzed with statistical comparisons, regression models[11], or multivariate methods such as machine learning algorithms. For the latter methods, the decision-making is most frequently modeled as a classification problem with the classes MDD and HC. The approaches range from classical supervised algorithms that usually operate on hand-crafted biomarkers[18] to more complex deep learning models that often but not exclusively operate on the raw time-series data or their frequency domain representation, and inherently derive biomarkers in the training process[6]. While for the former methods, the most common way to derive biomarkers is the application of feature importance algorithms, the latter are often analyzed with methods of Explainable Artificial Intelligence (XAI)[19], another step to introduce variability in data analysis. In most studies, the data recording and processing pipeline consists of one specific choice out of several possible alternatives. Yet, different methodical decisions during this process can have an impact on the findings, rendering comparison between studies hard[18]. One approach to reconcile the different findings and demonstrate robustness of the analysis of EEG data to the analytical variability is the multiverse analysis[20]. Neuroscience researchers have just started to use this approach to investigate the reproducibility and robustness of specific biomarkers[11], or to recommend an optimized processing pipeline[21]. This analysis strategy demonstrates that high number of degrees of freedom in conducting EEG studies is one likely cause for the lack of *reproducibility*, *replicability,* and *robustness* in findings[7,8].

One controversial group of biomarkers for MDD are measures derived from the alpha band. Alpha waves are connected to relaxation and dominantly present when eyes are closed, and their characteristics are frequently used in studies aimed at diagnosing MDD[5]. While some studies suggest that biomarkers from the alpha band perform better than markers from other frequency bands[22,23], other studies found that they perform worse[24–26]. Furthermore, it is disputed whether alpha activity is higher or lower in MDD subjects compared to HCs[22,23]. Noteworthy, all these studies used different processing methods, therefore their results are not comparable[27,28]. Yet, to translate results from research to diagnosis support in a clinical setting, objective physiological biomarkers with high

discriminatory power, replicability across datasets, and to some degree robustness to analytical variability are required.

The goal of this study is to assess two of the main types of *reproducibility* on EEG alpha band biomarkers for the support of MDD diagnosis. We focus our analysis on classical machine learning algorithms with hand-crafted biomarkers to mimic a diagnostic scenario while retaining the explainability of the models, and the close connection to studies using statistical analysis on those biomarkers. We also limit our investigation to alpha markers, the most prominently discussed EEG biomarkers for MDD[5]. For the purpose of *replication*, we use two publicly available datasets with eyes-closed resting-state EEG and harmonize them[29] to limit preprocessing variability at this point. To demonstrate *robustness* of the biomarkers against the processing pipeline, we systematically investigate the influence of a selection of preprocessing steps, different features, and classification algorithms with a multiverse analysis. The main analytic goals are to find biomarkers that have high diagnostic value and are replicable across datasets, as well as shedding light on the influence of processing steps and finding markers that are robust against those different processing choices. This approach can readily be transferred to optimize the processing pipeline of biomarkers with high diagnostic value. This study allows us to reconcile conflicting results in literature that putatively arise from processing differences across studies and provide an approach for finding reproducible EEG biomarkers for the diagnosis of MDD, or any other neuroimaging biomarker for psychiatric diseases.

Methods

Data

Two publicly available datasets (D1[30]; D2[31]) with MDD patients (D1: n=33 (presumably age: 40.3±12.9 ; female: n=17[32]), D2: n=24 (age: 30.9±10.4; female: n=11)) and HCs (D1: n=30 (presumably age: 38.3±15.6; female: n=9), D2: n=29 (age: 31.5±9.2; female: n=9)) were downloaded in August and October 2021, respectively. The study designs were approved by the respective ethics committees (D1: Human ethics committee of the Hospital Universiti Sains Malaysia (HUSM); D2: Ethics Committee for Biomedical Research at the Lanzhou University Second Hospital) and participants provided written informed consent. All patients were diagnosed based on DSM-IV criteria. Patients from both datasets did not take any anti-depression medication during the two weeks before recordings were taken. Both datasets contain 5-minute resting-state EEG time series data with eyes closed. D1 was recorded with 19 electrodes placed according to the 10-20 system[14] with linked-ear-reference and 256 Hz sample frequency. D2 was recorded with 128 electrodes in the Geodesic Sensor Net[15] with Cz as reference electrode and 250 Hz sample frequency. Lastly, MDD patients from D1 underwent two questionnaires, i.e. the Beck Depression Inventory-II (BDI-II)[33] and the Hospital Anxiety and Depression Scale (HADS)[34] with scores of 20.6±8.6 and 10.7±2.4, respectively. Patients from D2, in contrast, were assessed with the Patient Health Questionnaire-9item (PHQ-9).

Cleaning and Dataset Harmonization

To establish a common ground for the analyses, we harmonized the datasets as far as possible using the Python Package *MNE*[35] (Version: 1.2.3). We matched the 128 electrodes from D2 to correspond to the 10-20 system according to the supplementary material provided by the dataset. Electrodes not contained in both datasets were omitted, leaving 13 common channels: seven (pre)frontal (Fp1/2, F3/4, F7/8, Fz), two central (C3/4), two parietal (P3/4), and two occipital (O1/2). Afterward, both datasets were re-referenced to average for the later independent component analysis (ICA) and D1 was resampled to 250 Hz to match the sampling rate of D2. To harmonize the spectral bandwidth, both datasets were bandpass filtered from 1 to 40 Hz as common denominator of the different filters applied to the datasets before upload. Artifact components, i.e. eye blinks, muscle artifacts, heartbeat, line noise, and channel noise, were extracted by ICA[36], automatically labeled with *ICLabel*[37] (Version: 0.4), and EEG signals were reconstructed without the artifact components. Note that artifacts can be removed using various techniques[6,38]. However, this is a very extensive methodological topic in itself and their variations were therefore not included in our multiverse analysis.

Multiverse Analysis

To investigate the impact of processing choice on the classification results, we adapted a multiverse analysis[20]. For each step (also referred to as factor) of the processing pipeline, we implemented different variations (also referred to as factor levels) and combined them fully combinatorically. This procedure led us to a multiverse with 864 parallel processing paths (Table 1).

Table 1: Processing steps / factors and their variations / factor levels included in the multiverse analysis

Processing Stage	**Processing Step / Factor**	**Variations / Factor Levels**	**Multiverse Analysis Conditions (# paths total)**
Preprocessing	Normalization	subject-wise; channel-wise; none	3 (3)
	Segment Length	5s; 10s; 15s; 20s	4 (12)
	Outlier segment removal (for all conditions)		
Feature Extraction	Biomarker	11 alpha band (8-13 Hz) biomarker (cf. Table 2) and one composite biomarker	12 (144)
	Subsampling to fixed number of 10 segments (for all conditions)		
	Aggregation	None; median over 10 segments	2 (288)
Classification	Algorithm	Support Vector Machine (SVM); Logistic Regression (LR); Random Forest (RF)	3 (864)
			864 paths

Preprocessing

Normalization is a common method to preprocess EEG signals before classification[6]. To compare the impact of normalization methods, we applied three variations of *z*-normalization: 1) No normalization. 2) Subject-wise normalization is applied to level putative differences across subjects. For this, all channels of each subject are *z*-normalized jointly to align only the subjects of the datasets while retaining the between-channel differences for each subject. 3) Channel-wise normalization is frequently used when the raw time-series data is used in order to level putative differences across EEG traces. Here, each channel of each subject is *z*-normalized separately. Note that channel-wise normalization destroys spatial information based on absolute values of the power of the signal (**absolute bandpower**, **envelope median**, **envelope interquartile range**, **envelope variance**, **envelope**

range). Biomarkers based on distribution (**envelope skewness**, **envelope kurtosis**), or biomarkers normalized with respect to the full band (**relative centroid frequency**, **relative bandpower**) or a certain frequency (**absolute centroid frequency**, **peak frequency**) are not affected (cf. Table 2).

After normalizing, we split the data into non-overlapping segments of 20s. To identify outlier segments and exclude them from further analyses, we identified the maximum and minimum values of each segment for each subject. An outlier is defined as having values exceeding or falling below, respectively, the mean of the maximum/minimum values plus/minus two times the standard deviation of the maximum/minimum values. Subjects with less than 10 remaining segments were excluded from further analyses, while subjects with more than 10 segments were subsampled to 10 to prevent over- or underrepresentation of single subjects. This left 30 (MDD: 53%) / 33 (MDD: 58%) subjects for D1/D2. Splitting recordings into segments of the same duration is a common method to augment data, yet there is no best practice on how long these segments should be and reported time spans vary widely. Shen et al.[26] have demonstrated that segment length has a significant impact on classification results. Therefore, we also considered the first 5, 10, and 15 seconds of each segment for our analysis. Note that segments can also be cut in an overlapping fashion, but this additional variation would have blown our analysis out of proportion.

Feature Extraction

We extracted eleven biomarkers (Table 2) from the included segments. To compile a comprehensive set of alpha band biomarkers, we calculated spectral biomarkers common for diagnosis of psychiatric diseases[5]. We only chose biomarkers that can be calculated for each electrode (i.e. EEG channel) individually to keep a clear structure in the analysis. To obtain the spectral power characteristics of the signal, we used Welch's method with Hanning window, 50% overlap, and a window length of 512 data points. This results in 2.05s window duration and a spectral resolution of 0.5 Hz. We set the typical frequency range of the alpha band from 8 to 13 Hz[8]. To calculate the envelope of the signal, we first extracted the alpha band with a Butterworth bandpass filter of either 4^{th} or 8^{th} order since pretests showed deviating results. This step is not included in the statistical analyses of the multiverse, since it is not used for every biomarker. The envelope biomarkers were calculated based on the envelope of the time-domain signal[39], which was calculated with the Hilbert transformation using the absolute values. Since the upper and lower envelope of the alpha band had nearly identical absolute values, we only used the upper envelope for further analyses. The actual biomarkers were calculated on the distributions of the data points from the envelope. Because the calculation of some of the biomarkers is inconsistently or not at all described in the literature, we compiled a detailed overview over their construction to enable *analytical reproducibility* (Table 2). Biomarkers were calculated using the Python toolboxes *scipy*[40] (Version: 1.9.3), *numpy*[41] (Version: 1.23.5), and the function *bandpower* adopted from

Vallat and Walker[42]. To enable the *analytical reproducibility* of this study, source code for biomarker calculation is provided (see supplementary files "calculate_biomarkers.ipynb" and "alpha_markers.py").

*Table 2: Biomarkers used in this study, their description, calculation, and sources/studies where those biomarkers were used for MDD diagnosis or characterization. * indicates the biomarker that are invariant to normalization.*

Biomarker	**Description**	**Used In**
absolute bandpower	Total power of the alpha band. Welch periodogram, and integration using Simpson's rule (adopted from [42])	[22–26]
relative bandpower*	The absolute power of the alpha band divided by the absolute power of the EEG signal (i.e. from 1 to 40 Hz) (adopted from [42])	[24]
absolute centroid frequency*	Center of mass of the alpha bandpower spectrum. Welch periodogram normalized to unity total power, and calculating the power-weighted sum of the frequencies of the alpha band	[24,26]
relative centroid frequency*	Absolute centroid frequency of the alpha band divided by the absolute centroid frequency of the total EEG band (here: from 1 to 40 Hz)	[24]
peak frequency*	The frequency of the alpha band spectral component with the highest power. Welch periodogram, select frequency with maximum power[43]	[24,26,39]
envelope kurtosis*	Butterworth bandpass filter, Hilbert transform, and calculating the kurtosis of the absolute values of the transformed signal	[39]; on raw signal: [24]
envelope skewness*	Butterworth bandpass filter, Hilbert transform, and calculating the skewness of the absolute values of the transformed signal	[39]; on raw signal: [24]
envelope median	Butterworth bandpass filter, Hilbert transform, and calculating the median of the absolute values of the transformed signal	[39]
envelope interquartile range	Butterworth bandpass filter, Hilbert transform, and calculating the interquartile range between the 25th and the 75th quartile of the absolute values of the transformed signal	[39]

envelope variance	Butterworth bandpass filter, Hilbert transform, and calculating the variance of the absolute values of the transformed signal	[39]; on raw signal: [24]
envelope range	Butterworth bandpass filter, Hilbert transform, and calculating the range of the absolute values of the transformed signal	[39]

We processed both the eleven biomarkers separately, and their vectorial combination into one **composite biomarker**, resulting in twelve biomarkers total.

Biomarker values were calculated individually for each single time segment (and EEG channel). Since there are 10 segments per subject, this means that one subject is represented with 10 different values, which effectively corresponds to a data augmentation process commonly encountered in the field of machine learning. Alternatively, we pursued an aggregation strategy by taking the median over each subject's 10 segments, resulting in one single and putatively more robust biomarker value per subject (and EEG channel). This is a typical approach for analyses with statistical inference tests.

Classification

The two datasets were used to train individual classification models with three commonly used and simple classification algorithms: Logistic Regression (LR), Support Vector Machine (SVM) with a linear kernel, and Random Forest (RF). When a model was trained with one biomarker, its input was a 13-dimensional feature vector containing the biomarker values for each EEG channel. The composite biomarker resulted in a 156-dimensional feature vector (12 biomarkers x 13 channels), which was normalized before classification. The classification models were trained with six-fold cross-validation on a per-subject split basis[29], resulting in five to six subjects per test dataset. To counteract imbalance of the class distribution, the balanced mode was used, and data was classified as MDD or HC. Note that the partitioning in training and test dataset was kept identical for all paths of the multiverse analysis to ascertain comparability. Accuracy was calculated for each model; the raw values can be found in the supplementary material ("metrics.csv") for analytical reproducibility. We used *sklearn*[44] (Version: 1.1.3) for all classification-related implementations.

Statistics

For a quantitative analysis of the replication across datasets, we calculated analyses of variance (ANOVA) with dataset as between-subject factor, and all other factors as within-subject factors. To investigate the effect of the processing steps on biomarker performance we used ANOVAs with the

factors of the multiverse analysis as within-subjects factors for each dataset separately and subsequent ANOVAs or *t*-tests for post hoc analyses whenever appropriate. Furthermore, we used one-tailed *t*-tests against chance level, i.e. an accuracy of 50%, to assess the stability of each biomarker across the paths of the multiverse analysis. We refer to this analysis later as stable paths analysis, which combines *internal reproducibility* across folds and high performance of these models. All statistical tests were conducted with *statsmodel*[45] (Version: 0.14.0). Reported values are mean ± standard deviation unless stated otherwise. We do not correct the statistical results but report different thresholds of alpha levels to facilitate comparison to other studies.

Results

Replication of the classification results with two datasets

The overall classification performance differs significantly between datasets (cf. Figure 1; $F_{1,10}$=8.626; p<.05). Classifiers operating on D1 achieve a mean accuracy of 57.3±14.2% across all paths with a rather wide performance range from 15.0% to 85.6% mean accuracy for the individual paths. The overall performance of classifiers operating on D2 is at chance level with 51.0±8.5% with mean accuracies for the individual paths ranging from 26.7% to 73.3%. We find a significant interaction on the performance between the datasets and segment length (Figure 1B; $F_{3,12}$=4.551, p<.05) and biomarker (Figure 1E; $F_{11,44}$=3.496, p<.01). However, it should be considered that even the three top performing biomarkers (**relative centroid frequency**, **composite biomarker**, and **envelope skewness**) only show a meaningful above-average separation between HCs and MDDs for D1, i.e. random effects in classifiers for biomarkers performing around chance level might bias overall statistical effects.

Impact of processing steps on classification performance

The choice of biomarker significantly affects the classification performance for both datasets (Figure 1E; D1: $F_{11,55}$=21.583, p<.001, D2: $F_{11,55}$=4.031, p<.001). Therefore, we present detailed analyses for the individual biomarker in subsequent paragraphs. The classification performance of D1 significantly depends on the classification algorithm (Figure 1F; $F_{2,8}$=31.298, p<.001), with RF significantly outperforming the other two algorithms on the grand mean level (all t_5>5.796, p<.01). The classification performance of D2 significantly depends on segment length (Figure 1B; $F_{3,12}$=6.489, p<.01), which is driven by the significantly better performance of the 15s segments on the grand mean level (vs. 5s and 20s: t_5>4.712, p<.01). The processing steps normalization and aggregation do not affect classification accuracy for neither dataset (Figure 1A, C).

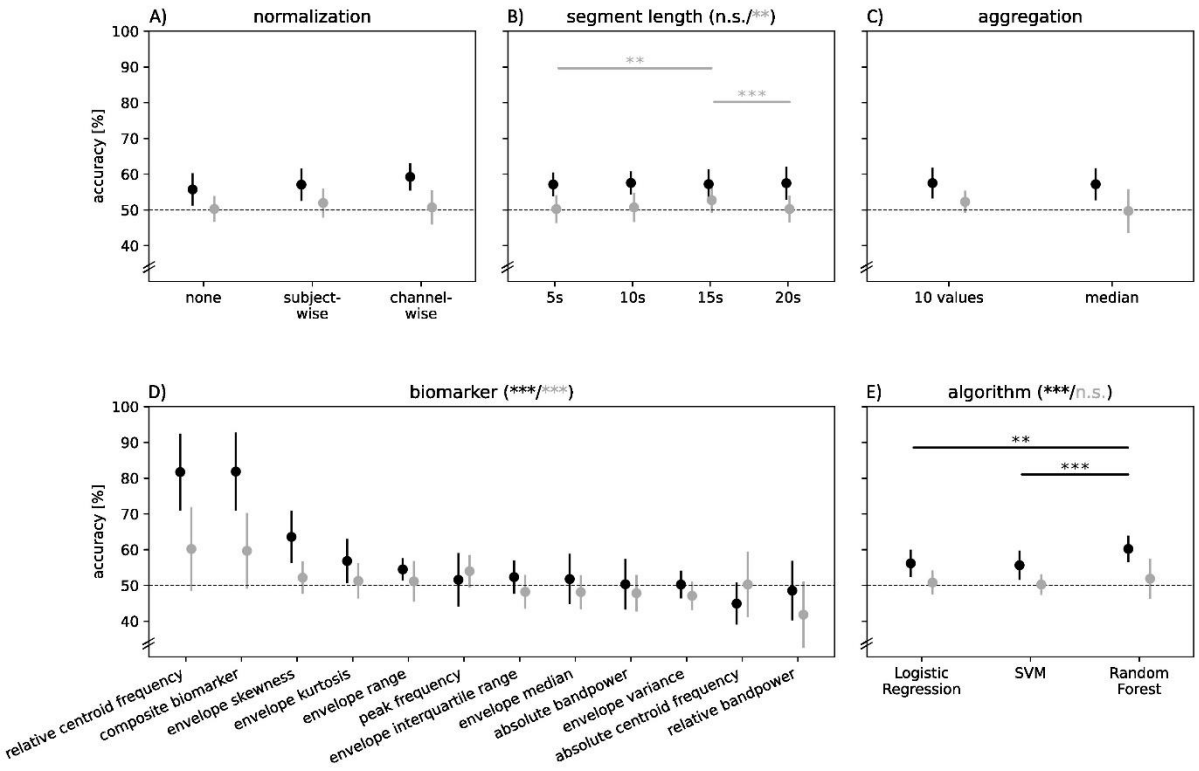

*Figure 1 Effects of processing steps, biomarkers and classification algorithms for D1 and D2. Values are averaged across factors for each fold. Error bars depict standard deviations across folds. Horizontal bars and stars depict post hoc tests with *<.05, **<.01, ***<.001, all uncorrected. Biomarkers are sorted according to their mean performance across datasets. This order is retained for all subsequent figures. Color coding applies to all elements in the figure. Abbreviations: main = main effect, D1/2 = dataset 1/2, SVM = Support Vector Machine.*

The differential performance of the biomarkers is modulated by segment length (Figure 2 top row; D1: $F_{33,165}$=2.011, p<.01, D2: $F_{33,165}$=2.644, p<.001) and algorithm (Figure 2 bottom row; D1: $F_{22,110}$=8.134, p<.001, D2: $F_{22,110}$=3.188, p<.001) in both datasets. However, post-hoc tests demonstrate that the three top biomarkers in D1 (**relative centroid frequency**, **composite biomarker**, and **envelope skewness**) are largely robust against those preprocessing steps (Figure 2A, C; all p>.05). Some of the worse performing biomarkers are significantly influenced by the choice of segment length: the **peak frequency** in both datasets (all $F_{3,15}$>5.171, p<.05), and the **composite biomarker** and **absolute centroid frequency** in D2 only (all $F_{3,15}$>4.149, p<.05). The algorithm has a significant effect on the classifiers using **absolute centroid frequency** and **relative bandpower** in both datasets (all $F_{2,10}$>4.608, p<.05), on the classifiers from **envelope kurtosis**, **envelope range**, **envelope interquartile range**, **envelope median**, and **envelope variance** in D1 only (all $F_{2,10}$>6.186, p<.05), and on the classifiers from **relative centroid frequency** and the **composite biomarker** in D2 only (all $F_{2,10}$>5.967, p<.05).

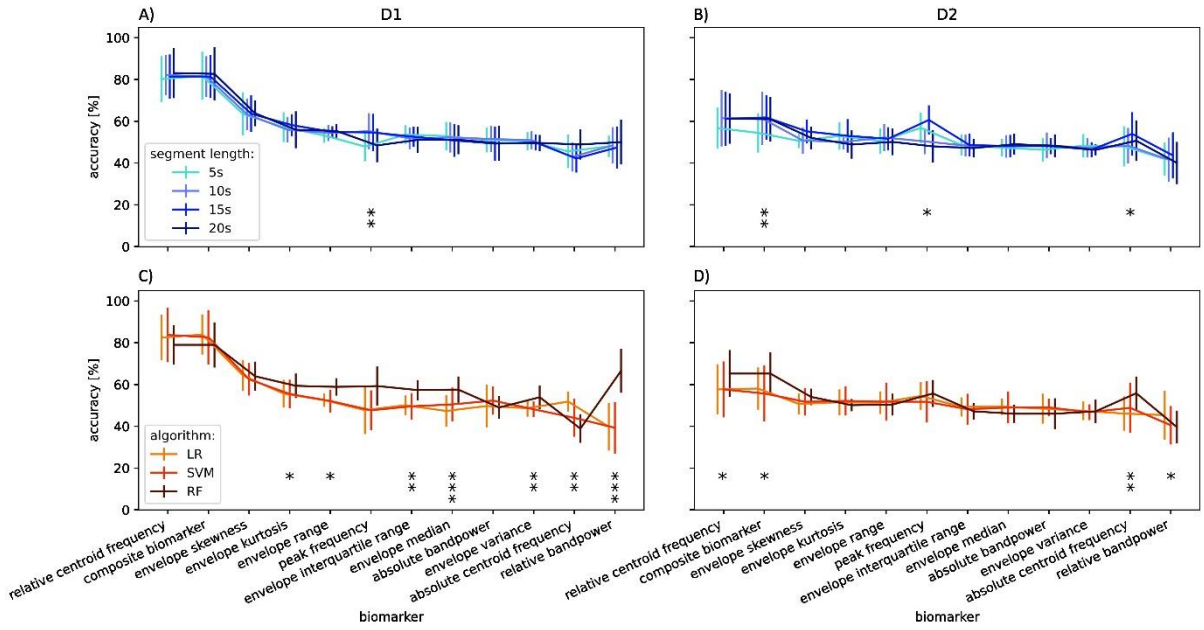

*Figure 2 Interaction effects of segment length x biomarker (top row) and algorithm x biomarker (bottom row) for D1 (left) and D2 (right). Values are averaged across factors for each fold. Error bars depict standard deviations. Stars depict an effect of the post-hoc ANOVA for the corresponding biomarker with *<.05, **<.01, ***<.001, all uncorrected. Abbreviations: D1/2 = dataset 1/2, LR = Logistic Regression, SVM = Support Vector Machine, RF = Random Forest.*

In addition to the two interactions common to both datasets, we find some differential effects for the other processing steps. Since those are dataset-specific and thus not generalizable, we only report them without further details. The raw accuracy data is provided in the supplementary material ("metrics.csv"), allowing further detailed analysis. Biomarkers are differentially affected by normalization in D1 only ($F_{22,110}$=3.744, p<.001). The interaction effects between aggregation x segment length and aggregation x algorithm are unique for D2 (both p<.05).

A detailed analysis of the influence of the processing steps on the top three biomarkers reveals that the classifiers operating on the **relative centroid frequency** are not affected by any of the processing steps in D1 (all main effects p>.073) and solely by algorithm in D2 ($F_{2,10}$=7.541, p<.05). The **composite biomarker** is also not affected by any of the preprocessing steps in D1 (all main effects p>.082), but by segment length, and algorithm in D2 (both p<.01). Lastly, we find a significant effect of the choice of normalization and aggregation on **envelope skewness** in D1 (both p<.05).

Robustness of biomarkers across paths and folds

The less impact the processing steps have on the diagnostic accuracy of a biomarker, the more *robust to analytical variability* is this biomarker (smaller range in Figure 3A). Even the most robust biomarker of this study, the **relative centroid frequency** has a performance range of nearly 9.3% for D1 and 20.0%

for D2. The least robust biomarkers are **envelope variance** (D1: range=61.0%) and **envelope skewness** (D2: range=46.7%).

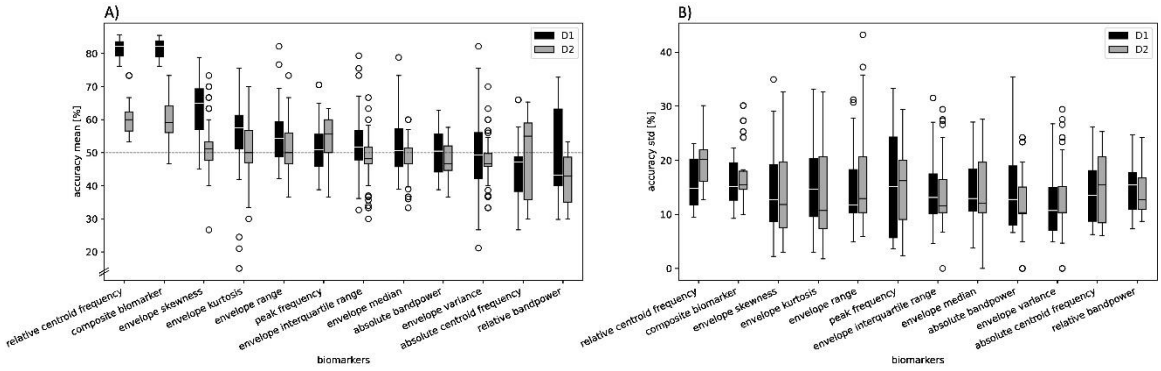

Figure 3 A) Robustness to analytical variability assessed with the spread of classification accuracies across paths. Mean accuracies for each path are averaged across folds. Dotted line depicts chance level. B) Replicability within datasets assessed with the variability of classification accuracies across folds within each path. Standard deviations across the six folds are shown for each path. Data points are depicted as outliers when their values extend beyond 1.5 times the interquartile range from the first and third quartile, respectively.

In order to demonstrate the *replicability* of the performance results within the datasets, we calculated the standard deviations of classification accuracies across folds obtained by the six-fold cross-validation for each path (Figure 3B). The lowest mean standard deviation across folds within paths is 12.1% for **envelope variance** in D1 and 12.0% for **absolute bandpower** in D2. The highest variability we find for the **relative centroid frequency** (D1: std=15.3%, D2: std=19.9%). However, these variations need to be considered with caution in this study since each test set for cross-validation only contained five to six subjects.

Examining the *robustness* of biomarkers within paths from the perspective of a quasi-meta-analysis, we find that only 11.8% of all diagnostic classifiers, i.e. 11.8% of all paths across biomarkers and datasets, perform significantly better than chance level on our most lenient alpha level of 0.01 (Figure 4A lowest dotted line). Along this line, there is no biomarker that surpasses this threshold on all paths. Furthermore, only one biomarker, the **peak frequency** in D2, contains paths that perform better than our most conservative alpha level (Figure 4A highest dotted line, p<.0001).

The *t*-statistic combines the performance of a classifier with its *internal replicability*, acknowledging the stochastic nature in the data sampling process. While inference testing is standard in medicine and neuroscience, biomarker studies from the machine-learning field sometimes neglect *replicability* in favor of performance. We find a strong positive correlation between performance (i.e. mean accuracy) and *robustness* (i.e. *t*-value) for the **relative centroid frequency** in D1 (r=.745; Figure 4B), the biomarker

with the highest performing classifiers and therefore chosen as an example. This demonstrates that even though there is a strong relationship between the two metrics, the latter should not be neglected.

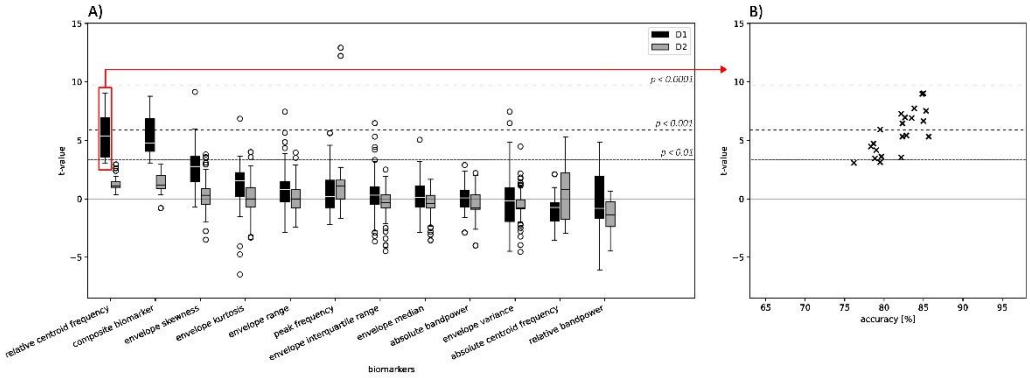

*Figure 4 Stable path analysis. A) Robustness across folds assessed with t-statistics for each path of each biomarker. B) Exemplary correlation between classification accuracy and the robustness exemplary for the **relative centroid frequency** (r=.745). Abbreviations: D1/2 = dataset 1/2. Data points are depicted as outliers when their values extend beyond 1.5 times the interquartile range from the first and third quartile, respectively. The dashed horizontal lines depict the varying degrees of the alpha level.*

Stable paths analysis

We define a stable path as a combination of processing steps that yields a diagnostic classifier performing significantly above chance level on an alpha level below 0.0001 (Figure 4A). This significance threshold was chosen to account for multiple testing in a multiverse analysis without over-emphasizing our specific design choices. Since no paths in D1 surpass this threshold, we relaxed the threshold to 0.001 for this dataset and focused on the **relative centroid frequency** for two reasons: First, there are 30 paths and not just outliers above the threshold (Figure 5) and second, it is more parsimonious than the **composite biomarker** with 156 input features. In those paths, every variation of the processing steps is represented, except for RF as classification algorithm. Median aggregation was only included in two paths as well. The four paths with the highest classification performance (accuracy = 85%) have one processing step in common: classification with the SVM algorithm. The four paths with the highest *robustness* (t >= 7.51) share only the processing of the ten segments individually. The highest classification accuracy combined with robustness was achieved with 10 individual 15s segments and SVM as classification algorithm. For this dataset and biomarker, this combination might constitute the ideal processing pipeline for a good performing and robust biomarker.

seg len [s]	5	5	5	10	10	10	15	15	20	20
aggr	indiv	indiv	median	indiv	indiv	median	indiv	indiv	indiv	indiv
algo	LR	SVM	SVM	LR	SVM	SVM	LR	SVM	LR	SVM
acc [%]	82	82	79	83	85	85	84	**85**	84	85

t		6.45	7.26	5.92	6.97	8.97	6.66	7.74	**9.02**	6.92	7.51

*Figure 5 Stable paths for the **relative centroid frequency** in D1. Abbreviations: seg len [s] = segment length in seconds, aggr = aggregation, algo = algorithm, acc [%] = accuracy in percent, t = t-value of tests against chance level, indiv = 10 individual segments, median = median of 10 segments, LR = Logistic Regression, SVM = Support Vector Machine. Note that normalization is omitted here because the biomarker is invariant to normalization.*

For D2, the two stable paths for **alpha peak frequency** both result in a performance of only 62% classification accuracy. Note that the calculation of this biomarker is also independent of normalization. The preprocessing pipelines with the stable paths both utilize segment lengths of 15s and median aggregation. The choice of classification algorithm yields slight differences in replicability (Logistic Regression: t=12.9; SVM: t=12.2).

Breakdown of relative centroid frequency

Since the **relative centroid frequency** stands out with its high and robust performance for both datasets, we subjected this biomarker to further investigations. The best performing path across both datasets for the **relative centroid frequency** is achieved with 10 individual 20s segments, therefore we chose this preprocessing pipeline for the analysis. In D1, the **relative centroid frequency** ranges below one consistently across electrodes for MDD patients, while it ranges consistently above one for HCs (Figure 6A). This means that the centroid of the alpha band is smaller than the centroid of the whole frequency spectrum in MDD patients and vice versa for HCs. **Relative centroid frequency** is composed of the **absolute centroid frequency** of the alpha band as well as the (absolute) centroid frequency of the broadband signal (here: 1 to 40Hz). Interestingly, the former does not perform well as a biomarker. Therefore, we investigated the latter (Figure 6B), which was not included in our original biomarker set because it is not a marker of the alpha band. This data suggests that the centroid frequency of the broadband signal might separate well between MDD and HC and that the discriminatory information is in the full signal rather than in the alpha band. However, these findings do not seem replicable with D2 (Figure 6C) due to a rather high spread of the MDD patients across the full band in D2 (Figure 6D).

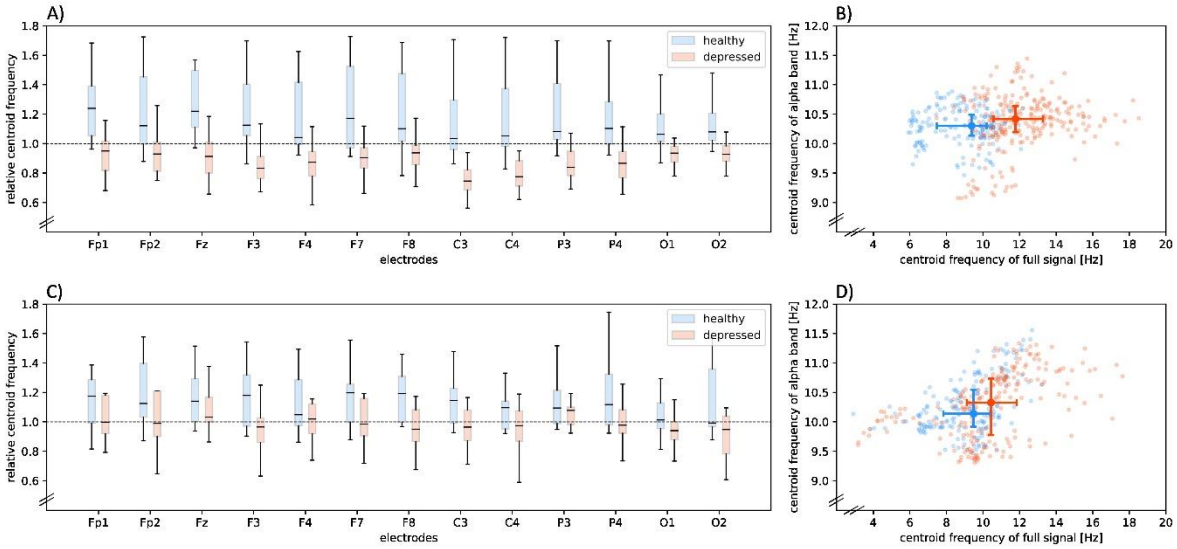

Figure 6 ***Relative centroid frequencies*** *of all electrodes for D1 (A) and D2 (C), respectively. The dotted line represents a relative centroid frequency of one, meaning that the centroid frequency of the alpha band corresponds to the centroid frequency of the full band. Relationship between the centroid frequency of the full signal (1 to 40Hz) and the centroid frequency of the alpha band (8 to 13Hz) for D1 (B) and D2 (D) for each subject and each channel separately. Data points are depicted as outliers when their values extend beyond 1.5 times the interquartile range from the first and third quartile, respectively. Error bars depict interquartile ranges in the x- and y-direction.*

Discussion

This study assesses several types of *reproducibility*[9] of alpha band EEG biomarkers for MDD diagnosis. It mainly demonstrates the *robustness* of several biomarkers by assessing their diagnostic performance across a multiverse of possible processing options as well as their *replicability* with two separate publicly available datasets. Using only biomarkers from the EEG alpha band, we achieve diagnostic accuracies that range from chance level up to 85% depending on the dataset, preprocessing variations, biomarker, and classification algorithm applied.

Replicability of biomarker research is important for generalizability and application in a clinical setting. The performances of both datasets were significantly impacted by the biomarker choice with the **relative centroid frequency** being the best performing and rather robust biomarker across datasets. Further *replication* was severely hampered by the classifiers on D2 performing on average only on chance level; a similar effect was observed on this dataset by other studies as well[11,29]. To assess the *replicability* within datasets, we used six-fold cross validation to receive a range of classification results instead of randomly drawing one single training- and test-set. These results were further used to substantiate our findings. Limited *replicability* across datasets despite the same analysis pipeline as we find in this study might have sources related or unrelated to the disease. The patients of both datasets have been diagnosed with a similar procedure, and both patient groups did not take anti-depressant medications two weeks prior to the study. However, the studies were conducted in different countries, the patients from D1 were on average about ten years older, and detailed symptoms cannot be compared due to the lack of detailed neuropsychological tests common to both datasets. Given the heterogeneity of MDD[10] and the small sample sizes, the variability of patients within and across datasets cannot be assessed properly and might have contributed to the limited *reproducibility*. Furthermore, there are differences in the recording setup, electrode placement, and electrode referencing. It has been shown that even electrodes from the same placement record different underlying brain regions between subjects[16], questioning the comparability of different electrode placements. Further studies are needed to systematically assess all these impact factors.

In order to understand the physiological underpinnings of a heterogeneous disease like MDD and find *reproducible* biomarkers, we need to know much better: Who are our subjects? The variability in symptoms and their expression needs to be assessed thoroughly and represented in the data. This holds for both patients and healthy controls, labels that are in fact too simple for capturing the complex physiology and phenology of a disorder such as MDD[46]. This endeavor necessitates much larger datasets to achieve some convergence of the findings. The better characterized the datasets are, the higher the chances that data can also be pooled across studies. This was unfortunately not possible in

this work despite harmonization of the datasets, a problem faced by other studies using these two datasets as well[29].

The vast variability in processing and analysis pipelines for neuroimaging data is another potential source for lack of *reproducibility*, which is addressed with meta- and multiverse-analyses, or multi-analyst studies[47]. With our multiverse analysis, we find an influence of some of the investigated processing steps, and observe that some of the biomarkers that perform on chance level on average have an acceptable performance and *robustness* in individual pipelines, e.g. the alpha **peak frequency**[48,49]. If such a pipeline was accidentally chosen in a traditional study only implementing one pipeline, this biomarker could be reported as a good EEG biomarker for MDD. Conversely, a biomarker with a bad performance in one pipeline might be disregarded for further analyses despite its potential, even though a specific processing choice might explain the bad performance. One example here is the impact of classification algorithms. Random forest performs better overall in D1 and is especially advantageous for some processing paths and biomarkers in both datasets. This indicates that the data space in these cases is not as well linearly separable since the other two algorithms are linear classifiers. Nonetheless, the most stable paths investigated here include the two linear classifiers. The impact of segment length has been shown to influence classification performance[26], a finding we partially *replicate* but not consistently. The tendency for better performance with longer segments suggests that the discriminatory features of MDD may be embedded in the course of time. Nonetheless, dividing recordings into sections of such length reduces the amount of data available for training; therefore, this is always a trade-off. Performing a multiverse analysis is advantageous for processing options where the best variation is not known and can strengthen the credibility of a finding if it is rather *robust* across the multiverse space. Conversely, the multiverse analysis can also be used to optimize the path for a given dataset and application. While this study provides the multiverse analysis across five processing steps, further steps such as artifact removal, feature extraction and selection as well as evaluation of the models offer room for further variations and need to be considered in future research.

The multiverse analysis of this study was restricted to alpha band biomarkers because of contradictions in literature regarding their performance[5]. The best performing while still *robust* biomarker in this study is the **relative centroid frequency**. Surprisingly, this biomarker is only represented very sparsely in the depression literature so far. In two studies it was included as putative biomarker and was either selected by one of four feature selection algorithms for the beta band and Fp2 electrode[24], or not reported whether it was included in the final feature set[50]. The **relative centroid frequency** is composed of the **absolute centroid frequency** and the centroid frequency of the broadband signal restricted from 1 to 40 Hz. Interestingly, the **absolute centroid frequency** performs substantially worse than the **relative centroid frequency** in our study, suggesting that the selectivity for MDD patients might be

driven by the total centroid frequency of the broadband. Our additional analyses demonstrate that this marker seems to separate the two clinical groups well in D1. Since we focused our analyses on the alpha band biomarkers, we did not pursue this further. However, though it is also the best performing biomarker for D2, we cannot *replicate* the good separation of clinical groups by the broadband centroid frequency in this dataset. A comparison of subfigures 6B and 6D may provide a potential explanation for this discrepancy: For full-band centroid frequencies higher than 6Hz, both datasets share the tendency of higher values in the MDD group compared to HCs. In contrast, we do not observe any subjects with a full-band centroid frequency below 6Hz in D1, whereas D2 features a distinct cluster of mainly MDD patients in this region. Considering that the recordings were taken in closed-eye resting-state, such remarkably low full-band centroid frequency values could plausibly be an indicator of drowsiness, or even sleep[5], impeding consistency between study conditions. Since there is no chance of confirming this hypothesis post-hoc, it remains purely speculative. But this ambiguity underlines again the vital importance of controlling and reporting study and recording conditions in a rigorous way as a fundamental prerequisite to ensure comparability of data across studies. To conclude, the **relative centroid frequency** of the alpha band as well as the centroid frequency of the broadband signal constitute interesting depression biomarkers that warrant further investigation.

A further biomarker that stands out is the **peak frequency** with outstanding stable paths in D2. However, we find its performance and *robustness* in D1 subpar. Alpha **peak frequency** is repeatedly discussed as a diagnostic biomarker for mental diseases[8,51,52], although there is evidence that there are substantial differences across individuals. Even though this biomarker is investigated intensively, findings are still conflicting[49]. We can replicate this conflict with our analyses but cannot resolve it since the differences arise mainly between datasets. All other individual biomarkers perform overall not sufficiently for consideration in a diagnostic scenario. Combining the individual markers into a **composite biomarker** constitutes a common method to construct a high-performance biomarker[24–26,39]. In this study, however, we do not find a distinctive advantage of the **composite biomarker**. It is noteworthy that this biomarker had 156 dimensions, which is out of proportion given the small sample size and a plausible explanation for the inferior performance[53]. Extensive feature selection, however, was out of the scope of this study given the complexity of the multiverse analysis.

The characteristics of the alpha band exhibit substantial inter-personal variations and are associated with age, neurological diseases, and memory performance[51]. A possible mitigation of these inter-individual differences might be to define alpha not by fixed frequency boundaries but rather by taking the individual alpha peak(s) into account[54]. Additionally, further biomarkers need to be included in future studies, searching for *reproducibly* high performing biomarkers for the support of depression diagnosis. Those biomarkers might be derived from other frequency bands, or come from completely

different categories such as biomarkers from nonlinear dynamics, connectivity features, or entropy characteristics to provide additional characterization of abnormal brain activity in MDD[5].

To conclude, this study demonstrates the large influence of choice of processing steps and their combinations. A multiverse approach in analyses of EEG resting-state data is therefore recommended, at least for the processing steps where an informed decision about a specific option cannot be made. Furthermore, we find one biomarker candidate that has shown a *robust* and *reproducible* high performance for MDD diagnostic support. This candidate, the **relative centroid frequency**, has previously not gained much attention in EEG research, yet is a marker that warrants further investigation. The restricted *replicability* of our findings with two datasets mirrors the inconsistencies in the field very well and highlights the necessity for large and well-curated EEG datasets for MDD research. Along this line, this study is based on rather small datasets, which renders our findings primarily a methodological showcase restricted in generalizability.

References

1. Scott KM, de Jonge P, Stein DJ, Kessler RC. Mental Disorders Around the World: Facts and Figures From the WHO World Mental Health Surveys. *AJP*. 2018;175(9):911-912. doi:10.1176/appi.ajp.2018.18050506

2. Moffitt TE, Caspi A, Taylor A, et al. How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective *versus* retrospective ascertainment. *Psychol Med*. 2010;40(6):899-909. doi:10.1017/S0033291709991036

3. Cai H, Han J, Chen Y, et al. A Pervasive Approach to EEG-Based Depression Detection. *Complexity*. 2018;2018:e5238028. doi:10.1155/2018/5238028

4. Abi-Dargham A, Moeller SJ, Ali F, et al. Candidate biomarkers in psychiatric disorders: state of the field. *World Psychiatry*. 2023;22(2):236-262. doi:10.1002/wps.21078

5. de Aguiar Neto FS, Rosa JLG. Depression biomarkers using non-invasive EEG: A review. *Neuroscience & Biobehavioral Reviews*. 2019;105:83-93. doi:10.1016/j.neubiorev.2019.07.021

6. Dev A, Roy N, Islam MdK, et al. Exploration of EEG-Based Depression Biomarkers Identification Techniques and Their Applications: A Systematic Review. *IEEE Access*. 2022;10:16756-16781. doi:10.1109/ACCESS.2022.3146711

7. Greco C, Matarazzo O, Cordasco G, Vinciarelli A, Callejas Z, Esposito A. Discriminative Power of EEG-Based Biomarkers in Major Depressive Disorder: A Systematic Review. *IEEE Access*. 2021;9:112850-112870. doi:10.1109/ACCESS.2021.3103047

8. Newson JJ, Thiagarajan TC. EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies. *Front Hum Neurosci*. 2019;12:521. doi:10.3389/fnhum.2018.00521

9. Botvinik-Nezer R, Wager TD. Reproducibility in Neuroimaging Analysis: Challenges and Solutions. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 2023;8(8):780-788. doi:10.1016/j.bpsc.2022.12.006

10. Marx W, Penninx BWJH, Solmi M, et al. Major depressive disorder. *Nat Rev Dis Primers*. 2023;9(1):44. doi:10.1038/s41572-023-00454-1

11. Kołodziej A, Magnuski M, Ruban A, Brzezicka A. No relationship between frontal alpha asymmetry and depressive disorders in a multiverse analysis of five studies. Shackman A, Baker CI, eds. *eLife*. 2021;10:e60595. doi:10.7554/eLife.60595

12. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365-376. doi:10.1038/nrn3475

13. Feng C, Thompson WK, Paulus MP. Effect sizes of associations between neuroimaging measures and affective symptoms: A meta-analysis. *Depression and Anxiety*. 2022;39(1):19-25. doi:10.1002/da.23215

14. Jasper HH. The 10/20 international electrode system. *EEG and Clinical Neurophysiology*. 1958;10(2):370-375.

15. Tucker DM. Spatial sampling of head electrical fields: the geodesic sensor net. *Electroencephalography and Clinical Neurophysiology*. 1993;87(3):154-163. doi:10.1016/0013-4694(93)90121-B

16. Scrivener CL, Reader AT. Variability of EEG electrode positions and their underlying brain regions: visualizing gel artifacts from a simultaneous EEG-fMRI dataset. *Brain and Behavior*. 2022;12(2):e2476. doi:10.1002/brb3.2476

17. Vincent KM, Xie W, Nelson CA. Using different methods for calculating frontal alpha asymmetry to study its development from infancy to 3 years of age in a large longitudinal sample. *Developmental Psychobiology*. 2021;63(6):e22163. doi:10.1002/dev.22163

18. Čukić M, López V, Pavón J. Classification of Depression Through Resting-State Electroencephalogram as a Novel Practice in Psychiatry: Review. *J Med Internet Res*. 2020;22(11):e19548. doi:10.2196/19548

19. Theissler A, Spinnato F, Schlegel U, Guidotti R. Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions. *IEEE Access*. 2022;10:100700-100724. doi:10.1109/ACCESS.2022.3207765

20. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. *Perspect Psychol Sci*. 2016;11(5):702-712. doi:10.1177/1745691616658637

21. Clayson PE, Baldwin SA, Rocha HA, Larson MJ. The data-processing multiverse of event-related potentials (ERPs): A roadmap for the optimization and standardization of ERP processing and reduction pipelines. *NeuroImage*. 2021;245:118712. doi:10.1016/j.neuroimage.2021.118712

22. Hosseinifard B, Moradi MH, Rostami R. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Computer Methods and Programs in Biomedicine*. 2013;109(3):339-345. doi:10.1016/j.cmpb.2012.10.008

23. Lee PF, Kan DPX, Croarkin P, Phang CK, Doruk D. Neurophysiological correlates of depressive symptoms in young adults: A quantitative EEG study. *Journal of Clinical Neuroscience*. 2018;47:315-322. doi:10.1016/j.jocn.2017.09.030

24. Cai H, Sha X, Han X, Wei S, Hu B. Pervasive EEG diagnosis of depression using Deep Belief Network with three-electrodes EEG collector. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. ; 2016:1239-1246. doi:10.1109/BIBM.2016.7822696

25. Mohammadi M, Al-Azab F, Raahemi B, et al. Data mining EEG signals in depression for their diagnostic value. *BMC Medical Informatics and Decision Making*. 2015;15(1):108. doi:10.1186/s12911-015-0227-6

26. Shen J, Zhao S, Yao Y, Wang Y, Feng L. A novel depression detection method based on pervasive EEG and EEG splitting criterion. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. ; 2017:1879-1886. doi:10.1109/BIBM.2017.8217946

27. Mahato S, Paul S. Electroencephalogram (EEG) Signal Analysis for Diagnosis of Major Depressive Disorder (MDD): A Review. In: Nath V, Mandal JK, eds. *Nanoelectronics, Circuits and Communication Systems*. Lecture Notes in Electrical Engineering. Springer; 2019:323-335. doi:10.1007/978-981-13-0776-8_30

28. Robbins KA, Touryan J, Mullen T, Kothe C, Bigdely-Shamlo N. How Sensitive Are EEG Results to Preprocessing Methods: A Benchmarking Study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 2020;28(5):1081-1090. doi:10.1109/TNSRE.2020.2980223

29. Savinov V, Sapunov V, Shusharina N, Botman S, Kamyshov G, Tynterova A. EEG-based depression classification using harmonized datasets. In: *2021 Third International Conference Neurotechnologies and Neurointerfaces (CNN)*. ; 2021:93-95. doi:10.1109/CNN53494.2021.9580293

30. Mumtaz W. MDD Patients and Healthy Controls EEG Data (New). Published online 2016:903228416 Bytes. doi:10.6084/M9.FIGSHARE.4244171.V2

31. Cai H, Gao Y, Sun S, et al. MODMA dataset: a Multi-modal Open Dataset for Mental-disorder Analysis. *Sci Data*. 2022;9(1):178. doi:10.1038/s41597-022-01211-x

32. Mumtaz W, Xia L, Mohd Yasin MA, Azhar Ali SS, Malik AS. A wavelet-based technique to predict treatment outcome for Major Depressive Disorder. Hu D, ed. *PLoS ONE*. 2017;12(2):e0171409. doi:10.1371/journal.pone.0171409

33. Mukhtar F, PS Oei T. Exploratory and confirmatory factor validation and psychometric properties of the Beck Depression Inventory for Malays (BDI-Malay) in Malaysia. *Malaysian Journal of Psychiatry*. 2008;17(1):58.

34. Yusoff N, Low WY, Yip CH. Psychometric properties of the Malay Version of the hospital anxiety and depression scale: a study of husbands of breast cancer patients in Kuala Lumpur, Malaysia. *Asian Pac J Cancer Prev*. 2011;12(4):915-917.

35. Gramfort A, Luessi M, Larson E, et al. MEG and EEG data analysis with MNE-Python. *Front Neurosci*. 2013;7. doi:10.3389/fnins.2013.00267

36. Gramfort A, Luessi M, Larson E, et al. MNE software for processing MEG and EEG data. *NeuroImage*. 2014;86:446-460. doi:10.1016/j.neuroimage.2013.10.027

37. Li A, Feitelberg J, Saini AP, Höchenberger R, Scheltienne M. MNE-ICALabel: Automatically annotating ICA components with ICLabel in Python. *Journal of Open Source Software*. 2022;7(76):4484. doi:10.21105/joss.04484

38. Yasin S, Hussain SA, Aslan S, Raza I, Muzammel M, Othmani A. EEG based Major Depressive disorder and Bipolar disorder detection using Neural Networks:A review. *Comput Methods Programs Biomed*. 2021;202:106007. doi:10.1016/j.cmpb.2021.106007

39. Poil SS, De Haan W, van der Flier W, Mansvelder H, Scheltens P, Linkenkaer-Hansen K. Integrative EEG biomarkers predict progression to Alzheimer's disease at the MCI stage. *Frontiers in Aging Neuroscience*. 2013;5. Accessed March 5, 2024. https://www.frontiersin.org/articles/10.3389/fnagi.2013.00058

40. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2

41. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2

42. Vallat R, Walker MP. An open-source, high-performance tool for automated sleep staging. Peyrache A, Büchel C, Bagur S, eds. *eLife*. 2021;10:e70092. doi:10.7554/eLife.70092

43. Angelakis E, Lubar JF, Stathopoulou S, Kounios J. Peak alpha frequency: an electroencephalographic measure of cognitive preparedness. *Clinical Neurophysiology*. 2004;115(4):887-897. doi:10.1016/j.clinph.2003.11.034

44. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(null):2825-2830.

45. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. In: ; 2010:92-96. doi:10.25080/Majora-92bf1922-011

46. Park SC, Kim YK. Challenges and Strategies for Current Classifications of Depressive Disorders: Proposal for Future Diagnostic Standards. In: Kim YK, ed. *Major Depressive Disorder: Rethinking and Understanding Recent Discoveries*. Springer; 2021:103-116. doi:10.1007/978-981-33-6044-0_7

47. Botvinik-Nezer R, Holzmeister F, Camerer CF, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. 2020;582(7810):84-88. doi:10.1038/s41586-020-2314-9

48. Voetterl HTS, Sack AT, Olbrich S, et al. Alpha peak frequency-based Brainmarker-I as a method to stratify to pharmacotherapy and brain stimulation treatments in depression. *Nat Mental Health*. 2023;1(12):1023-1032. doi:10.1038/s44220-023-00160-7

49. Zhou P, Wu Q, Zhan L, et al. Alpha peak activity in resting-state EEG is associated with depressive score. *Front Neurosci*. 2023;17. doi:10.3389/fnins.2023.1057908

50. Mohan R, Perumal S. Classification and Detection of Cognitive Disorders like Depression and Anxiety Utilizing Deep Convolutional Neural Network (CNN) Centered on EEG Signal: Traitement du Signal. *Traitement du Signal*. 2023;40(3):971-979. doi:10.18280/ts.400313

51. Klimesch W. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*. 1999;29(2):169-195. doi:10.1016/S0165-0173(98)00056-3

52. Lefebvre A, Delorme R, Delanoë C, et al. Alpha Waves as a Neuromarker of Autism Spectrum Disorder: The Challenge of Reproducibility and Heterogeneity. *Front Neurosci*. 2018;12. doi:10.3389/fnins.2018.00662

53. Altman N, Krzywinski M. The curse(s) of dimensionality. *Nature Methods*. 2018;15(6):399-400. doi:10.1038/s41592-018-0019-x

54. Olejarczyk E, Bogucki P, Sobieszek A. The EEG Split Alpha Peak: Phenomenological Origins and Methodological Aspects of Detection and Evaluation. *Front Neurosci*. 2017;11. doi:10.3389/fnins.2017.00506

Supplementary Material

Code for biomarker calculation: “calculate_biomarkers.ipynb” and “alpha_markers.py”

Raw accuracies of diagnostic classification models: “metrics.csv”