

Original Research Articles

Reproducibility of electroencephalography biomarkers for diagnosis of major depressive disorder

Yasmin Hollenbenders, MSc^{1,2} , Friedrich Ph. Carrle, MSc^{1,2} , Roman Mähler, MSc^{1,2} , Christoph Maier, PhD³, Alexandra Reichenbach, PhD^{1,2,3} ^a

¹ Center for Machine Learning, Heilbronn University, Heilbronn, Germany, ² Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany, ³ Medical Informatics, Heilbronn University, Heilbronn, Germany

Keywords: major depressive disorder, electroencephalography, biomarker, diagnosis, decision-support, classification, multiverse analysis, reproducibility, replicability, robustness

<https://doi.org/10.52294/001c.165175>

Aperture Neuro

Vol. 6, 2026

Major depressive disorder (MDD) and other psychiatric diseases can greatly benefit from objective decision support in diagnosis and therapy. Machine learning approaches based on biomarkers extracted from electroencephalography (EEG) have the potential to serve as low-cost decision support systems. Although this approach has shown promise, inconsistent findings regarding the diagnostic value of those biomarkers impede their clinical translation. Therefore, the *replicability* and *robustness* of these biomarkers need to be established first. We employ a multiverse analysis to systematically investigate the *robustness* against variations in six data processing steps, which may be one source of contradictory findings. These steps are artifact removal, normalization, time-series segment length, biomarker from the alpha band, aggregation, and classification algorithm. To ensure *replicability* of our findings, we analyze two publicly available EEG datasets with eyes-closed resting-state data containing 25/18 MDD patients and 23/14 healthy control subjects. The accuracies of diagnostic classifiers range from 81% to chance level, dependent on dataset and combination of processing steps. We observe that the selection and combination of processing steps significantly influences the results. Overall, the *replicability* of our findings across the two datasets was inconsistent. This study is a showcase for the advantages of employing a multiverse approach in EEG data analysis and advocates for larger, well-curated datasets 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, epidemiologic data suggests that a substantial number of cases is unreported, which might be due to current diagnostic practices, stigmatization, or public health policies.² In order to advance knowledge of the disorder, identify its biological underpinnings, improve diagnostic procedures, and develop more effective treatments, biomarkers are needed. Electroencephalography (EEG) constitutes a non-invasive and cost-efficient method to provide such biomarkers by recording brain physiology. Indeed, studies have shown that biomarkers extracted from EEG signals can distinguish between MDD patients and healthy

controls (HC).³⁻⁵ However, results contradict each other regarding discriminatory features extracted from these signals.^{3,5} The heterogeneity of the disorder as well as methodological differences across studies likely contribute substantially to these inconsistent findings,^{5,6} and complicate reproducibility.

Resolving these contradictions has increasingly moved into the focus of neuroscience research by promoting the importance of reproducibility. Botvinik-Nezer et al.⁷ distinguish three complementary dimensions of reproducibility, two of which are applied in this study. *Replicability* captures whether implementing the same analytical approach to different datasets yields comparable results. *Robustness* to analytical variability characterizes the stability of findings for the same data across different analytical approaches. Variability can, therefore, enter the data-processing pipeline

^a Corresponding Author:

Alexandra Reichenbach, PhD (alexandra.reichenbach@hs-heilbronn.de)
Center for Machine Learning, Heilbronn University
Heilbronn, Germany

through decisions made at multiple stages. Choices about population selection, study design, and data acquisition, including technical factors, can affect *replicability*, whereas decisions about preprocessing, feature extraction, and the final distinction between MDD and HC may challenge the *robustness* of findings.

Replicability is influenced by characteristics of data collections, including but not limited to participant selection, diagnostic assessment and labeling, symptom documentation, sample size, and electrode placement. These parameters are fixed once data have been recorded and, therefore, cannot be modified when working with publicly available open datasets. Nevertheless, *replicating* findings using public data or newly collected samples, in combination with the original processing pipeline, remains essential for corroborating findings.⁷

Once data is recorded, preprocessing steps such as filtering, artifact removal, normalization, segmenting, sub-sampling or augmentation, and data aggregation are often applied to the EEG signal. Although this sequence represents commonly used steps, the selection of steps and the specific methodological choices within each step can vary considerably, potentially leading to divergent findings and challenging *robustness*.⁸ Although artifact removal is a central step in EEG data preprocessing, the appropriate degree of intervention during this process is still a matter of considerable debate. Many studies rely on fully automated⁹ or (semi-)manual cleaning procedures,^{10,11} whereas other researchers argue for minimizing artifact removal to preserve more of the original neural activity.¹² Along the same line, normalization approaches range from omitting normalization entirely,⁹ to applying participant-wise normalization to harmonize datasets,¹³ or to implementing channel-wise normalization in order to stabilize input distributions for machine-learning (ML) models.¹⁴ Each strategy embodies distinct assumptions about inter-individual variability and model sensitivity. Furthermore, segmentation of EEG time-series is frequently employed in ML-based EEG analysis to increase the amount of training data. Here, the segment length can significantly influence model performance,¹⁵ highlighting the importance of carefully selecting temporal windows. Finally, while ML approaches generally benefit from larger datasets, statistical analyses often take the opposite route by aggregating data to robust estimators to reduce noise and emphasize stable patterns.

Alpha waves are associated with relaxation and predominantly observed when eyes are closed.¹⁶ Biomarker from the alpha band are as prominent in MDD research³ as they are subject to controversial discussion.¹⁷ While some studies suggest that biomarkers from the alpha band perform better than markers from other frequency bands,^{18,19} other studies found that they perform worse.^{15,20,21} Furthermore, it is disputed whether alpha activity is higher or lower in MDD participants compared to HCs.^{18,19} Yet, all these studies used different processing methods, rendering their results non-comparable.^{17,22} Even though the alpha band is one of the most consistently defined frequency bands, usually specified from 8 to 13 Hz, its definition can range anywhere within the borders of 6 to 14 Hz, introducing one

source of analytical variability.⁶ Additionally, biomarkers sharing the same name are calculated in different ways such as the frequently reported alpha asymmetry,²³ further hampering reproducibility.⁷ To ensure the reliability of findings regarding the neurophysiological underpinnings of MDD and their deviation to the neurophysiology of HC, objective biomarkers not only with high discriminatory power but also *replicability* across datasets, and to some degree of *robustness* to analytical variability are required.

The analytical approach for differentiating between patients and HCs represent another key variability across studies. These differences can be analyzed, e.g., with statistical comparisons, or with ML algorithms that constitute an increasingly popular multivariate approach. The latter are frequently applied to mimic a clinical decision-making process by modeling a classification problem with the class labels MDD and HC. The approaches range from classical supervised algorithms that usually operate on hand-crafted biomarkers²⁴ to more complex deep learning models that often, but not exclusively, operate on the raw time-series data or their frequency domain representation to inherently derive biomarkers in the training process.⁴

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.²⁴ One approach to reconcile the different findings and demonstrate *robustness* of the analysis of EEG data to the analytical variability is the multiverse analysis.²⁵ This approach has been adopted in applied neuroscience to reconcile contradictory findings about alpha asymmetry in MDD,²⁶ or to recommend an optimized processing pipeline.²⁷ This analysis strategy enables the demonstration that the high number of degrees of freedom in conducting EEG studies is one cause for the lack of reproducibility in findings.^{5,6}

The goal of this study is to assess two types of reproducibility on EEG alpha band biomarkers for the support of MDD diagnosis. In order to mimic a diagnostic scenario yet still retain close connection to studies using statistical analyses on those biomarkers, we adopt a classification approach with classical ML algorithms based on hand-crafted EEG markers. To keep the focus on the analytical variability, we limit our investigation to linear alpha markers, the most frequently discussed category of EEG biomarkers for MDD.³ For the purpose of *replication*, we utilize the two most frequently used publicly available MDD datasets²⁸ with eyes-closed resting-state EEG and harmonize them.¹³ To assess the *robustness* of the biomarkers to analytical variability, we systematically investigate the influence of several preprocessing steps, different features, and classification algorithms by means of a multiverse analysis. This study does not examine assumptions or predictions regarding specific biomarkers.

METHODS

DATA

Two publicly available datasets (D1²⁹; D2³⁰) with MDD patients (D1: n=29 (presumably age: 40.3±12.9 ; female: n=17),²⁹ D2: n=24 (age: 30.9±10.4; female: n=11)) and HCs (D1: n=28 (presumably age: 38.3±15.6; female: n=9),²⁹ D2: n=29 (age: 31.5±9.2; female: n=9)) were downloaded in August and October 2021, respectively. All information about D1 were derived from the original study describing the dataset,²⁹ because no accompanying information was provided with the dataset. The study designs were approved by the respective ethics committees (D1: Human ethics committee of the Hospital Universiti Sains Malaysia; 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 D2 did not take any anti-depression medication during the 2 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 system³⁰ with linked-ear-reference and 256 Hz sample frequency. D2 was recorded with 128 electrodes in the Geodesic Sensor Net³¹ with Cz as reference electrode and 250 Hz sample frequency.

DATASET HARMONIZATION

To establish a common ground for the analyses and limit preprocessing variability, we harmonized the datasets as far as possible using the Python Package *MNE*³² (version: 1.2.3). We matched the 128 electrodes from D2 to correspond to the 10-20 system according to the supplementary material provided with the dataset. Only data from the 13 overlapping electrode locations was included in the analysis: 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 common average, a prerequisite for the independent component analysis (ICA) in the cleaning step. 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 to include the filters originally applied to the two datasets.

MULTIVERSE ANALYSIS

We adapted a multiverse analysis²⁵ to investigate the impact of processing choices on the classification results. For each step (for analysis purposes also referred to as factor) of the processing pipeline, we implemented different options (also referred to as factor levels) and combined them fully. This resulted in a multiverse comprising 4752 parallel processing paths (Table 1) for each dataset. Note that the chosen processing steps and variations are only a selection of possible options and not an extensive list.

CLEANING

EEG data is susceptible to environmental, technical, and physiological interference.³³ Three approaches for artifact removal were carried out: 1) No artifact removal.¹² 2) Fully automatic removal of artifact components: EEG signals were separated into components with an ICA,³⁴ automatically labeled with *ICLabel*³⁵ (version: 0.4), and reconstructed without the components affected by artifacts, i.e., eye blinks, muscle artifacts, heartbeat, line noise, and channel noise. 3) Customized artifact removal with manual revision: An ICA was performed to identify common artifacts like eye and muscle movement as well as low-frequency anomalies and dataset specific technical artifacts. Identified artifact components were removed or attenuated using wavelet denoising. The pre-cleaned data was manually revised to exclude remaining high and low frequency artifacts by flagging affected time frames.

PREPROCESSING

Normalization is applied to EEG signals in order to level amplitudes and make participants more comparable.⁴ To compare the impact of normalization methods, we applied three variations of z-normalization: 1) No normalization. 2) Participant-wise normalization is applied to level putative differences across participants. For this, all channels of each participant are z-normalized jointly to align only the participants of the datasets with each other while retaining the between-channel differences for each participant. 3) Channel-wise normalization is frequently used to level putative differences across EEG traces. Each channel of each participant is z-normalized separately. Note that channel-wise normalization destroys spatial information based on absolute values of the power of the signal. Furthermore, biomarkers based on distribution, a certain frequency, or calculated relating to the full band are not affected by any of the normalization methods.

After normalizing, we split the data into non-overlapping segments of 20 s. All segments containing flagged artifacts were removed. This procedure was applied both to the cleaning path with manually revised data as well as to the paths without and with automatic artifact removal. This balances the comparability across datasets with a putative bias introduced by this approach. We selected the number of segments to maximize both the amount of usable data and the number of remaining participants, resulting in eight randomly subsampled segments per participant. Participants with less than eight remaining segments were excluded from subsequent analyses. This left 48 (MDD: 52%) / 32 (MDD: 56%) participants 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. Therefore, we additionally considered the first 5, 10, and 15 s of each segment for the analysis.

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]
Cleaning	Artifact removal	None; automatic; manual revision	3 [3]
Preprocessing	Normalization	None; participant-wise; channel-wise	3 [9]
	Butterworth filter	4 th order; 8 th order	2 [18]
	Segment length	5 s; 10 s; 15 s; 20 s	4 [72]
	Remove segments with artifacts (flagged by manual revision) and subsampling of the segments (for all paths)		
Feature Extraction	Biomarker	11 alpha band (8-13 Hz) Biomarker (cf. Table 2)	11 [792]
	Aggregation	None; median over 8 segments	2 [1584]
Classification	Algorithm	Support vector machine; logistic regression; random forest	3 [4752]
			4752 paths

The processing stage (first column) categorizes the processing steps (second column) for clarity. The third column depicts the variations used for each step. The fourth column indicates the number of variations and the number of combinations / paths of the multiverse at that point in the processing pipeline. The preprocessing steps are described in detail in the Multiverse Analysis section of the methods.

FEATURE EXTRACTION

To compile a comprehensive set of alpha band biomarkers, we calculated eleven spectral biomarkers common for diagnosis of psychiatric disorders (Table 2).⁵ We limited the analysis to biomarkers that can be calculated individually for each electrode resp. EEG channel to avoid combinatoric explosion of the number of possible features. The frequency range was set to 8 to 13 Hz, the typical range of the alpha band.⁶ 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 2.05 s window duration and a spectral resolution of 0.5 Hz. To calculate the envelope of the signal, we first extracted the alpha band with a Butterworth bandpass filter of either 4th or 8th order. The envelope biomarkers were calculated based on the envelope of the time-domain signal,³⁶ which was calculated using the absolute values of the Hilbert transform. The actual biomarkers characterize the distributions of the envelope data points. Because the calculation of some of the biomarkers is inconsistent across studies or not properly described in the literature, we compiled a detailed overview of their construction to enable analytical reproducibility (Table 2). Biomarkers were calculated using the Python toolboxes *scipy*³⁷ (version: 1.9.3), *numpy*³⁸ (version: 1.23.5), and the function *bandpower* adopted from Vallat and Walker.³⁹ The accompanying source code for biomarker calculation further ensures analytical reproducibility (see Supplementary Materials "python_code_for_biomarker_calculation.pdf").

Biomarker values were calculated individually for each time segment and EEG channel. Since there are eight segments per participant, one participant is represented with eight different values, which effectively corresponds to a data augmentation process commonly encountered in the field of ML. Alternatively, we pursued an aggregation strategy by taking the median over each participant's values, resulting in one single and putatively more robust biomarker

value per participant and EEG channel. The latter 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 well suited for smaller datasets: logistic regression (maximum 200 iterations), support vector machine (SVM) with a linear kernel (default values), and random forest (maximum depth 2). Models were trained for each individual biomarker. For each model, the input was a 13-dimensional feature vector containing the biomarker value for each EEG channel. The class labels were the diagnostic groups MDD and HC. The models were trained with six-fold cross-validation stratified by diagnosis on a per-participant split basis,¹³ resulting in five to eight participants per test dataset. To counteract imbalance of the class distribution, the balanced mode was used. Note that the partitioning in training and test dataset was kept identical for all paths of the multiverse analysis to ascertain comparability. Balanced accuracy of each model was used for the analysis. We used *sklearn*⁴¹ (version: 1.1.3) for all classification-related implementations.

ANALYTICAL APPROACH

In this study, we assess two types of reproducibility,⁷ both from different perspectives. The *replicability* was assessed between as well as within datasets. *Replication* between datasets is assessed by using two publicly available datasets and observing convergences and divergences between those across all further *robustness* analyses. *Replication* within datasets is analyzed by using six-fold cross validation and comparing results between folds. *Robustness* is assessed for both datasets regarding processing steps and biomarkers. To assess the *robustness* of the classification performance against variations in the processing steps, we conducted analyses of variance (ANOVAs) including the factors of the

Table 2. Biomarkers used in this study, their description, calculation, and sources/studies where those biomarkers were used for diagnosis or characterization.

Biomarker	Description	Used In
Absolute band power	Total power of the alpha band Welch periodogram, and integration using Simpson's rule (adopted from ³⁹)	15, 18-21
Relative band power*	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 ³⁹)	20
Absolute centroid frequency*	Center of mass of the alpha band power spectrum Welch periodogram, and calculation of the power-weighted sum of the frequencies of the alpha band	15,20
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)	20
Peak frequency*	The frequency of the alpha band spectral component with the highest power Welch periodogram, select frequency with maximum power ⁴⁰	15,20, 36
Envelope kurtosis*	Butterworth bandpass filter, Hilbert transform, and calculating the kurtosis of the absolute values of the transformed signal	36, on raw signal: 20
Envelope skewness*	Butterworth bandpass filter, Hilbert transform, and calculating the skewness of the absolute values of the transformed signal	36, on raw signal: 20
Envelope median	Butterworth bandpass filter, Hilbert transform, and calculating the median of the absolute values of the transformed signal	36
Envelope interquartile range	Butterworth bandpass filter, Hilbert transform, and calculating the interquartile range between the 25 th and the 75 th quartile of the absolute values of the transformed signal	36
Envelope variance	Butterworth bandpass filter, Hilbert transform, and calculating the variance of the absolute values of the transformed signal	36, on raw signal: 20
Envelope range	Butterworth bandpass filter, Hilbert transform, and calculating the range of the absolute values of the transformed signal	36

* Indicates biomarkers that are invariant to normalization.

multiverse as within-participant factors and subsequent ANOVAs or *t*-tests for post hoc analyses whenever appropriate. We categorized the ANOVAs into two groups of biomarkers: those affected by normalization (A-BM: **envelope median, range, variance, interquartile range** and the **absolute band power**) and those unaffected by normalization (U-BM: **absolute** and **relative centroid frequency, envelope skewness** and **kurtosis, relative band power, and peak frequency**). ANOVAs for U-BMs are conducted without the normalization factor. The *robustness* of the individual biomarkers was further assessed with three approaches: 1) Analysis of the performance range of each path for each biomarker. 2) One-tailed *t*-tests against chance level (i.e., 50% accuracy) to test each path for above chance level classification performance. 3) Correlation analysis between paths with biomarker as variables to analyze biomarker patterns across paths.

Additional analyses include Fischer's exact test to compare groups with nominal scaled data. ANOVAs were conducted with *statsmodel*⁴² *AnovaRM* (version: 0.14.0), post-

hoc tests and *t*-tests against chance level with *pinguoin*⁴³ *ttest* (version: 0.5.5). Reported values represent mean $\pm$ standard deviation unless stated otherwise. Statistical results are not corrected; instead, we report multiple alpha thresholds to facilitate comparison to other studies.

GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

Copilot (Microsoft, Redmond, Washington, USA, version: 2.20260326.20.0) was used selectively to enhance the clarity and precision of the wording in the manuscript.

RESULTS

BUTTERWORTH FILTER ORDER DOES NOT AFFECT CLASSIFICATION PERFORMANCE

Since the Butterworth filter was only used to calculate the envelope biomarkers and the filter order has no effect on their classification results (all envelope biomarkers; D1: $p >$

.88; D2: $p > .35$) (Figure S1), we limit further analyses to the 8th filter order. This reduces the multiverse to 2376 paths.

CLASSIFICATION RESULTS ARE NOT REPLICABLE ACROSS DATASETS

The overall classification performance varied significantly between datasets ($t_{10} = 4.760$; $p < .001$). Classifiers operating on D1 achieve a mean accuracy of $59.4 \pm 9.4\%$ across all paths with a wide performance range from 35.4% to 87.5% mean accuracy for the individual paths. The overall performance of classifiers operating on D2 is at chance level with $49.3 \pm 8.8\%$ with mean accuracies for the individual paths ranging from 18.9% to 87.2%. Note that some statistical effects for D2 might not be meaningful when the accuracies perform on average around chance level.

ARTIFACT REMOVAL IMPACTS CLASSIFICATION PERFORMANCE

During manual revision of the EEG signals, we found technical artifacts (sinusoidal artifacts at ~5 Hz across channels (Figure S2 and Figure S3) and exceptionally high spikes (Figure S4 and Figure S5) in D1 that were not removed by the automatic artifact removal. Importantly, those technical artifacts were unequally distributed across groups (before participant removal: MDD: 1 (3.4%) / HC: 12 (42.8%); $p < .001$; after participant removal: MDD: 1 (4.0%) / HC: 8 (34.8%); $p = .009$). For D1, the artifact removal step has a significant impact only on the overall performance of the classifiers operating on the U-BMs ($F_{2,10} = 4.201$, $p = .047$). However, biomarker performance of both groups is differentially affected by it (interaction artifact removal x biomarker, Figure 1A; U-BM: $F_{10,50} = 10.969$, $p < .001$; A-BM: $F_{8,40} = 2.423$, $p = .031$). Analysis of D2 neither reveals an impact of the artifact removal step per se (U-BM and A-BM: $F_{2,10} < 3.006$, $p > .095$), yet biomarkers are differentially modulated by the chosen method to remove artifacts (interaction artifact removal x biomarker, Figure 1B: U-BMs: $F_{10,50} = 3.295$, $p = .002$; A-BMs: $F_{8,40} = 2.892$, $p = .012$). Because of the bias due to the technical artifacts in D1, the remaining analysis includes only the results of the classifiers operating on the manually revised data, leaving 792 paths.

IMPACT OF PROCESSING STEPS ON CLASSIFICATION PERFORMANCE

First, we assess whether the *robustness* of the classification results against the variations of the individual processing steps is *replicable* across datasets. The A-BMs of both datasets are significantly affected by normalization (Figure 2A, D1 (black squares): $F_{2,10} = 6.529$, $p = .015$; D2 (grey squares): $F_{2,10} = 4.534$, $p = .040$). Additionally, the performance of the individual biomarkers is differentially modulated by normalization (Figure 3A, D1: $F_{8,40} = 3.779$, $p = .002$; Figure 3B, D2: $F_{8,40} = 3.210$, $p = .006$). These main and interaction effects are the only one that both datasets have. However, the pattern of influence differs across datasets. The only other commonality across datasets is that the pro-

cessing step aggregation affects classification accuracy for neither dataset nor biomarker group (Figure 2C).

For D1, the performance of classifiers with U-BMs were not affected by any processing steps (all factors $p > .06$). The classification performance of the A-BMs is significantly affected by segment length (Figure 2B, black square: $F_{3,15} = 3.522$, $p = .041$), and algorithm (Figure 2E, black square: $F_{2,10} = 4.606$, $p = .038$).

D2 is more severely affected by processing steps. The performance of both biomarker groups is significantly affected by biomarker (Figure 2D, U-BMs (grey circle): $F_{5,25} = 3.371$, $p = .018$; A-BMs (grey squares): $F_{4,20} = 13.102$, $p < .001$). Furthermore, the performance of the individual biomarkers is differentially modulated by algorithm (U-BMs: $F_{10,50} = 2.777$, $p = .008$; A-BMs: $F_{8,40} = 4.836$, $p < .001$). The performance of the U-BMs significantly depends on segment length (Figure 2B, grey circle: $F_{3,15} = 5.567$, $p = .009$), and algorithm (Figure 2E, grey circle: $F_{2,10} = 9.947$, $p = .004$). The performance of the individual A-BMs is differentially modulated by segment length ($F_{12,60} = 9.427$, $p < .001$). The interaction effects between normalization x algorithm and segment length x aggregation are also unique for the A-BMs of D2 (both interactions $p < .003$).

BIOMARKERS ARE DIFFERENTIALLY MODULATED BY PREPROCESSING STEPS

To examine more closely how individual processing steps affect each biomarker, we conducted separate ANOVAs for all biomarkers. Three A-BMs are in both datasets influenced by the same set of processing steps. However, each of them is additionally affected differently by one processing step that differs between datasets. The **absolute band power** is affected by segment length (D1: $F_{3,15} = 3.355$, $p = 0.047$; D2: $F_{3,15} = 4.130$, $p = .026$) and additionally in D1 affected by normalization (D1: $F_{2,10} = 6.236$, $p = 0.017$). **Envelope range** is affected by segment length (D1: $F_{3,15} = 6.057$, $p = 0.007$; D2: $F_{3,15} = 12.772$, $p < .001$) and algorithm (D1: $F_{2,10} = 4.178$, $p = 0.048$; D2: $F_{2,10} = 5.714$, $p = .022$), and additionally in D2 affected by normalization (D2: $F_{2,10} = 4.172$, $p = .048$). The last one of this group, **envelope interquartile range**, is affected by normalization (D1: $F_{2,10} = 5.029$, $p = 0.031$; D2: $F_{2,10} = 14.171$, $p = .001$) and additionally in D2 affected by algorithm (D2: $F_{2,10} = 7.833$, $p = .009$). Furthermore, the three U-BMs **relative centroid frequency**, **relative band power**, and **absolute centroid frequency** are not affected by any processing steps in both datasets (all factors, D1: $p > .08$; D2: $p > .11$).

The remaining five biomarkers were differently affected in D1 and D2. **Envelope median** (A-BM) is affected by normalization in D1 ($F_{2,10} = 6.610$, $p = .015$), and by segment length in D2 ($F_{3,15} = 4.782$, $p = .016$). **Envelope variance** (A-BM) is also affected by normalization in D1 ($F_{2,10} = 6.567$, $p = 0.015$) but not affected at all in D2 (all factors, $p > .09$). This pattern is reversed in the remaining three U-BMs, which are not affected in D1 (all factors, $p > .05$). Yet in D2, **envelope skewness** and **envelope kurtosis** are affected by algorithm ($F_{2,10} = 14.262$, $p = .001$; $F_{2,10} = 6.744$, $p = .014$; respectively), and **peak frequency** is affected by segment length ($F_{3,15} = 4.016$, $p = .028$).

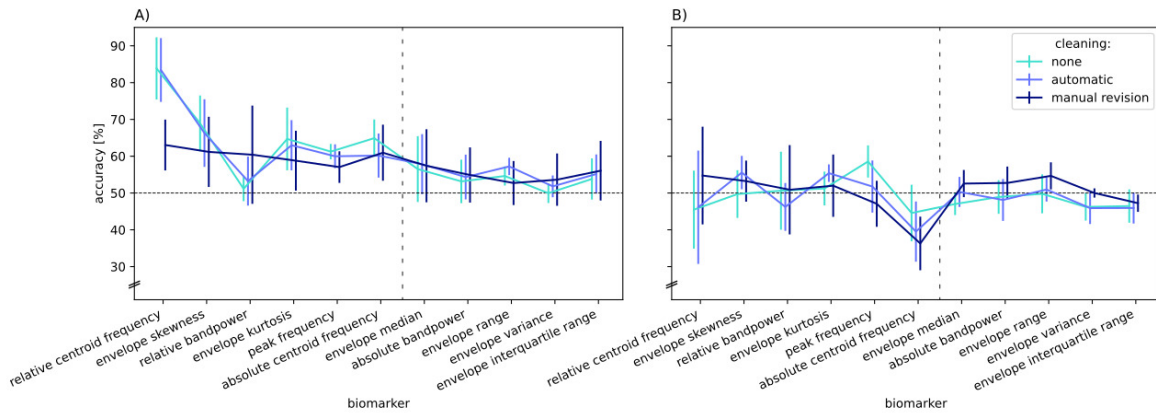


Figure 1. Interaction effects of artifact removal x biomarker on classification performance (balanced accuracy) for A) D1 and B) D2 for 8th Butterworth filter order. Values are averaged across factors for each fold. Error bars depict standard deviations across folds. Biomarkers are grouped by U-BMs (left of the vertical dotted line) and A-BMs (right) and then sorted according to their mean performance across datasets. Abbreviations: D1/2 = dataset 1/2, U-BMs = biomarkers unaffected by normalization, A-BMs = biomarkers affected by normalization.

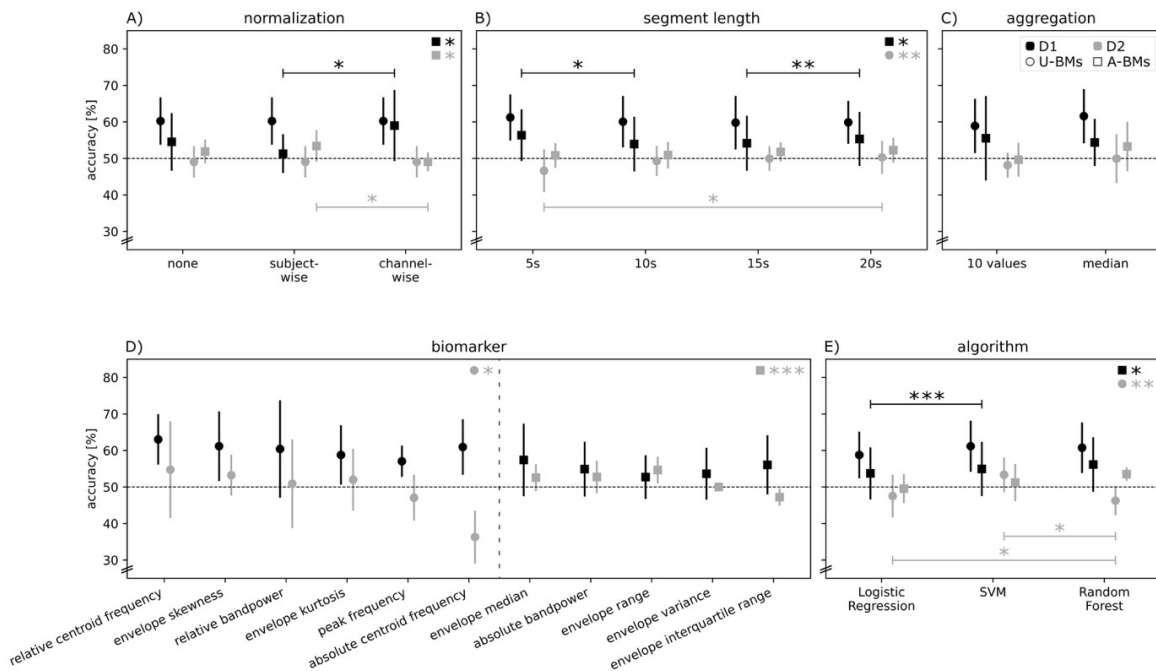


Figure 2. Effects of preprocessing steps, biomarkers, and classification algorithms on classification performance (balanced accuracy) for D1 (black) and D2 (grey) for Butterworth filter order of 8 and manually revised data. Values are averaged across factors for each fold. Error bars depict standard deviations across folds. Main effects for each dataset and biomarker group separate are depicted in the top right corner. Horizontal bars depict post hoc tests. Stars depict an effect of the analyses with * $p < .05$, ** $p < .01$, *** $p < .001$, all uncorrected. Biomarkers are grouped by U-BMs (circle markers) and A-BMs (square markers) and then 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: D1/2 = dataset 1/2, SVM = Support Vector Machine, U-BMs = biomarkers unaffected by normalization, A-BMs: biomarkers affected by normalization.

MOST BIOMARKERS ARE NOT ROBUST AGAINST CHANCE LEVEL

The smaller the impact of the processing steps on the diagnostic accuracy of a biomarker, the more *robust* it is to analytical variability (smaller range in [Figure 4A](#)). The most

robust biomarkers vary between datasets with **envelope kurtosis** (14.6%) for D1 and **relative centroid frequency** (13.9%) for D2. The least *robust* biomarker is the **envelope range** with 27.6% in D1 and 43.9% in D2.

The *t*-statistic combines the performance of a classifier with its *internal replicability*, acknowledging the stochastic

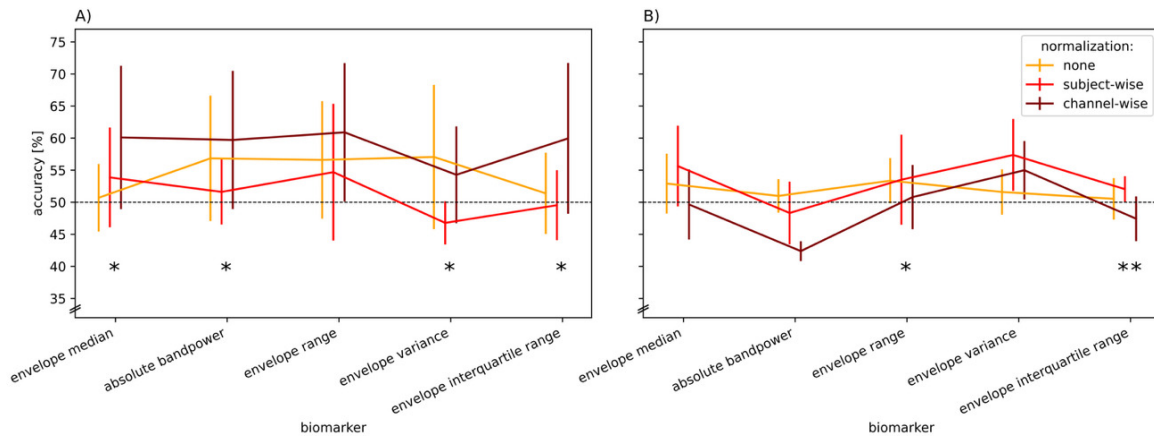


Figure 3. Interaction effects of normalization x biomarker on classification performance (balanced accuracy) for A) D1 and B) D2 for 8th Butterworth filter order and manually revised data only for A-BMs. Values are averaged across factors for each fold. Error bars depict standard deviations across folds. Stars depict an effect of the post-hoc ANOVA for the corresponding biomarker with * $p < .05$, ** $p < .01$, *** $p < .001$, all uncorrected. Abbreviations: D1/2 = dataset 1/2, A-BMs = biomarkers affected by normalization

nature in the data sampling process (Figure 4B). While inference testing is standard in medicine and neuroscience, biomarker studies from the ML field tend to neglect *replicability* in favor of performance.⁴⁴ Examining the *robustness* of biomarkers within paths from the perspective of a quasi-meta-analysis, we find that only 19.8% of all diagnostic classifiers, i.e., 19.8% of all paths across biomarkers and datasets, perform significantly better than chance level on our most lenient alpha level of 0.05 (Figure 4B lowest dashed line). Along this line, there is no biomarker that surpasses this threshold on all paths. Furthermore, only one biomarker, the **relative centroid frequency** in D1, contains one path (segment length: 15 s, aggregation: median, algorithm: SVM) that performs better than our most conservative alpha level (Figure 4B, highest dashed line: $p < .0001$). Notably, just four biomarkers in D1 (**relative centroid frequency**, **envelope skewness**, **envelope kurtosis**, **absolute centroid frequency**) obtain only positive t -values, meaning that all their paths perform above chance level. Another interesting behavior show biomarkers like the **envelope kurtosis** in D2, where 83.3% of all paths perform below our most lenient alpha level of 0.05 yet one path (segment length: 20 s; aggregation: individual values, algorithm: random forest) exceeds our second most conservative level of 0.001.

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 5). The lowest median of the standard deviations is 10.8% for **peak frequency** in D1 and 10.6% for **envelope kurtosis** in D2. The highest variability we find for the **relative band power** in D1 (median of std = 19.4%) and **absolute centroid frequency** in D2 (median of std = 16.8%). However, these variations need to be considered with caution in this study since each test set for cross-validation only contained five to eight participants, i.e., the resolution for the accuracy ranges between 12.5% and 20%.

To summarize, we observed four different kinds of performance patterns in the analyses. Mostly *robust* to analytical variability: The biomarker performs significantly above chance-level for most paths (Figure 4B, e.g., D1 **relative centroid frequency**). Lack of *replicability* across datasets: The biomarker performs above chance level in one dataset and below in the other (Figure 4A, e.g., **absolute centroid frequency** ($t_{71} = 31.973$, $p < .001$)). Limited *replicability* within datasets: The performance within a dataset and path shows high variance across folds (Figure 5, e.g., D2 **envelope kurtosis**). Not *robust* to analytical variability: The biomarker performance diverges widely within a dataset depending on the chosen processing path (Figure 4A, e.g., D2 **envelope range**).

BIOMARKER PATTERNS

Another perspective on the impact of the processing steps on the biomarkers is presented by a correlation analysis between the different paths. The focus here is on the biomarker patterns – that is, whether the biomarkers exhibit similar or divergent performance relative to one another.

The correlations are in general higher within (Figure 6 and Figure S8) than across datasets (Figure S9). Channel-wise normalization of D2 seems to align the classification patterns to the ones of D1. Furthermore, 5 s segments and channel-wise normalization (in D1 as well as in D2) increase correlation between datasets.

For D1, normalization variants exert distinct effects on the resulting biomarker patterns (Figure 6). Paths employing participant-wise normalization exhibit high correlations without marked differences across the other factors. Paths without normalization also show a high degree of internal consistency, except for those combining median as aggregation step with random forest as algorithm. In these cases, the biomarker patterns diverge from the other non-normalized paths. Under channel-wise normalization, we observe that paths using random forest display bio-

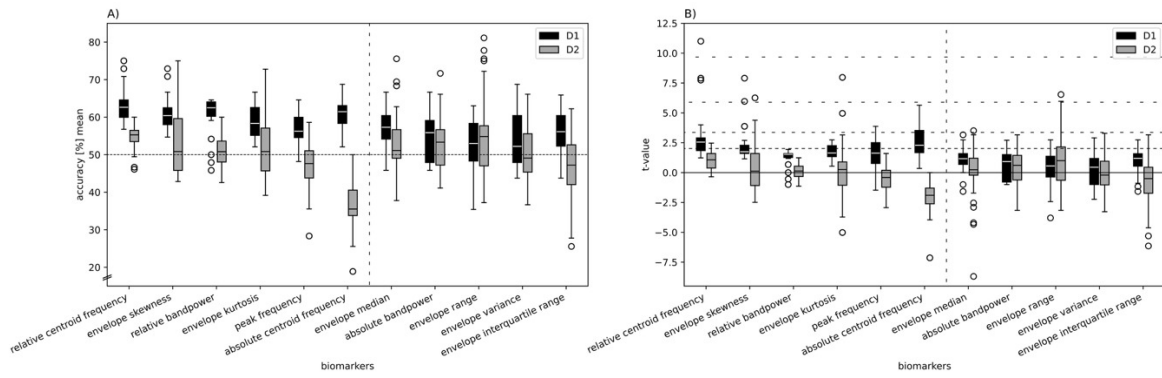


Figure 4. Robustness to analytical variability. A) Classification accuracies (balanced accuracy) across paths ($n = 72$) with Butterworth filter order of 8 and manually revised data. Mean accuracies for each path are averaged across folds. Horizontal dotted line depicts chance level. Vertical dotted line separates U-BMs (left) from A-BMs (right). B) t-statistics for each path of each biomarker ($n = 72$) with Butterworth filter order of 8 and manually revised data. Horizontal dashed lines depict varying degrees of alpha levels from 0.05 (lowest and narrowest dashed line), 0.01, 0.001, to 0.0001 (highest and widest dashed line). Data points are depicted as outliers when their values extend beyond 1.5 times the interquartile range from the first and third quartile, respectively. Abbreviations: D1/2 = dataset 1/2, U-BMs = biomarkers unaffected by normalization, A-BMs: biomarkers affected by normalization.

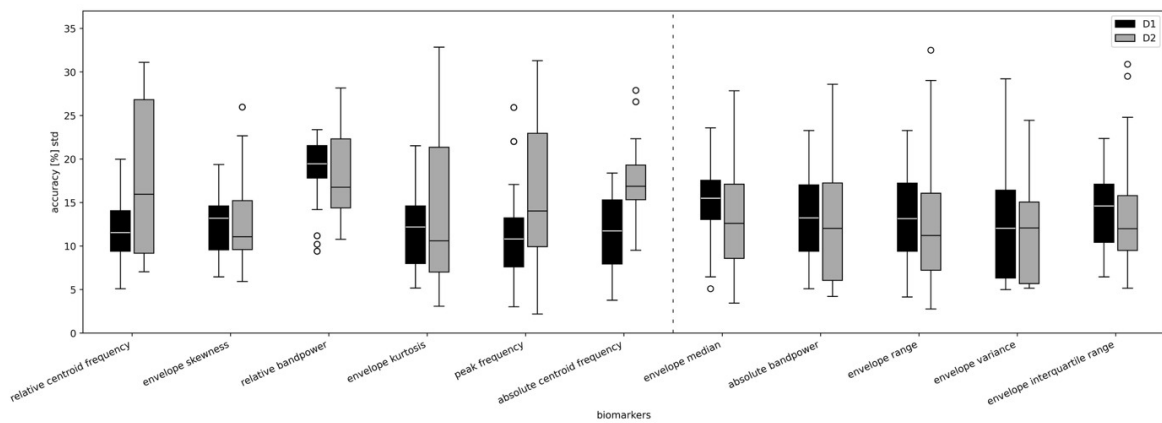


Figure 5. Replicability within datasets. Variability of classification accuracies (balanced accuracy) across folds within each path ($n = 72$) with Butterworth filter order of 8 and manually revised data. Standard deviations across the six folds are shown for each path. Vertical dotted line separates U-BMs (left) from A-BMs (right). Data points are depicted as outliers when their values extend beyond 1.5 times the interquartile range from the first and third quartile, respectively. Abbreviations: D1/2 = dataset 1/2, U-BMs = biomarkers unaffected by normalization, A-BMs: biomarkers affected by normalization.

marker patterns that are less consistent with the remaining pipelines than those using alternative algorithms.

The raw values of the classification results can be found in the Supplementary Material (“metrics.csv”) for analytical reproducibility.

DISCUSSION

This study assesses two types of reproducibility⁷ across processing pipelines for alpha band EEG biomarkers used in a diagnostic task for MDD detection. It evaluates the *robustness* to analytical variability of those biomarkers by assessing their diagnostic performance across a multiverse of possible processing pipelines, as well as their *replicability* with two separate publicly available datasets. Including

only linear biomarkers from the EEG alpha band, we achieve diagnostic accuracies that range from 81% down to chance level depending on the dataset, preprocessing variations, biomarkers, and classification algorithm applied. Even though the datasets used in this study are rather small, they are the most widely analyzed publicly available datasets for EEG biomarker research in MDD.²⁸ More than a third of studies in this field in the last 5 years base their findings on these datasets, more than 75% of the studies based on public datasets use these data and from those studies, more than 75% use them as standalone datasets without further data to *replicate* their findings.

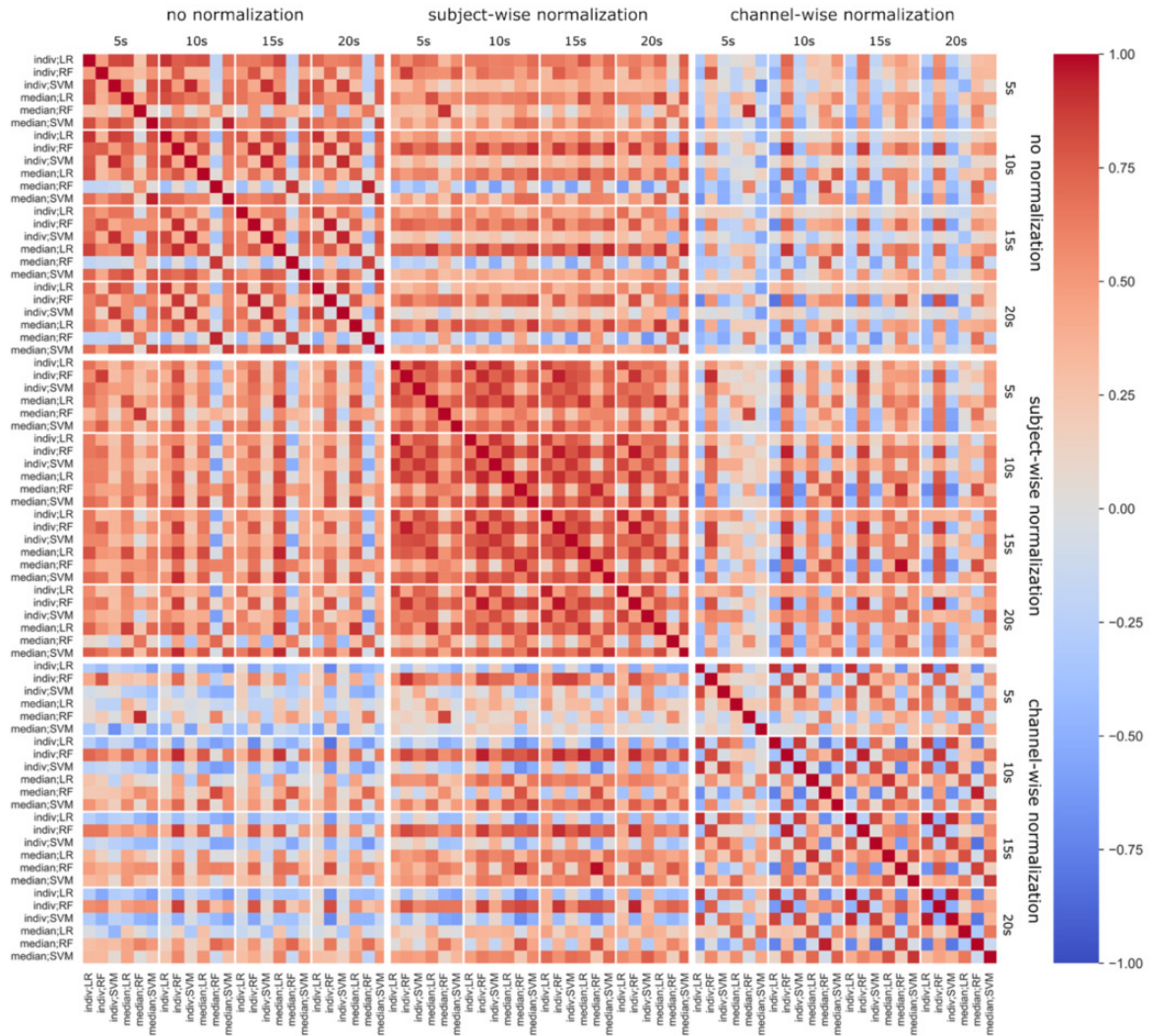


Figure 6. Correlation of classification performance (balanced accuracy) of paths in D1 with Butterworth filter order of 8 and manually revised data (n=72). Path names are constructed as follows: aggregation; algorithm. Big squares depict normalization variants; small squares depict various segment lengths. Abbreviations: LR = Logistic Regression, RF = Random Forest, SVM = Support Vector Machine.

REPLICATION OF FINDINGS WITH TWO PUBLICLY AVAILABLE DATASETS

Replicability of biomarker research is important for generalizability and application in a clinical setting. The *replication* of findings across datasets was largely not possible in this study despite their harmonization. It was severely hampered by the classifiers trained on D2 performing on average only on chance level. Similar findings have been reported before when working with these datasets; D2 shows either a behavior contrary to other datasets analyzed²⁶ or a performance around chance level for approaches that work with other datasets.¹³ Limited *replicability* across datasets despite the same analysis pipeline might have sources related to the heterogeneity of the disease, differences in the study design, or in the data acquisition.

A multifaceted disorder such as MDD is heterogeneous in terms of severity and symptoms, progression, (drug) treatment (success), and the presence of comorbidities.⁴⁵ Sample sizes of overall 30 to 50 participants⁶ are very likely too small to capture the variability observed in MDD. Furthermore, the label MDD is assigned on various criteria, including clinical diagnoses based on the DSM-IV manual or by sum-scores of questionnaire-based assessments,²⁶ which cannot reflect the prevailing heterogeneity of the condition. This holds for both patients and HCs, labels that are in fact too simple for capturing the complex physiology and phenology of a disorder such as MDD.⁴⁶ As a consequence, the effect sizes of studies are often inflated,^{7,47, 48} leading to overestimation of the selectivity of the biomarkers studied. The patients of both datasets in this study have been diagnosed based on DSM-IV criteria but cannot be compared with regard to other clinical factors since in-

formation about them is missing for one or both datasets. Given the small sample sizes in this study, patient variability both within and across datasets may not have been adequately captured, likely contributing to the limited *replicability*. In order to understand the physiological underpinnings of a heterogeneous disease like MDD and find reproducible biomarkers, we need to characterize participants much better. The variability in symptoms and their expression as well as comorbidities need to be assessed thoroughly and represented in the data.¹⁰ The better characterized the datasets are, the higher the chances that data can also be pooled across studies. Complementary, dedicated studies to systematically assess these impacting factors are needed. Along the same line, the study populations need to be well characterized with additional information, e.g., demographic data or other factors about the current state of the participants. The datasets for this study were collected in different countries and the patients from D1 were on average about 10 years older. These differences might have further contributed to the lack of *reproducibility*.

Comparability of studies may further be hampered by the lack of standardization in the recording of EEG data. There are differences in the recording setup, electrode placement, and electrode referencing. Technical differences like the electrode placements, e.g., the quasi-standard of the EEG-10-20/10/5 system³⁰ or the Geodesic Sensor Net configuration³¹ might be other sources of variability. It has been shown that even electrodes from the same landmark-standardized placements record different underlying brain regions between participants,⁴⁹ questioning the comparability of participants as well as different electrode placements. More variability in technical recording properties comes from the choice of the electrode type, sampling frequency, and acquisition hardware. Moreover, settings like the time of the day, temperature of the room, or resting-state conditions can introduce further variability. Obtaining robust neurophysiological representations^{50,51} across all these variations requires data sharing, and collaborative efforts are sorely needed.⁷

Replicability cannot just be assessed across datasets, but also within. 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. This revealed high fluctuations of classification accuracy even within datasets, confirming the notion of high heterogeneity within the groups MDD and HC. Yet, these findings are limited by the small sample size leading to 5–8 participants per test group, which means 2–4 participants per group (MDD/HC). One falsely classified participant can thus lead to an accuracy drop of 12.5% to 20%. These measures of variability should therefore not be over-interpreted and serve merely as an illustration what kind of statements would be possible with this approach.

ROBUSTNESS TO ANALYTICAL VARIABILITY

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 mul-

tivariate-analyses, or multi-analyst studies.⁸ Performing a multivariate analysis is advantageous when multiple equally valid processing options are available.²⁵ If a finding is rather *robust* across the multiverse space this can strengthen the credibility of this finding. If it is not, it might resolve discrepant previous findings. With our multivariate analysis we find a marked influence of some of the investigated processing steps.

The most notable effect on the classification was introduced by data cleaning. Manual revision of D1 significantly reduced the performance of some biomarkers, especially of the group unaffected by normalization (U-BM). Insights from the revision process reveal technical artifacts (spikes and 5 Hz-noise) in this dataset that are not captured by automatic ICA-based artifact detection algorithms like ICLabel and that are unequally distributed across diagnostic groups. It is therefore recommended to always review the data after automatic artifact removal. Even though D1 is frequently used,²⁸ we did not find a study that commented on this problem. From the studies included in a recent review,²⁸ nine of them report methods not suitable for removing these specific artifacts, and three of them did not mention those type of artifact despite performing manual artifact removal. Although we removed those artifacts, our approach might introduce a bias since we excluded segments with artifacts in all paths, solely examined through the manual revision. With this approach we, therefore, removed detected artifacts also in paths where these artifacts would not have been removed.

Normalization has a significant influence in both datasets, yet with opposing effects. Interestingly, in D1 channel-wise normalization outperforms participant-wise normalization, even though this destroys the spatial patterns across electrode locations. Given the performance of D2 around chance level, we refrain from interpreting effects here even when they are statistically significant. The impact of segment length has been shown to influence classification performance,¹⁵ a finding we partially *replicate*, yet 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 processing choice is always a trade-off. Random forest performs better overall in D2, yielding paths for biomarkers to perform robustly with a high accuracy. This indicates that the data space in these cases is not as well linearly separable since the other two algorithms are linear classifiers. While this study provides the multivariate analysis across six processing steps, further variations to them and further steps in general such as feature extraction and selection as well as evaluation of the models offer room for further variations and need to be considered in future research.

BIOMARKERS

The performances of both datasets were significantly impacted by the biomarker choice, yet “good performing biomarkers” vary substantially between datasets and processing pipelines. If a “good pipeline” was accidentally chosen

in a traditional study only implementing one pipeline, a biomarker performing well in this pipeline would be reported as a good EEG biomarker for MDD even though it might perform badly with all other processing options. 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.

Even though the explanatory power of our analysis is limited by the size of the datasets, we find some biomarkers whose pattern demonstrate different types of robustness that are methodologically interesting. **Absolute centroid frequency** is in D1 very robust to analytical variability with classification performances above chance level across all pipelines, presents a high within-dataset replicability, and some paths with excellent performance. Even replication with D2 is possible to a certain degree. Similarly, **envelope skewness**, **envelope kurtosis**, and **absolute centroid frequency** exhibit a similar pattern for D1 but they fail to replicate completely in D2. In contrast, we 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 **envelope range** in D2. Those type of analyses clearly demonstrate the power of the multiverse approach and provide the opportunity to further reproducibility and translatability of biomarker research.

The set of biomarkers chosen for this study constitutes only a small fraction of possible EEG biomarker. Furthermore, the characteristics of the alpha band exhibit substantial inter-personal variations that are amongst other factors associated with age and cognitive function.⁵² 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.⁵³ Additional biomarkers can be derived from other frequency bands, or come from different categories such as nonlinear biomarkers, connectivity features, or entropy measures and provide complementary characterization of aberrant brain activity in MDD.³

CONCLUSION

To conclude, this study demonstrates the large influence of choice of processing steps and their combinations for two publicly available EEG resting-state datasets. 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. 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. However, given the frequent use of these two datasets and the fact that this size of dataset is well within the range of data used for current MDD biomarker studies²⁸ renders the current study a wake-up call for critical evaluation of the state of the field.

.....

DATA AND CODE AVAILABILITY STATEMENT

Code for biomarker calculation is available at: “python_code_for_biomarker_calculation.pdf”

Raw accuracies of diagnostic classification models are available at: “raw_classification_metrics.csv”

FUNDING SOURCES STATEMENT

YH was funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the ZIM program under grant number KK5207801BM0. RM was funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the ZIM program under grant number KK5207802SA4.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

Submitted: March 31, 2026 CDT. Revised: May 20, 2026 CDT.

Accepted: June 11, 2026 CDT. Published: July 29, 2026 CDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

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](https://doi.org/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](https://doi.org/10.1017/S0033291709991036)
3. 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](https://doi.org/10.1016/j.neubiorev.2019.07.021)
4. 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](https://doi.org/10.1109/ACCESS.2022.3146711)
5. 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](https://doi.org/10.1109/ACCESS.2021.3103047)
6. 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](https://doi.org/10.3389/fnhum.2018.00521)
7. 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](https://doi.org/10.1016/j.bpsc.2022.12.006)
8. 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](https://doi.org/10.1038/s41586-020-2314-9)
9. 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](https://doi.org/10.1109/ACCESS.2022.3146711)
10. Keil A, Debener S, Gratton G, et al. Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*. 2014;51(1):1-21. doi:[10.1111/psyp.12147](https://doi.org/10.1111/psyp.12147)
11. Hatz F, Hardmeier M, Bousleiman H, Rüegg S, Schindler C, Fuhr P. Reliability of fully automated versus visually controlled pre- and post-processing of resting-state EEG. *Clinical Neurophysiology*. 2015;126(2):268-274. doi:[10.1016/j.clinph.2014.05.014](https://doi.org/10.1016/j.clinph.2014.05.014)
12. Delorme A. EEG is better left alone. *Sci Rep*. 2023;13(1):2372. doi:[10.1038/s41598-023-27528-0](https://doi.org/10.1038/s41598-023-27528-0)
13. Savinov V, Sapunov V, Shusharina N, Botman S, Kamyshev 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](https://doi.org/10.1109/CNN53494.2021.9580293)
14. Ellis CA, Sancho ML, Miller RL, Calhoun VD. Identifying EEG Biomarkers of Depression with Novel Explainable Deep Learning Architectures. In: Longo L, Lapuschkin S, Seifert C, eds. *Explainable Artificial Intelligence*. Springer Nature Switzerland; 2024:102-124. doi:[10.1007/978-3-031-63803-9_6](https://doi.org/10.1007/978-3-031-63803-9_6)
15. 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](https://doi.org/10.1109/BIBM.2017.8217946)
16. Başar E, Schürmann M, Başar-Eroglu C, Karakaş S. Alpha oscillations in brain functioning: an integrative theory. *International Journal of Psychophysiology*. 1997;26(1):5-29. doi:[10.1016/S0167-8760\(97\)00753-8](https://doi.org/10.1016/S0167-8760(97)00753-8)
17. 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](https://doi.org/10.1007/978-981-13-0776-8_30)
18. 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](https://doi.org/10.1016/j.cmpb.2012.10.008)

19. 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](https://doi.org/10.1016/j.jocn.2017.09.030)
20. 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](https://doi.org/10.1109/BIBM.2016.7822696)
21. 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](https://doi.org/10.1186/s12911-015-0227-6)
22. 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](https://doi.org/10.1109/TNSRE.2020.2980223)
23. 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](https://doi.org/10.1002/dev.22163)
24. Ćukić 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](https://doi.org/10.2196/19548)
25. 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](https://doi.org/10.1177/1745691616658637)
26. 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](https://doi.org/10.7554/eLife.60595)
27. 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](https://doi.org/10.1016/j.neuroimage.2021.118712)
28. Mähler R, Reichenbach A. Beyond the Label “Major Depressive Disorder” – Detailed Characterization of Study Population Matters for EEG-Biomarker Research. *medRxiv*. Published online March 17, 2025:2025.03.17.25324119. doi:[10.1101/2025.03.17.25324119](https://doi.org/10.1101/2025.03.17.25324119)
29. 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](https://doi.org/10.1371/journal.pone.0171409)
30. Jasper HH. The 10/20 international electrode system. *EEG and Clinical Neurophysiology*. 1958;10(2):370-375.
31. 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](https://doi.org/10.1016/0013-4694(93)90121-B)
32. 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](https://doi.org/10.3389/fnins.2013.00267)
33. Jiang X, Bian GB, Tian Z. Removal of Artifacts from EEG Signals: A Review. *Sensors (Basel)*. 2019;19(5):987. doi:[10.3390/s19050987](https://doi.org/10.3390/s19050987)
34. 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](https://doi.org/10.1016/j.neuroimage.2013.10.027)
35. Li A, Feitelberg J, Saini AP, Höchenberger R, Scheltienne M. MNE-ICLabel: Automatically annotating ICA components with ICLabel in Python. *Journal of Open Source Software*. 2022;7(76):4484. doi:[10.21105/joss.04484](https://doi.org/10.21105/joss.04484)
36. 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. doi:[10.3389/fnagi.2013.00058](https://doi.org/10.3389/fnagi.2013.00058)
37. 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](https://doi.org/10.1038/s41592-019-0686-2)
38. 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](https://doi.org/10.1038/s41586-020-2649-2)
39. 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](https://doi.org/10.7554/eLife.70092)
40. 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](https://doi.org/10.1016/j.clinph.2003.11.034)

41. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(null):2825-2830.
42. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. In: ; 2010:92-96. doi:[10.25080/Majora-92bf1922-011](https://doi.org/10.25080/Majora-92bf1922-011)
43. Vallat R. Pingouin: statistics in Python. *Journal of Open Source Software*. 2018;3(31):1026. doi:[10.21105/joss.01026](https://doi.org/10.21105/joss.01026)
44. Christodoulou E, Reinke A, Andr   P, et al. False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims. *arXiv*. Published online August 6, 2025. doi:[10.48550/arXiv.2505.04720](https://doi.org/10.48550/arXiv.2505.04720)
45. 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](https://doi.org/10.1038/s41572-023-00454-1)
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](https://doi.org/10.1007/978-981-33-6044-0_7)
47. 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](https://doi.org/10.1038/nrn3475)
48. 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](https://doi.org/10.1002/da.23215)
49. 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](https://doi.org/10.1002/brb3.2476)
50. Salido Ruiz RA, Ranta R, Louis-Dorr V. EEG montage analysis in the Blind Source Separation framework. *Biomedical Signal Processing and Control*. 2011;6(1):77-84. doi:[10.1016/j.bspc.2010.06.007](https://doi.org/10.1016/j.bspc.2010.06.007)
51. Xiang TY, Wang WK, Zhou XH, et al. Cross-Dataset Montage Alignment via Electroencephalogram Source Imaging for Enhanced Motor Decoding. *IEEE Transactions on Cognitive and Developmental Systems*. Published online 2025:1-14. doi:[10.1109/TCDS.2025.3624805](https://doi.org/10.1109/TCDS.2025.3624805)
52. 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](https://doi.org/10.1016/S0165-0173(98)00056-3)
53. 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](https://doi.org/10.3389/fnins.2017.00506)

SUPPLEMENTARY MATERIALS

Supplementary Materials

Download: <https://apertureneuro.org/article/165175-reproducibility-of-electroencephalography-biomarkers-for-diagnosis-of-major-depressive-disorder/attachment/354006.docx>

Raw Classification Metrics

Download: <https://apertureneuro.org/article/165175-reproducibility-of-electroencephalography-biomarkers-for-diagnosis-of-major-depressive-disorder/attachment/354007.csv>

Python Code for Biomarker Calculation

Download: <https://apertureneuro.org/article/165175-reproducibility-of-electroencephalography-biomarkers-for-diagnosis-of-major-depressive-disorder/attachment/354008.pdf>
