Cross-Method Analysis of Explainability in EEG-Based Major Depressive Disorder Classification

Antonia Šarčević, Alan Jović, Nikolina Frid
University of Zagreb Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{antonia.sarcevic, alan.jovic, nikolina.frid}@fer.unizg.hr

Jakša Vukojević
University Psychiatric Hospital Vrapče
Bolnička 32, 10000 Zagreb, Croatia
jaksa.vukojevic@bolnica-vrapce.hr

Abstract. *Deep learning models have shown promising performance for EEG-based Major Depressive Disorder (MDD) detection, but their clinical use requires methods that can provide insight into model predictions. This study compares multiple post-hoc explainability techniques, including gradient-, perturbation-, and Shapley-based approaches, applied to an InceptionTime classifier trained on resting-state clinical EEG recordings. The results reveal partial agreement between attribution methods, alongside notable differences arising from their underlying methodological assumptions. These findings indicate that post-hoc explanations provide complementary but method-dependent perspectives on model behavior, highlighting the importance of assessing cross-method consistency when interpreting EEG-based depression classifiers.*

Keywords. Major Depressive Disorder, EEG, Explainable Artificial Intelligence, Deep Learning, InceptionTime

1 Introduction

Major depressive disorder (MDD) is among the most prevalent psychiatric disorders worldwide (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018). Despite extensive research, diagnosis remains primarily based on clinical assessment, motivating the search for objective measures that could support MDD evaluation. In this context, electroencephalography (EEG) has attracted considerable interest as a non-invasive and accessible modality for assessing MDD. Numerous studies have reported alterations in spectral, connectivity, and nonlinear EEG characteristics (Čukić et al., 2020; Kaiser et al., 2018; Simmatis et al., 2023); however, no objective biological marker has yet been established for routine clinical use (Kendler, 2016).

The increasing availability of EEG data has motivated the application of machine learning methods capable of identifying complex patterns that may not be apparent through conventional analysis alone. Both classical feature-based approaches and more recent deep learning models have reported promising results for MDD detection (Aleem et al., 2022; Avots et al., 2022; Dutua and Sultana, 2025; Wang et al., 2024). However, modern deep learning and time-series classification models learn high-dimensional representations whose relationship to underlying neurophysiological processes is often unclear, limiting direct human interpretation of model behavior (Elton, 2020; Jovic et al., 2025). Such limitations are particularly important in high-stakes domains such as healthcare (Rudin, 2019).

To improve the interpretability of the model, post-hoc explainability methods are commonly applied to black-box models (Linardatos et al., 2021; Marcinkevičs and Vogt, 2023). Widely used approaches include Shapley-value-based methods, gradient-based attribution methods, and perturbation-based techniques (Lundberg and Lee, 2017; Sundararajan et al., 2017; Zeiler and Fergus, 2014). Beyond improving transparency, such methods may help relate model predictions to physiologically meaningful EEG characteristics and support scientific interpretation of learned representations (Hicks et al., 2021; Mulc et al., 2025). However, these approaches rely on different assumptions and approximation strategies, which may lead to substantially different explanations for the same prediction (Jovic et al., 2025; Krishna et al., 2022). Consequently, explanation stability, robustness, and physiological plausibility remain important concerns. Despite these challenges, EEG studies commonly apply a single explainability technique, while systematic investigation of agreement and disagreement between multiple post-hoc explainability methods remains relatively limited.

In this work, we apply multiple post-hoc explainability methods to an InceptionTime model (Ismail Fawaz

et al., 2020) trained for EEG-based MDD detection. We investigate whether different explainability methods produce convergent attribution patterns and examine the extent to which their explanations remain consistent across the same classification task. Through this analysis, we assess both the potential usefulness and the limitations of post-hoc explainability for interpreting black-box EEG models in depression detection.

2 Materials and Methods

2.1 Dataset Description

EEG recordings were obtained at the University Psychiatric Hospital Vrapče using a Nihon Kohden EEG-1200K system from 140 adults: 70 patients with current moderate-to-severe MDD diagnosed by senior psychiatrists according to ICD-10 and 70 controls screened to exclude psychiatric or neurological disorders and psychopharmacotherapy. Groups were matched for age and sex as closely as possible (MDD: 37 women, 36.86±10.22 years; 33 men, 45.24±12.10; controls: 32 women, 35.88±9.96; 38 men, 36.16±11.35). Further cohort criteria are reported in (Mulc et al., 2025). Signals were sampled at 200 Hz from 19 electrodes positioned according to the international 10–20 system. Each 30-minute session began with resting-state recording and continued with activation protocols. Preprocessing included band-pass filtering (0.1–40 Hz), common-average re-referencing, and ICA-based artifact removal using ICLabel. The first 300 seconds of resting-state EEG were segmented into 2-second windows with a 0.5-second (25%) overlap, corresponding to a 1.5-second stride.

2.2 Model Training and Evaluation

The InceptionTime architecture (Ismail Fawaz et al., 2020) was selected due to its strong performance in time-series classification tasks (Middlehurst et al., 2024), including EEG analysis (Kendjel et al., 2026). Experiments were conducted using the InceptionTime implementation provided by the tsai library (Oguiza, 2023), which employs parallel convolutions with multiple kernel sizes to capture temporal patterns at different temporal scales and residual connections to facilitate optimization.

Channel-wise z-score normalization was performed using statistics computed exclusively from the training data within each fold. Model evaluation was conducted using subject-level stratified 5-fold cross-validation, with all epochs from each participant retained in the same fold to prevent information leakage. Within each training fold, 10% of subjects were reserved for validation. Epochs inherited the corresponding participant-level diagnosis as weak supervision, while inference and evaluation remained subject-level. For each subject, epoch-wise logits were aggregated using a trimmed-mean operator and converted into a single probability estimate before assigning the final class label. All reported metrics were calculated from these subject-level predictions. Training was performed using the Adam optimizer, cross-entropy loss, and One-Cycle learning-rate scheduling, and the checkpoint with the lowest validation loss was retained for subsequent explainability analysis.

2.3 Explainability Framework

To analyze model behavior, we evaluated five post-hoc explainability methods representing three major attribution paradigms: DeepSHAP (Lundberg and Lee, 2017; Shrikumar et al., 2017), Integrated Gradients (IG) (Sundararajan et al., 2017), Grad-CAM (Selvaraju et al., 2017), Occlusion (Zeiler and Fergus, 2014), and Permutation Feature Importance (PFI) (Breiman, 2001; Fisher et al., 2019). For each EEG segment, channel–time attributions were aggregated into channel-level importance values and then averaged across segments, subjects, and cross-validation folds. Signed attributions were converted to absolute values before channel-level aggregation; the resulting profiles therefore represent attribution magnitude rather than direction.

DeepSHAP was implemented using Captum's `DeepLiftShap` module with a background reference set of 40 randomly selected training samples per fold.

Integrated Gradients was computed using 100 interpolation steps. To assess baseline sensitivity, both zero-valued and Gaussian-noise baselines were evaluated.

Grad-CAM was applied to the final Inception block prior to global average pooling. To obtain channel-level relevance estimates, Grad-CAM was supplemented with InputXGradient attribution.

Occlusion analysis was performed using sliding temporal masks of size $(1, 50)$ and stride $(1, 25)$.

Permutation Feature Importance was computed by evaluating each EEG channel independently. For a selected channel, its temporal sequence was randomly permuted across the evaluation batch while all remaining channels were left unchanged. Channel importance was then quantified by the resulting decrease in target-class prediction confidence.

Frequency-band attribution was derived from the DeepSHAP and IG channel–time maps. A real FFT (rFFT) was applied along the temporal axis, and the resulting magnitudes were pooled into delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz) bands. Each score array was min–max normalized to the range 0–1; arrays with a range below 10^{-8} were set to zero. Class-specific scores were averaged across segments, subjects, and folds and then summed across channels within each region. The heatmaps therefore represent cumulative regional spectral attribution magnitude, which may depend partly on the number of channels assigned to each region, rather than EEG band

power.

Explanation robustness was evaluated using the coefficient of variation (CoV) for fold-wise variability, the Gini coefficient for attribution concentration, and Kendall's τ for cross-method ranking agreement. Higher Gini values indicate greater concentration. Fold-wise rank stability was additionally assessed using an overlap-based SRA score adapted from the sequential ranking principle of Ekstrøm et al. (2019). For each fold pair, the top-k Jaccard overlap at depth k was weighted by $1/k$, normalized by $\sum_{k=1}^{d} 1/k$, and averaged across fold pairs. SRA ranges from 0 to 1, with higher values indicating greater agreement. Regional SRA applies the same procedure after aggregating channels into cortical regions. In addition, a spatial-temporal variation metric (STV) (Mortaheb et al., 2019) was used to quantify the smoothness of attribution maps across neighboring EEG channels.

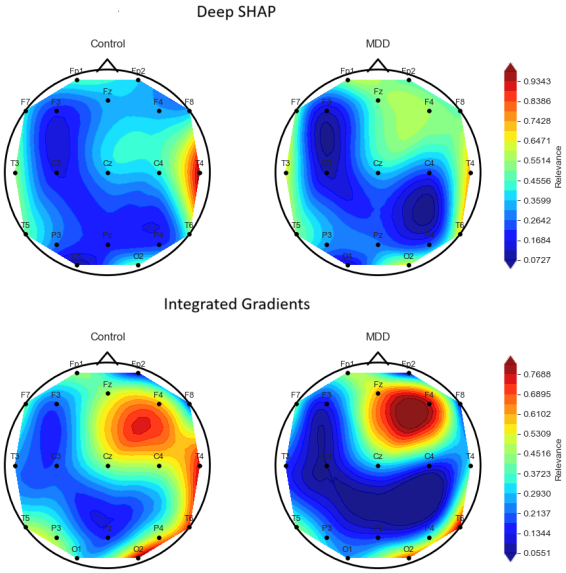


Figure 1. Topographic attribution maps

3 Results

3.1 Classification Performance

Classification performance was evaluated using subject-level predictions aggregated across EEG segments within a stratified 5-fold cross-validation framework. Mean performance metrics are summarized in Table 1. The model achieved strong and consistent performance across folds, with an average ROC AUC of 0.976 and PR AUC of 0.980. High precision (0.926) and recall (0.914) further indicate reliable identification of both MDD and control subjects. These internally validated subject-level results provide a suitable basis for analyzing model behavior but do not establish clinical validity or external generalizability.

Table 1. Subject-level classification performance (mean $\pm$ standard deviation).

Metric	Mean $\pm$ STD
Accuracy	0.914 ± 0.048
Precision	0.926 ± 0.032
Recall	0.914 ± 0.048
F1-score	0.913 ± 0.051
ROC AUC	0.976 ± 0.019
PR AUC	0.980 ± 0.014

3.2 Explainability Results

For each explainability method, attribution maps were computed for individual EEG segments and aggregated across subjects and cross-validation folds to obtain global class-level relevance profiles for the Control and MDD groups. Preliminary temporal attribution analysis did not reveal consistent patterns across folds or subjects. Consequently, the following analysis focuses on aggregated spatial and spectral attribution trends. Detailed visualizations are presented for DeepSHAP and Integrated Gradients, while channel-level agreement across all evaluated explainability methods is examined through the subsequent cross-method comparison and consensus analyses.

3.2.1 Global attribution patterns

DeepSHAP and Integrated Gradients (IG) produced broadly similar spatial attribution patterns, indicating partial agreement between the two explainability approaches (Fig. 1). Both methods highlighted the importance of right-hemisphere regions and revealed a broader spatial attribution distribution in the MDD group compared to Controls. For DeepSHAP, attributions in the Control group remained comparatively localized around the right temporal region, with T4 emerging as the most influential channel. In contrast, the MDD group exhibited a wider attribution pattern involving additional right-hemisphere channels, particularly T6, O2, F4, and the midline frontal channel Fz. Integrated Gradients showed stronger emphasis on frontal regions, especially F4 and Fz, in both groups and more prominently in MDD. While right temporal relevance remained evident for Controls, IG also attributed greater importance to parietal and occipital regions, including T5, T6, and O2.

The regional-frequency attribution maps generated using DeepSHAP (Fig. 2) indicate substantial contributions of beta-band activity in both groups. In the MDD group, attribution was more broadly distributed across delta-, theta-, and alpha-band components, with the strongest emphasis observed in frontal regions.

The regional-frequency heatmaps generated using IG (Fig. 3) indicate substantial contributions from frontal and temporal regions across multiple frequency bands. Compared to the Control group, the MDD

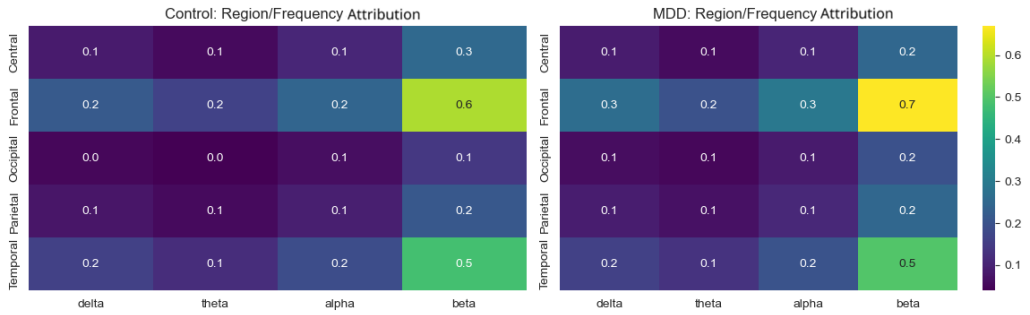


Figure 2. DeepSHAP regional and frequency-band attribution heatmap

group exhibited greater relative attribution, particularly within the delta and beta bands.

3.2.2 Stability and cross-method agreement

Table 2 and Fig. 4 summarize the stability and agreement metrics. Occlusion yielded the highest channel-level and regional SRA values (0.428 and 0.574), indicating the strongest fold-wise ranking agreement, while Permutation yielded the lowest (0.303 and 0.397). Occlusion exceeded IG by 0.093 and 0.107 and Permutation by 0.125 and 0.177 at the channel and regional levels, respectively. These comparisons are descriptive rather than inferential.

CoV values ranged from 0.548 to 0.578, indicating comparable, but not necessarily low, fold-wise variability. Gini ranged from 0.185 for DeepSHAP, indicating the most diffuse attribution profile, to 0.234 for Occlusion, indicating the most concentrated. STV varied only from 0.299 to 0.310, indicating comparable spatial smoothness across methods.

Table 2. Quantitative evaluation of explainability-method stability and agreement

Method	SRA↑	Reg. SRA↑	CoV↓	Gini↑	STV↓
IG	0.335	0.467	0.565	0.226	0.299
DeepSHAP	0.349	0.515	0.578	0.185	0.308
Grad-CAM	0.355	0.423	0.560	0.218	0.310
Occlusion	0.428	0.574	0.548	0.234	0.303
PFI	0.303	0.397	0.561	0.216	0.308

Agreement between explainability methods was further assessed using pairwise Kendall's Tau correlations (Fig. 5). The strongest agreement was observed between IG and Grad-CAM ($\tau = 0.93$), while IG and Permutation ($\tau = 0.89$) and Grad-CAM and Permutation ($\tau = 0.87$) also demonstrated substantial ranking similarity. In contrast, DeepSHAP showed noticeably lower agreement with the remaining approaches, with correlation values ranging from 0.36 to 0.44.

These results indicate that gradient-based and perturbation-based methods produced partially convergent attribution rankings, whereas DeepSHAP yielded comparatively distinct feature attribution profiles.

3.2.3 Consensus attribution profile

To summarize attribution patterns across methods, channel-level attributions were averaged across samples and min–max normalized within each method to the range 0–1. The consensus score was the unweighted mean of the normalized values across methods. Equal weighting was used because no ground-truth attribution or validated basis for method-specific weights was available; methodological differences were examined separately through the cross-method analyses. Cross-method agreement was calculated as $1/(1 + \sigma_c)$, where σ_c is the standard deviation of the normalized method scores for channel c. Higher values indicate closer agreement, not statistical confidence. Because methods may agree on either high or low importance, agreement was interpreted jointly with consensus. Fig. 6 presents the channels with the highest consensus scores and their agreement values.

The aggregated rankings consistently prioritized the right temporoparietal channel T6 and the right occipital channel O2, followed by the right frontal channel F4, the midline frontal channel Fz, and the right temporal channel T4. These channels also exhibited relatively high cross-method agreement scores across the evaluated explainability methods.

4 Discussion

The explainability methods converged partially on a common spatial interpretation. Integrated Gradients and Grad-CAM estimate local output sensitivity, whereas Occlusion and PFI measure prediction changes following finite input perturbations. Their broadly comparable channel rankings suggest that these distinct probes recovered a common coarse spatial pattern of model reliance. DeepSHAP yielded a more distinct and evenly distributed profile, consistent with its allocation of contributions relative to a background set and its sensitivity to reference selection and correlations among EEG channel–time features. Across methods, right frontal (F4, Fz), right temporal/parietal (T6), and posterior (O2) channels were repeatedly emphasized and prioritized by the consensus

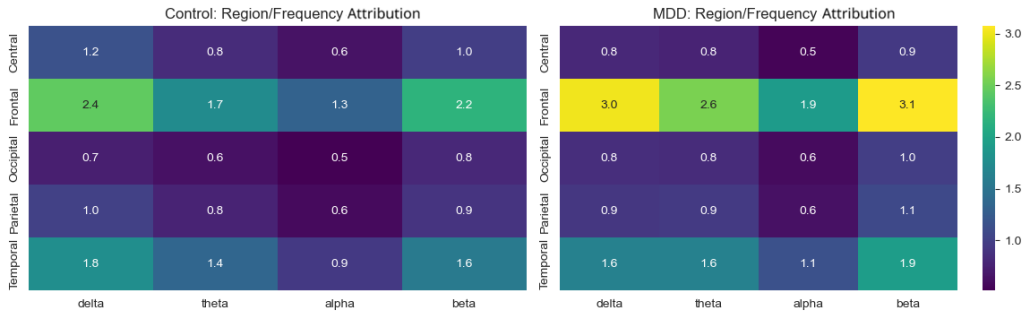


Figure 3. IG regional and frequency-band attribution heatmap

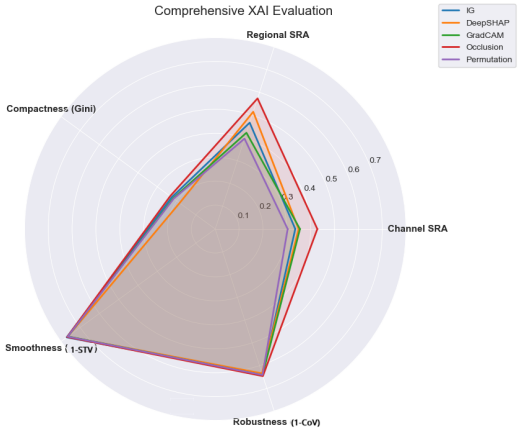


Figure 4. Descriptive comparison of explanation stability and agreement metrics. CoV and STV were transformed to align their visual direction with the agreement metrics.

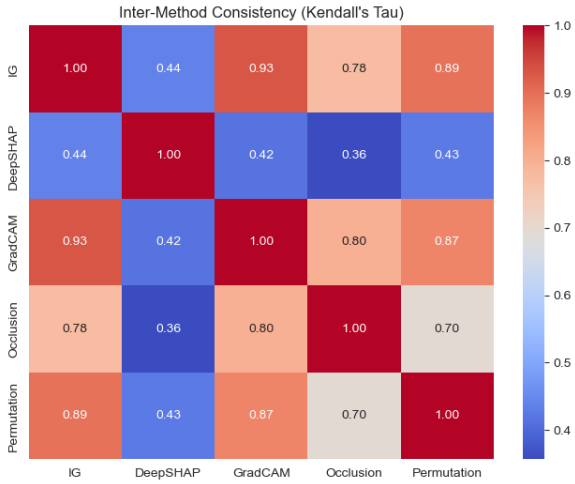


Figure 5. Inter-method consistency heatmap

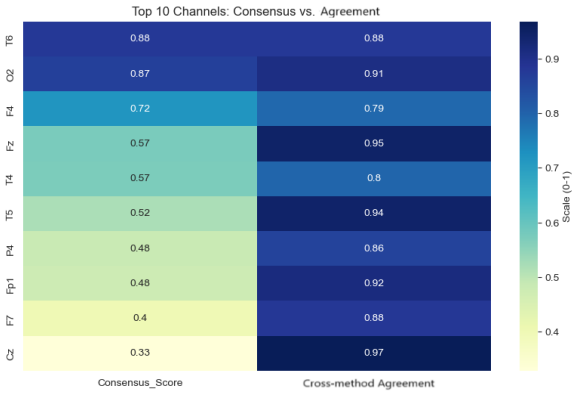


Figure 6. Consensus and cross-method agreement

analysis. Frequency-band rankings varied more substantially, indicating that spectral interpretations were less robust to method choice than coarse spatial patterns.

These spatial patterns are broadly consistent with previous EEG findings in MDD. Frontal asymmetry and altered right frontal activity have been associated with disrupted emotional regulation and cognitive control (Allen et al., 2018; Davidson, 1998; Torre-Luque and Bornas, 2017), while altered hemispheric organization and frontal network activity have also been reported in depressive populations (Du et al., 2023; Teng et al., 2022). Similarly, the emphasis on T6 and O2 is compatible with reported alterations in posterior alpha- and beta-band activity (Čukić et al., 2020; Fingelkurts et al., 2006; Simmatis et al., 2023). These correspondences support the physiological plausibility of the learned patterns but do not establish causal relationships or clinical biomarkers.

The practical value of multi-method explainability lies in model auditing and hypothesis generation. Agreement across methods with different assumptions provides stronger evidence of a stable model dependency than an individual explanation, whereas disagreement identifies interpretations requiring further validation. The distinct DeepSHAP profile and variable frequency-band rankings demonstrate that explanations remain method-dependent and may not faithfully represent the model's internal reasoning.

Several limitations affect these findings. Although the cohort of 140 participants is comparatively large for EEG-based MDD classification (Kendjel et al., 2026), its single-site origin and lack of external validation limit generalizability. Residual demographic

differences, medication effects, and other unmeasured participant characteristics may also influence classification and attribution patterns. The high subject-level AUC may partly reflect epoch aggregation and the distinction between moderate-to-severe MDD and screened healthy controls; performance in milder, comorbid, or external populations remains unknown. EEG results are also sensitive to acquisition conditions, preprocessing, and inter-subject variability. Finally, the analysis focused on global profiles and did not examine subject-specific explanations or transient temporal relevance.

Future work should evaluate classification and explanation stability across independent, multi-site cohorts, preprocessing pipelines, and attribution configurations. Inherently interpretable models, connectivity-aware architectures, and neurophysiological priors may further help distinguish reproducible signal structure from explainer-specific effects.

5 Conclusion

This study compared five post-hoc explainability methods applied to an InceptionTime model for EEG-based MDD classification. Integrated Gradients, Grad-CAM, Occlusion, and PFI produced broadly comparable channel rankings despite their different assumptions, whereas DeepSHAP yielded the most distinct and evenly distributed attribution profile. Nevertheless, several frontal, temporal, and posterior channels were consistently emphasized across methods, while frequency-band rankings showed greater variability.

The practical value of post-hoc explainability in this setting is not the direct identification of neurological biomarkers, but the auditing of black-box model behavior. Cross-method convergence identifies candidate model dependencies that are comparatively robust to explainer choice, whereas disagreement reveals interpretations that are method-dependent and should be treated cautiously. Multi-method analysis can therefore guide robustness checks, investigation of possible confounding or predictive shortcuts, and selection of hypotheses for independent validation, but cannot by itself establish causal neurophysiological relevance or clinical validity.

Acknowledgments

This work was supported by the European Union-NextGenerationEU project NPOO 581-16956 VISTA-Health, and the Croatian Science Foundation under the project number HRZZ-IP-2022-10-8241. Funders played no role in the study design, data collection, data analysis and interpretation, or the writing of this manuscript.

References

Aleem, Shumaila, Noor ul Huda, Rashid Amin, Samina Khalid, Sultan S. Alshamrani, and Abdullah Alshehri (2022). "Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions". In: *Electronics* 11.7. ISSN: 2079-9292. DOI: 10.3390/electronics11071111. URL: https://www.mdpi.com/2079-9292/11/7/1111.

Allen, John J. B., Philipp M. Keune, Michael Schönenberg, and Robin Nusslock (2018). "Frontal EEG alpha asymmetry and emotion: From neural underpinnings and methodological considerations to psychopathology and social cognition". In: *Psychophysiology* 55.1, e13028. DOI: https://doi.org/10.1111/psyp.13028.

Avots, Egils, Klāvs Jermakovs, Maie Bachmann, Laura Päeske, Cagri Ozcinar, and Gholamreza Anbarjafari (2022). "Ensemble approach for detection of depression using EEG features". In: *Entropy* 24.2, p. 211. DOI: 10.3390/e24020211.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Čukić, Milena, Miodrag Stokić, Slavoljub Radenković, Miloš Ljubisavljević, Slobodan Simić, and Danka Savić (2020). "Nonlinear analysis of EEG complexity in episode and remission phase of recurrent depression". In: *International Journal of Methods in Psychiatric Research* 29.2, e1816. DOI: 10.1002/mpr.1816.

Davidson, Richard J. (1998). "Affective Style and Affective Disorders: Perspectives from Affective Neuroscience". In: *Cognition and Emotion* 12.3, pp. 307–330. DOI: 10.1080/026999398379628.

Du, Yishan, Lingling Hua, Shui Tian, ZhongPeng Dai, Yi Xia, Shuai Zhao, HaoWen Zou, Xiaoqin Wang, Hao Sun, Hongliang Zhou, YingHong Huang, Zhi-Jian Yao, and Qing Lu (2023). "Altered beta band spatial-temporal interactions during negative emotional processing in major depressive disorder: An MEG study". In: *Journal of Affective Disorders* 338, pp. 254–261. ISSN: 0165-0327. DOI: 10.1016/j.jad.2023.06.001.

Dutua, Stefana and Alina Elena Sultana (2025). "Optimizing Depression Classification Using Combined Datasets and Hyperparameter Tuning with Optuna". In: *Sensors* 25.7, p. 2083. DOI: 10.3390/s25072083.

Ekstrøm, C. T., T. A. Gerds, and A. K. Jensen (2019). "Sequential Rank Agreement Methods for Comparison of Ranked Lists". In: *Biostatistics* 20.4, pp. 582–598. DOI: 10.1093/biostatistics/kxy017.

Elton, Daniel C. (2020). "Self-explaining AI as an Alternative to Interpretable AI". In: *Artificial General Intelligence*. Ed. by Ben Goertzel, Aleksandr I. Panov, Alexey Potapov, and Roman Yampolskiy.

Cham: Springer International Publishing, pp. 95–106. DOI: 10.1007/978-3-030-52152-3_10.

Fingelkurts, Alexander A., Andrew A. Fingelkurts, Heikki Rytsälä, Kirsi Suominen, Erkki Isometsä, and Seppo Kähkönen (2006). "Composition of brain oscillations in ongoing EEG during major depression disorder". In: *Neuroscience Research* 56.2, pp. 133–144. ISSN: 0168-0102. DOI: https://doi.org/10.1016/j.neures.2006.06.006.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici (2019). "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously". In: *Journal of Machine Learning Research* 20.177, pp. 1–81. URL: http://jmlr.org/papers/v20/18-760.html.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators (2018). "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159. Epub 2018 Nov 8; Erratum in: The Lancet. 2019;393(10190):e44, pp. 1789–1858. DOI: 10.1016/S0140-6736(18)32279-7.

Hicks, Steven A., Jonas L. Isaksen, Vajira Thambawita, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Inga Strümke, Christina Ellervik, Morten Salling Olesen, Torben Hansen, Claus Graff, Niels-Henrik Holstein-Rathlou, Pål Halvorsen, Mary M. Maleckar, and Jørgen K. Kanters (2021). "Explaining deep neural networks for knowledge discovery in electrocardiogram analysis". In: *Scientific Reports* 11.1, p. 10949. DOI: 10.1038/s41598-021-90285-5.

Ismail Fawaz, Hassan, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean (2020). "Inceptiontime: Finding alexnet for time series classification". In: *Data mining and knowledge discovery* 34.6, pp. 1936–1962. URL: https://doi.org/10.1007/s10618-020-00710-y.

Jovic, Alan, Nikolina Frid, Karla Brkic, and Mario Cifrek (2025). "Interpretability and accuracy of machine learning algorithms for biomedical time series analysis – a scoping review". In: *Biomedical Signal Processing and Control* 110, p. 108153. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2025.108153. URL: https://www.sciencedirect.com/science/article/pii/S1746809425006640.

Kaiser, Andreas Kurt, Maria-Theresa Gnjezda, Stephanie Knasmüller, and Wolfgang Aichhorn (2018). "Electroencephalogram alpha asymmetry in patients with depressive disorders: current perspectives". In: *Neuropsychiatric Disease and Treatment* 14. PMID: 29928121, pp. 1493–1504. DOI: 10.2147/NDT.S137776. eprint: https://www.tandfonline.com/doi/pdf/10.2147/NDT.S137776. URL: https://www.tandfonline.com/doi/abs/10.2147/NDT.S137776.

Kendjel, Saša, Nikolina Frid, Eda Kalafatić, Krešimir Jozić, and Alan Jović (2026). "ALFABEATS: A Machine Learning Based Decision Support Platform for Biomedical Time Series Analysis". In: *Emerging Trends in Computational Biology, Biomedical Engineering, and Health Informatics*. Cham: Springer Nature Switzerland, pp. 213–224. ISBN: 978-3-032-22199-5. DOI: 10.1007/978-3-032-22199-5_15.

Kendler, Kenneth S. (2016). "The Phenomenology of Major Depression and the Representativeness and Nature of DSM Criteria". In: *American Journal of Psychiatry* 173.8, pp. 771–780. DOI: 10.1176/appi.ajp.2016.15121509. eprint: https://psychiatryonline.org/doi/pdf/10.1176/appi.ajp.2016.15121509. URL: https://psychiatryonline.org/doi/abs/10.1176/appi.ajp.2016.15121509.

Krishna, Satyapriya, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju (2022). "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective". In: *Trans. Mach. Learn. Res.* 2024. URL: https://api.semanticscholar.org/CorpusID:246485817.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2021). "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1. ISSN: 1099-4300. DOI: 10.3390/e23010018. URL: https://www.mdpi.com/1099-4300/23/1/18.

Lundberg, Scott M. and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 4768–4777. ISBN: 9781510860964. URL: https://dl.acm.org/doi/10.5555/3295222.3295230.

Marcinkevičs, Ričards and Julia E. Vogt (2023). "Interpretable and explainable machine learning: A methods-centric overview with concrete examples". In: *WIREs Data Mining and Knowledge Discovery* 13.3, e1493. DOI: https://doi.org/10.1002/widm.1493. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1493. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1493.

Middlehurst, Matthew, Patrick Schäfer, and Anthony Bagnall (2024). "Bake off redux: a review and ex-

perimental evaluation of recent time series classification algorithms". In: *Data Mining and Knowledge Discovery* 38.4, pp. 1958–2031. DOI: 10.1007/s10618-024-01022-1.

Mortaheb, Sepehr, Jitka Annen, Camille Chatelle, Helena Cassol, Geraldine Martens, Aurore Thibaut, Olivia Gosseries, and Steven Laureys (2019). "A Graph Signal Processing Approach to Study High Density EEG Signals in Patients with Disorders of Consciousness". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4549–4553. DOI: 10.1109/EMBC.2019.8856436.

Mulc, Damir, Jaksa Vukojevic, Eda Kalafatic, Mario Cifrek, Domagoj Vidovic, and Alan Jovic (2025). "Opportunities and Challenges for Clinical Practice in Detecting Depression Using EEG and Machine Learning". In: *Sensors* 25.2. ISSN: 1424-8220. DOI: 10.3390/s25020409.

Oguiza, Ignacio (2023). *tsai - A state-of-the-art deep learning library for time series and sequential data*. Github. URL: https://github.com/timeseriesAI/tsai.

Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5, pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626. URL: https://dl.acm.org/doi/10.1007/s11263-019-01228-7.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). "Learning important features through propagating activation differences". In: *International conference on machine learning*. PMlR, pp. 3145–3153. URL: https://dl.acm.org/doi/10.5555/3305890.3306006.

Simmatis, Leif, Emma E. Russo, Joseph Geraci, Irene E. Harmsen, Nardin Samuel, et al. (2023). "Technical and clinical considerations for electroencephalography-based biomarkers for major depressive disorder". In: *npj Mental Health Research* 2.18. DOI: 10.1038/s44184-023-00038-7.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR, pp. 3319–3328. URL: https://dl.acm.org/doi/10.5555/3305890.3306024.

Teng, Chaolin, Mengwei Wang, Wei Wang, Jin Ma, Min Jia, Min Wu, Yuanyuan Luo, Yu Wang, Yiyang Zhang, and Jin Xu (2022). "Abnormal Properties of Cortical Functional Brain Network in Major Depressive Disorder: Graph Theory Analysis Based on Electroencephalography-Source Estimates". In: *Neuroscience* 506, pp. 80–90. ISSN: 0306-4522. DOI: https://doi.org/10.1016/j.neuroscience.2022.10.010.

Torre-Luque, Alejandro de la and Xavier Bornas (2017). "Complexity and Irregularity in the Brain Oscillations of Depressive Patients: A Systematic Review". In: *Neuropsychiatry* 7.5, pp. 279–290. DOI: 10.4172/Neuropsychiatry.1000232.

Wang, Yuwen, Yudan Peng, Mingxiu Han, Xinyi Liu, Haijun Niu, Jian Cheng, Suhua Chang, and Tao Liu (2024). "GCTNet: a graph convolutional transformer network for major depressive disorder detection based on EEG signals". In: *Journal of Neural Engineering* 21.3, p. 036042. DOI: 10.1088/1741-2552/ad5048.

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer, pp. 818–833. URL: https://link.springer.com/chapter/10.1007/978-3-319-10590-1_53.