Latent Predictors of Antidepressant Response from EEG via Semi-Supervised Autoencoders

JD Pruett
Stanford University
Department of Psychology

1 Introduction

Major depressive disorder (MDD) is a highly heterogeneous illness for which standard antidepressant treatments achieve remission in only a minority of patients (Watts et al. 2022). Landmark trials (e.g. STAR*D) and large surveys show that most patients fail to remit after one or several courses of antidepressants, and no single first-line medication is clearly superior his chronic treatment resistance means many patients endure prolonged trial-and-error therapy before finding an effective regimen (Sinyor, Schaffer, and Levitt 2010). It has therefore become a major goal to develop personalized predictors of treatment outcome so that each patient can be matched to the therapy most likely to work for them.

Machine learning has been applied to this problem by building prognostic models of treatment response based on patient features. Early efforts often relied on baseline clinical and demographic variables, which have shown only limited predictive power (van Mens et al. 2023; Jankowsky et al. 2023; Chen et al. 2023). In recent years, there has been growing interest in neurophysiological biomarkers – especially brain imaging – as inputs to these models (Cohen et al. 2021). For example, magnetic resonance imaging (MRI) studies have achieved moderate accuracy (AUC $\approx$0.84) in distinguishing responders from non-responders.

However, MRI is expensive and not easily scalable (Andrade et al. 2023). In contrast, electroencephalography (EEG) is a low-cost, portable measure of brain activity. Resting-state EEG captures large-scale neural oscillations and network dynamics that can reflect the brain's functional state (Sargent et al. 2021). Because it is cheap and rapid to acquire, EEG has potential as a clinical tool for guiding depression treatment decisions.

Yet most prior studies of machine learning applications to antidepressant EEG prediction relied on classifiers—such as LDA, k-nearest neighbors (kNN), radial kernel SVMs, or decision rule-based methods, and none employed deep latent variable models (Watts et al. 2022). Such models, however, may be better suited to learning temporally structured patterns in high-dimensional EEG data. In this project, we introduce a semi-supervised architecture for predicting individual response to treatment.

We first train a standard autoencoder using resting-state EEG to learn a compact latent representation of brain activity. We then develop a semi-supervised variant in which the encoder is optimized not only to reconstruct the EEG signal but also to predict clinical response, shaping the latent space to

capture features that are relevant for treatment outcomes. These learned representations are then fed into a simple classifier and evaluated on held-out data to assess generalization performance.

In addition, we compare two architectural variants—a standard autoencoder (AE) and a recurrent variational autoencoder (RVAE)—to investigate whether incorporating temporal dynamics improves the model's ability to predict treatment response. Broadly, this study permits the investigation of whether distributed, data-driven EEG signatures can generalize across medications and improve predictive performance beyond baseline clinical and demographic variables. More specifically, this design allows us to ask not only whether latent representations from EEG can improve predictive accuracy beyond clinical and demographic variables, but also how the structure of those representations—static vs. temporal—influences model performance. To our knowledge, this is the among the first applications of semi-supervised latent representation learning to antidepressant response prediction from EEG data. Optimistically, this line of research can offer a computationally scalable and neuroscientifically grounded approach for uncovering individualized biomarkers of treatment response.

2 Background and Related Work

A number of EEG-derived measures have been linked to treatment outcome. Classical analyses have focused on spectral band powers (e.g. alpha, theta) and their topographical patterns. For instance, frontal alpha asymmetry has long been studied in depression: greater right-frontal alpha (implying left-frontal hypoactivity) has been reported to predict better response to SSRIs in certain subgroups. In the iSPOT-D trial, Arns et al. (2016) found that this frontal alpha asymmetry predicted a favorable response to SSRIs (escitalopram/sertraline) in women but not in men. Other EEG features – such as rostral anterior cingulate theta power or global connectivity measures – have shown promise in isolated studies (Wu et al. 2020). However, no single EEG biomarker has proved robust or widely replicable across studies.

Meta-analyses indicate that, on average, machine-learning models using EEG can distinguish responders from non-responders with an accuracy of 84% (pooled, AUC). These models often use combinations of band powers, asymmetry indices, connectivity or entropy features, and typically identify non-responders more reliably than responders (Watts et al. 2022).

Recently, deep learning and latent-space methods have been applied to EEG for response prediction. For example, Wu and colleagues developed SELSER, a supervised latent-space model that optimizes spatial filters and band-power features to predict antidepressant outcome from resting EEG. Using placebo-controlled trial data, SELSER achieved significant prediction of symptom improvement to sertraline (Wu et al. 1011). Similarly, Schwartzmann et al. built an EEG-based classifier for SSRI response that showed promising accuracy for two different SSRIs (Schwartzmann et al. 2023). These works demonstrate the potential of data-driven latent representations of EEG to capture the complex neural signatures of treatment response.

The present model extends prior work in three key ways. First, we adapt a Recurrent Variational Autoencoder (RVAE) to model the temporal structure of EEG data. By combining the strengths of sequence learning and probabilistic generative modeling, RVAEs are well suited to capture the dynamic, high-dimensional patterns characteristic of EEG recordings. Second, we introduce a framework that integrates clinical features alongside learned EEG-based representations, enabling the model to draw on both physiological and contextual predictors of treatment response. Third, we implement an end-to-end semi-supervised architecture that not only reconstructs EEG input but also learns latent features optimized for clinical outcome prediction. Rather than treating the latent space purely as an intermediate compression layer, we explicitly train it to encode information relevant to the target variable—treatment response—by attaching a supervised prediction head to the latent representation.

3 Model

3.1 iSPOT-D Dataset

The International Study to Predict Optimized Treatment in Depression (iSPOT-D) is a large, multisite, open-label, randomized clinical trial designed to identify predictors of antidepressant treatment

response in major depressive disorder (MDD). The study enrolled over 1,000 adults aged 18–65 who met DSM-IV criteria for non-psychotic MDD and had either never received antidepressant treatment or had undergone a sufficient washout period. Participants were randomized to receive one of three commonly prescribed medications: escitalopram, sertraline, or venlafaxine-XR. Medication dosing followed standard therapeutic guidelines and was managed by treating clinicians. A comprehensive battery of clinical, cognitive, genetic, and neurophysiological assessments was administered, including both clinician-rated (e.g., HDRS-17) and self-report (e.g., QIDS-SR16) measures of depressive symptoms. Resting-state EEG was recorded at multiple timepoints—prior to treatment, one week after initiation, and again at six to eight weeks—along with additional follow-ups by telephone. Treatment response was assessed after eight weeks using established clinical criteria (e.g., $\geq 50\%$ reduction in HDRS-17). The present analysis focuses on EEG data from approximately 700 participants who completed baseline and follow-up recordings and had outcome data available.

3.2 Encoder/Decoder Architecture

We implemented both unsupervised and semi-supervised autoencoder architectures to learn latent representations from high-dimensional EEG data and clinical features, with the goal of predicting antidepressant treatment response. Each subject's input vector $\mathbf{x}_i \in \mathbb{R}^D$ comprised two components: (1) EEG-derived features extracted from two-minute resting-state recordings across 26 scalp electrodes sampled at 500 Hz, and (2) clinical features, including age, gender, race, assigned medication (escitalopram, sertraline, or venlafaxine), and 17 items from the Hamilton Depression Rating Scale (HAM-D), along with derived subscales indexing emotional, behavioral, and somatic symptom domains.

Our baseline unsupervised model is a standard autoencoder composed of an encoder function f_θ and decoder function g_ϕ. The encoder maps the input $\mathbf{x}_i$ to a latent representation $\mathbf{z}_i \in \mathbb{R}^L$: $\mathbf{z}_i = f_\theta(\mathbf{x}_i)$. The decoder then reconstructs the input from the latent code: $\hat{\mathbf{x}}_i = g_\phi(\mathbf{z}_i)$. The model is trained to minimize the reconstruction loss:$\mathcal{L}_{\text{recon}} = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$.

To predict clinical outcomes, we extend the model with semi-supervised variants. In these versions, the latent representation $\mathbf{z}_i$ is passed through a prediction head h_ψ to estimate the probability of treatment response: $\hat{y}_i = h_\psi(\mathbf{z}_i) = \sigma(\text{MLP}_{\text{pred}}(\mathbf{z}_i))$, where σ denotes the sigmoid activation function. The supervised component of the loss is the binary cross-entropy between predicted and true treatment response labels: $\mathcal{L}_{\text{sup}} = -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)$.

As shown in Figure 1, the total loss function used to train the semi-supervised model is a weighted combination of reconstruction and classification losses: $\mathcal{L}_{\text{semi}} = \lambda \mathcal{L}_{\text{recon}} + (1-\lambda)\mathcal{L}_{\text{sup}}$ where $\lambda \in [0, 1]$ balances the influence of unsupervised and supervised objectives and is tuned as a hyperparameter.

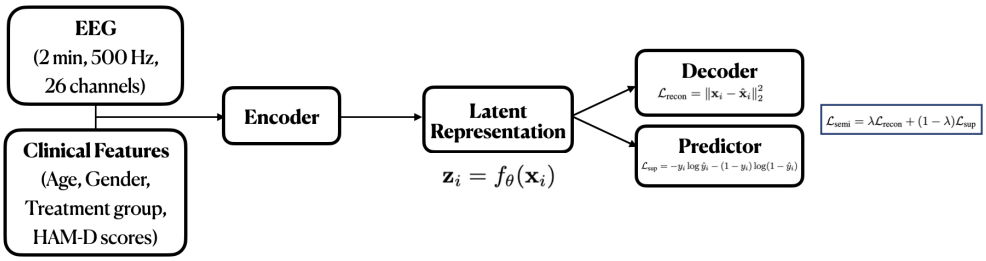


Figure 1: Encoder/Decoder Architecture for Unsupervised and Semi-Supervised Variants

The architecture of the encoder and decoder is symmetric and composed of fully connected layers. Specifically, the encoder comprises a multilayer perceptron (MLP) with dimensions:

$$\text{Linear}(D \to 512) \to \text{ReLU} \to \text{Linear}(512 \to 256) \to \text{ReLU} \to \text{Linear}(256 \to L)$$

The decoder reverses this structure, mapping from $\mathbb{R}^L$ back to $\mathbb{R}^D$. The predictor network in the semi-supervised model adds an additional branch:

$$\text{Linear}(L \to 32) \to \text{ReLU} \to \text{Linear}(32 \to 1) \to \text{Sigmoid}$$

After training, as Figure 2 highlights, the encoder is used to extract compact latent features from the input EEG and clinical data. These latent representations $\mathbf{z}_i \in \mathbb{R}^L$, which capture subject-level

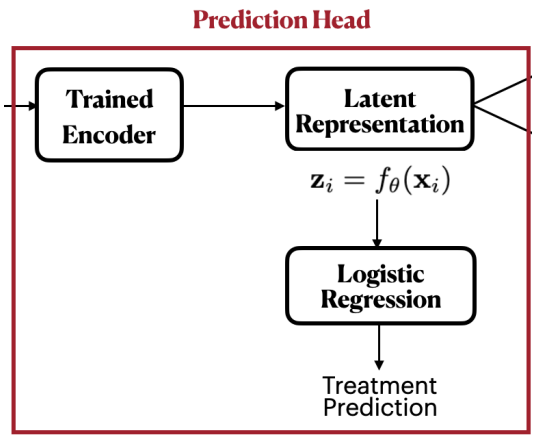


Figure 2: Binary Treatment Response Pipeline

structure in the data, are then used as input to a separate downstream classifier. Specifically, we extract latent vectors from the trained model using:

$$\mathbf{z}_i = f_\theta(\mathbf{x}_i)$$

where f_θ denotes the encoder network. These latent embeddings are aggregated over the dataset and serve as features in a standard logistic regression model for predicting binary treatment response. This two-stage approach allows us to evaluate the predictive value of the learned representations independently from the autoencoder's internal predictor (if using the semi-supervised model).

3.3 RVAE

To better capture the inherent temporal dynamics of EEG recordings, we implement a Recurrent Variational Autoencoder (RVAE)—a model that extends standard autoencoders by incorporating both sequential structure and probabilistic latent representations. Unlike static autoencoders, which flatten time-series data into static vectors, the RVAE explicitly models time-dependent patterns using a GRU-based encoder. This encoder processes the EEG sequence and produces a latent distribution parameterized by a mean and log-variance, from which latent codes are sampled via the reparameterization trick. This enforces a smooth, structured latent space that promotes generalization and robustness—key advantages of VAEs in modeling high-dimensional biosignals. A GRU-based decoder reconstructs the original sequence from the latent code, preserving temporal coherence in the reconstruction and ensuring that the learned representations reflect the underlying neural dynamics.

In our model, 2-minute resting-state EEG recordings are processed as sequences of 116 timesteps with 26 channels per frame. The encoder consists of a GRU that summarizes temporal dependencies into a hidden state, which is then transformed into the parameters $\boldsymbol{\mu}$ and $\log \boldsymbol{\sigma}$ of a diagonal Gaussian distribution. A latent vector $\mathbf{z}$ is sampled from this distribution using the reparameterization trick, enabling backpropagation through the stochastic node. The decoder GRU reconstructs the input by generating a sequence conditioned on $\mathbf{z}$, which is linearly mapped back to the electrode space. To align the latent space with clinically meaningful structure, a supervised prediction head maps $\mathbf{z}$ to the probability of antidepressant treatment response.

This semi-supervised RVAE is trained with a composite objective that combines (1) reconstruction loss to preserve signal fidelity, (2) Kullback-Leibler divergence to regularize the latent space toward a unit Gaussian prior, and (3) binary cross-entropy loss to encourage outcome prediction from latent variables:

$$\mathcal{L} = \lambda_{\text{recon}} \cdot \|\mathbf{x} - \hat{\mathbf{x}}\|^2 + \lambda_{\text{KL}} \cdot \text{KL}\left(q(\mathbf{z}|\mathbf{x}) \| \mathcal{N}(0, \mathbf{I})\right) + \lambda_{\text{sup}} \cdot \mathcal{L}_{\text{BCE}}(y, \hat{y})$$

4 Results

We evaluated the predictive performance of latent EEG representations learned through different autoencoder architectures compared to clinical features alone for antidepressant response prediction.

Table 1 summarizes the logistic regression performance across all model configurations on a held-out test set of 100 patients.

Table 1: Logistic Regression Results on Latent Representations Across Models

EEG Autoencoder Type	Input Features	Accuracy	ROC-AUC
None	Clinical	0.620	0.513
Unsupervised AE	EEG	0.610	0.533
Semi-Supervised AE	EEG	0.590	0.526
Semi-Supervised RVAE	EEG	0.620	0.518
Unsupervised AE	Clinical + EEG	0.590	0.417
Semi-Supervised AE	Clinical + EEG	0.600	0.557
Semi-Supervised RVAE	Clinical + EEG	0.580	0.456

The semi-supervised autoencoder with clinical features achieved the highest ROC-AUC of 0.557, representing a 0.044 improvement over clinical features alone (ROC-AUC = 0.513). EEG-only models demonstrated limited predictive value, with ROC-AUC scores ranging from 0.518 to 0.533. The unsupervised autoencoder achieved the best EEG-only performance at 0.533 ROC-AUC.

Among combined EEG+clinical approaches, performance varied considerably. The semi-supervised autoencoder with clinical features (0.557) outperformed the unsupervised autoencoder with clinical features (0.417) and semi-supervised RVAE with clinical features (0.456). Notably, two of the three combined models (unsupervised AE + clinical: 0.417; semi-supervised RVAE + clinical: 0.456) performed worse than clinical features alone.

Accuracy scores ranged from 0.580 to 0.620 across all models, with clinical-only and semi-supervised RVAE achieving the highest accuracy of 0.620. All ROC-AUC scores remained within 0.06 points of random chance performance (0.5).

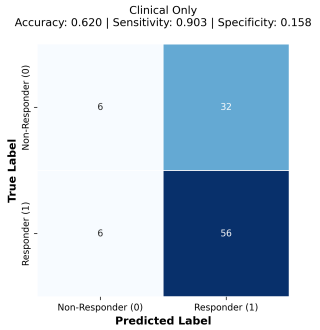
In order to optimize latent feature learning of the autoencoders, we first conducted a large hyperparameter grid search for each of the three models, testing 360, 1,200, and 1,280 configurations for the unsupervised vanilla AE, semi-supervised vanilla AE, and the semi-supervised RVAE models respectively. Hyperparameters included size of the latent dimension, dropout rate, learning rate, batch size, reconstruction vs. prediction weights, and beta (for the RVAE).

Figure 3 reveals a bias toward positive predictions and poor specificity across all six displayed model configurations. The test set contained 38 non-responders and 62 responders. Model specificity ranged from 0.000 to 0.316, indicating severe difficulty in correctly identifying non-responders. The semi-supervised RVAE (Figure 3f) was the most extreme case, classifying every single patient as a responder (38/38 false positives, 0% specificity). Even the best-performing model (semi-supervised AE + clinical, Figure 3b) correctly identified only 12 of 38 non-responders, yielding 26 false positives. This pattern suggests the models learned to predict the majority class rather than discovering meaningful discriminative features. The high sensitivity and accuracy values (0.774-1.000) are largely artifactual, resulting from the models' tendency to classify most or all patients as responders rather than from truly learning to separate the responder groups.

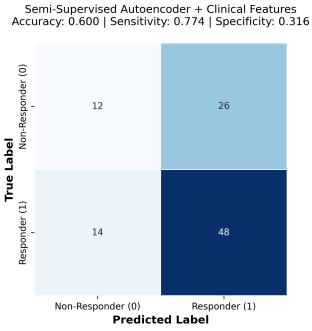
Figure 4 shows 2D PCA visualization of semi-supervised autoencoder latent features. The first two principal components captured 60.2% of total variance (PC1: 40.2%, PC2: 20.0%). Visual inspection reveals moderate clustering distinction between responders and non-responders in the training set, though that distinction disappears in the test set, suggesting overfitting.

5 Discussion

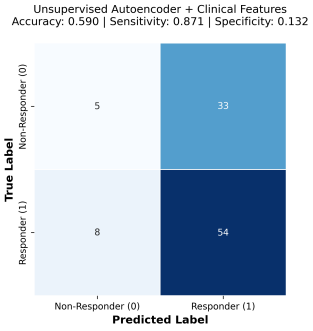
This study investigated whether autoencoder-based latent representations of resting-state EEG, with or without the inclusion of clinical features, could improve prediction of antidepressant treatment response over clinical features alone. While the semi-supervised autoencoder with clinical features yielded the best overall ROC-AUC (0.557), this improvement over the clinical-only baseline (0.513) was modest, and all models performed only slightly above chance. Furthermore, most models demonstrated poor specificity, frequently misclassifying non-responders as responders (Figure 3),



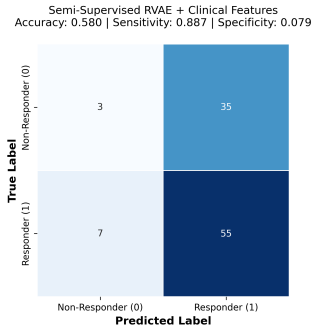
(a) Clinical Features



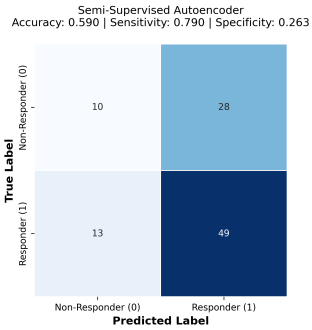
(b) Semi-Supervised AE Latent EEG and Clinical Features



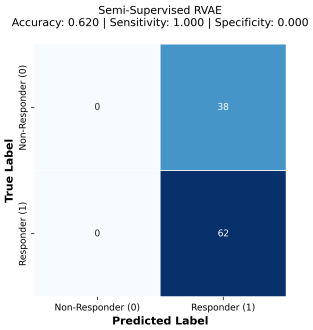
(c) Unsupervised AE Latent EEG and Clinical Features



(d) Semi-Supervised RVAE Latent EEG and Clinical Features



(e) Semi-Supervised AE Latent EEG Features



(f) Semi-Supervised RVAE Latent EEG Feature

Figure 3: Confusion Matrices

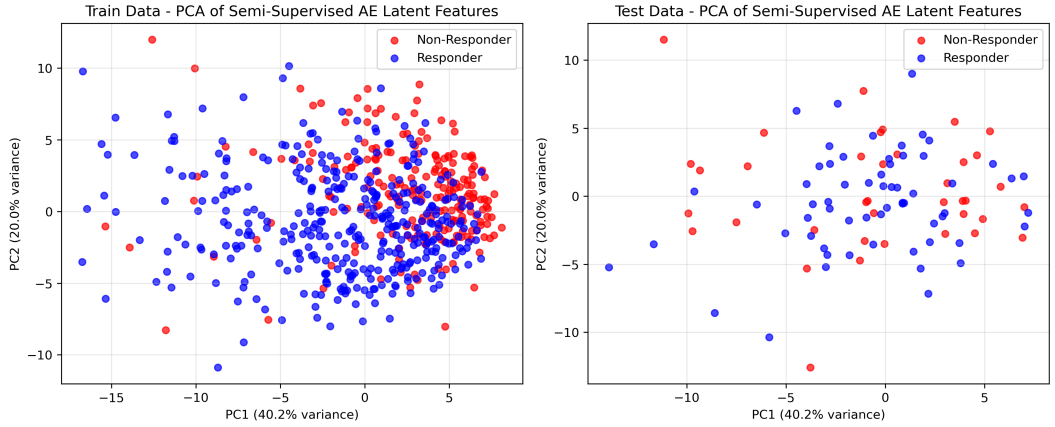


Figure 4: 2-PCA of semi-supervised autoencoder learned latent EEG features

suggesting the models learned to favor the majority class rather than extracting robust discriminative features.

Notably, the addition of EEG features did not consistently improve performance. Two of the three models that combined EEG and clinical inputs (unsupervised AE + clinical; semi-supervised RVAE + clinical) performed worse than the clinical-only baseline. This result contrasts with prior work suggesting that latent EEG representations can enhance predictive performance for antidepressant response (e.g., SELSER, Schwartzmann et al.). One possible explanation is that prior models were trained and evaluated on more homogeneous or placebo-controlled datasets, whereas our use of

a naturalistic, multicenter trial (iSPOT-D) likely introduced greater heterogeneity in both signal quality and patient presentation. Another is that our modeling was insufficiently regularized, yielding overfitting.

The semi-supervised RVAE architecture was designed to take advantage of the temporal structure in EEG recordings, but it did not outperform simpler MLP-based autoencoders. This suggests that either the temporal dynamics present in two minutes of resting EEG were not sufficiently informative for predicting treatment response, or that the model failed to extract relevant temporal features. It is also possible that the latent dimensionality or sequence encoding strategy (116 timepoints $\times$ 26 channels) limited the RVAE's ability to generalize from training to test sets.

The low ROC-AUC values and lack of latent space separability (Figure 4) further suggest that our models failed to isolate a strong EEG-based signal predictive of treatment response. Given that EEG biomarkers of response remain elusive in the broader literature, this may reflect inherent limitations of resting-state EEG for this task. Alternatively, it may indicate the need for more specialized feature engineering, integration of task-based EEG, or subject-specific baselining.

Figure 4 presents 2D principal component projections of the latent EEG features learned by the semi-supervised autoencoder. Together, the first two components capture 60.2% of the total variance in the latent space (PC1: 40.2%, PC2: 20.0%). The lack of clear margins between groups—even in training—also indicates that the learned latent space may not have captured strong response-related structure to begin with, aligning with the near-chance ROC-AUC scores observed in the quantitative evaluation. The relatively high explained variance by the first two PCs also suggests that class-discriminative information, if present, is not simply buried in higher-dimensional components, but likely absent or weak across the latent space as a whole.

Despite these limitations, the model framework introduced here offers a scalable and generalizable pipeline for latent representation learning on EEG in clinical populations. The semi-supervised formulation—combining reconstruction and outcome prediction—provides a flexible way to exploit both labeled and unlabeled data. However, the current results underscore the importance of model calibration and class balance.

Future work might experiment with better RVAE architectures and alternative EEG representations. Here we used broadband power, while more precise might would explore individual frequency bands or connectivity features. Finally, we suspect that additional penalization of overconfident majority-class predictions wold also improve results.

References

Andrade, S. M., da Silva-Sauer, L., de Carvalho, C. D., de Araújo, E. L. M., Lima, E. d. O., Fernandes, F. M. L., Moreira, K. L. d. A. F., Camilo, M. E., Andrade, L. M. M. d. S., Borges, D. T., da Silva Filho, E. M., Lindquist, A. R., Pegado, R., Morya, E., Yamauti, S. Y., Alves, N. T., Fernández-Calvo, B., & de Souza Neto, J. M. R. (2023). Identifying biomarkers for tDCS treatment response in alzheimer's disease patients: A machine learning approach using resting-state EEG classification [Publisher: Frontiers]. *Frontiers in Human Neuroscience*, *17*. https://doi.org/10.3389/fnhum.2023.1234168

Arns, M., Bruder, G., Hegerl, U., Spooner, C., Palmer, D. M., Etkin, A., Fallahpour, K., Gatt, J. M., Hirshberg, L., & Gordon, E. (2016). EEG alpha asymmetry as a gender-specific predictor of outcome to acute treatment with different antidepressant medications in the randomized iSPOT-d study. *Clinical Neurophysiology*, *127*(1), 509–519. https://doi.org/10.1016/j.clinph.2015.05.032

Chen, B., Jiao, Z., Shen, T., Fan, R., Chen, Y., & Xu, Z. (2023). Early antidepressant treatment response prediction in major depression using clinical and TPH2 DNA methylation features based on machine learning approaches. *BMC Psychiatry*, *23*, 299. https://doi.org/10.1186/s12888-023-04791-z

Cohen, S. E., Zantvoord, J. B., Wezenberg, B. N., Bockting, C. L. H., & van Wingen, G. A. (2021). Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: A systematic review and meta-analysis [Publisher: Nature Publishing Group]. *Translational Psychiatry*, *11*(1), 1–10. https://doi.org/10.1038/s41398-021-01286-x

Jankowsky, K., Krakau, L., Schroeders, U., Zwerenz, R., & Beutel, M. E. (2024). Predicting treatment response using machine learning: A registered report [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjc.12452]. *British Journal of Clinical Psychology*, *63*(2), 137–155. https://doi.org/10.1111/bjc.12452

Sargent, K., Chavez-Baldini, U., Master, S. L., Verweij, K. J. H., Lok, A., Sutterland, A. L., Vulink, N. C., Denys, D., Smit, D. J. A., & Nieman, D. H. (2021). Resting-state brain oscillations predict cognitive function in psychiatric disorders: A transdiagnostic machine learning approach. *NeuroImage: Clinical*, *30*, 102617. https://doi.org/10.1016/j.nicl.2021.102617

Schwartzmann, B., Dhami, P., Uher, R., Lam, R. W., Frey, B. N., Milev, R., Müller, D. J., Blier, P., Soares, C. N., Parikh, S. V., Turecki, G., Foster, J. A., Rotzinger, S., Kennedy, S. H., & Farzan, F. (2023). Developing an electroencephalography-based model for predicting response to antidepressant medication. *JAMA Network Open*, *6*(9), e2336094. https://doi.org/10.1001/jamanetworkopen.2023.36094

Sinyor, M., Schaffer, A., & Levitt, A. (2010). The sequenced treatment alternatives to relieve depression (STAR*d) trial: A review [Publisher: SAGE Publications Inc]. *The Canadian Journal of Psychiatry*, *55*(3), 126–135. https://doi.org/10.1177/070674371005500303

Van Mens, K., Lokkerbol, J., Wijnen, B., Janssen, R., de Lange, R., & Tiemens, B. (2023). Predicting undesired treatment outcomes with machine learning in mental health care: Multisite study. *JMIR Medical Informatics*, *11*, e44322. https://doi.org/10.2196/44322

Watts, D., Pulice, R. F., Reilly, J., Brunoni, A. R., Kapczinski, F., & Passos, I. C. (2022). Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis [Publisher: Nature Publishing Group]. *Translational Psychiatry*, *12*(1), 1–18. https://doi.org/10.1038/s41398-022-02064-z

Williams, L. M., Rush, A. J., Koslow, S. H., Wisniewski, S. R., Cooper, N. J., Nemeroff, C. B., Schatzberg, A. F., & Gordon, E. (2011). International study to predict optimized treatment for depression (iSPOT-d), a randomized clinical trial: Rationale and protocol. *Trials*, *12*(1), 4. https://doi.org/10.1186/1745-6215-12-4

Wu, W., Zhang, Y., Jiang, J., Lucas, M. V., Fonzo, G. A., Rolle, C. E., Cooper, C., Chin-Fatt, C., Krepel, N., Cornelssen, C. A., Wright, R., Toll, R. T., Trivedi, H. M., Monuszko, K., Caudle, T. L., Sarhadi, K., Jha, M. K., Trombello, J. M., Deckersbach, T., … Etkin, A. (2020). An electroencephalographic signature predicts antidepressant response in major depression [Publisher: Nature Publishing Group]. *Nature Biotechnology*, *38*(4), 439–447. https://doi.org/10.1038/s41587-019-0397-3