Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Classifying Alcoholism from Electroencephalography Data using High-Dimensional Penalized Regression

Jong Won Lee† and Nathaniel E. Helwig [*‡]

†Department of Psychology, University of Minnesota, Minneapolis, MN, 55455, USA
‡Department of Psychology and School of Statistics, University of Minnesota, Minneapolis, MN, 55455, USA
*Corresponding author. Email: helwig@umn.edu

Abstract

We propose a penalized regression framework for binary classification of high-dimensional EEG data. Using open-source data with n = 122 subjects (n_a = 77 alcoholics; n_c = 45 controls) and p = 14592 features (256 time × 57 channel), we demonstrate that our method can (a) reliably classify EEG patterns of alcoholics, and (b) produce solutions with spatiotemporal sparsity without the need to impose prior spatiotemporal information. This allows for the precise identification of which time points and channels best distinguish EEG patterns of controls and alcoholics.

Keywords: Logistic Regression, Random Forest, Regularization, Support Vector Machine, Variable Selection

1. Introduction

Classification of high-dimensional data is a challenging machine learning problem, given that traditional approaches to classification, e.g., discriminant analysis and logistic regression, require having more subjects n than predictors p. Historically, classification with high-dimensional features has required applying some dimension reduction approach, such as principal components analysis (Eckart & Young, 1936; Hotelling, 1933), to reduce the signal before applying classic classification techniques. Such dimension reduction or feature extraction approaches can produce effective classification rules, as long as the relevant part of the signal is retained in the reduced dimensional feature representation. However, guaranteeing that the relevant part of the signal is retained is a challenging problem. Unless a researcher has a priori knowledge about which feature subspaces are relevant (i.e., active) for the classification rule, some data driven technique must be used to identify the reduced dimensional feature set that best distinguishes the groups (Mwangi et al., 2014).

In high-dimensional situations with $n < p$, regularization and variable selection are often helpful for classification—particularly when the features are highly correlated with one another (Zou, 2019). Regularization methods that allow for variable selection have the benefit of determining the reduced dimensional feature set using a data-driven approach that optimizes a single (classification oriented) loss function. This is in contrast to a two-stage approach that performs dimension reduction of the feature set (stage 1) without considering the classification problem (stage 2). In this paper, we apply recent advances in elastic net regression (Helwig, 2025b, 2026) for the binary classification of spatiotemporal electroencephalography (EEG) data. Using a popular open-source EEG dataset (Begleiter, 1995; Kelly et al., 2025), we aim to identify which time points and EEG channels are most relevant for distinguishing EEG signals of controls and alcoholics. Using penalized logistic regression and support vector machines (SVMs), we demonstrate that our approach can develop *interpretable machine learning* rules that compete with state-of-the-art random forest implementations.

2. Penalized Regression Framework

2.1 Overview

Let $y_i \in \{-1, 1\}$ denote a binary response variable with the levels numerically encoded as -1 (failure) and +1 (success), and let $\mathbf{x}_i \in \mathbb{R}^p$ denote the predictor variables (or features) for the i-th observation. To connect the predictor and response variables, consider a linear model of the form

$$g(\mu_i) = \eta_i = \boldsymbol{\beta}^\top \mathbf{x}_i \tag{1}$$

where $\mu_i = E(y_i|\mathbf{x}_i)$ is the conditional expected value of the response variable y_i given the features $\mathbf{x}_i$, the function $g(\cdot)$ is some user-selected link function, which is monotonic and invertible, η_i is referred to as the linear predictor for the i-th observation, and $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ are unknown coefficients (or weights) that need to be learned from training data. In this paper, we consider the identity link function, i.e., $g(\mu) = \mu$, and the logit link function, i.e., $g(\mu) = \log(\frac{\mu}{1-\mu})$.

Given an independent sample of data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we propose to estimate the unknown regression coefficients $\boldsymbol{\beta}$ by minimizing a penalized loss function of the form

$$L_{\lambda,\alpha} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, \eta_i) + \sum_{j=1}^{p} \mathcal{P}_{\lambda,\alpha}(\beta_j) \tag{2}$$

where $\mathcal{L}$ is some loss function that measures the discrepancy between the response y_i and the linear predictor η_i, and $\mathcal{P}_{\lambda,\alpha}$ is some penalty function that measures the size/strength of the coefficients. Note that the loss function is assumed to satisfy $\mathcal{L}(y_i, \eta_i) \geq 0$ for all $y_i \in \{-1, 1\}$ and all $\eta_i \in \mathbb{R}$, and the penalty function is assumed to satisfy $\mathcal{P}_{\lambda,\alpha}(\beta_j) \geq 0$ for all $\beta_j \in \mathbb{R}$. The regularization parameter $\lambda \geq 0$ controls the overall influence of the penalty $\mathcal{P}_{\lambda,\alpha}$, whereas the regularization parameter $\alpha \in [0, 1]$ controls the balance between the L1 and L2 penalty components (later described).

2.2 Loss Functions

Three loss functions are considered in this paper: logistic regression loss, smoothed SVM loss, and squared SVM loss (see Figure 1). The logistic (logit) loss function for $y_i \in \{-1, 1\}$ has the form

$$\mathcal{L}^{(\text{logit})}(y_i, \eta_i) = \log\left(1 + \exp(-y_i\eta_i)\right) \tag{3}$$

which is mathematically equivalent to the log loss function used for logistic regression with $y_i \in \{0, 1\}$. The quadratically smoothed SVM loss function is defined as

$$\mathcal{L}^{(\text{svm1})}(y_i, \eta_i) = \begin{cases} \frac{1}{2\theta}(1 - y_i\eta_i)_+^2 & 1 - \theta < y_i\eta_i \\ 1 - y_i\eta_i - \theta/2 & y_i\eta_i \leq 1 - \theta \end{cases} \tag{4}$$

where $(x)_+ = \max(0, x)$ denotes the positive part of the input, and $\theta > 0$ is an additional tuning parameter that controls the rate of smoothing (Yang & Zou, 2015; T. Zhang, 2004). As $\theta \to 0$, the smoothed SVM loss function approaches the hinge loss function used in SVMs. We set $\theta = 1$ throughout this paper, which is a standard default value for this hyperparameter (Yang & Zou, 2015); note that the smoothed SVM behaves like a combination of the logit and squared SVM loss functions in this case (see Figure 1). Finally, the squared SVM loss function is defined as

$$\mathcal{L}^{(\text{svm2})}(y_i, \eta_i) = (1 - y_i\eta_i)_+^2 \tag{5}$$

which is the second power of the hinge loss function (Yang & Zou, 2015; T. Zhang, 2004).

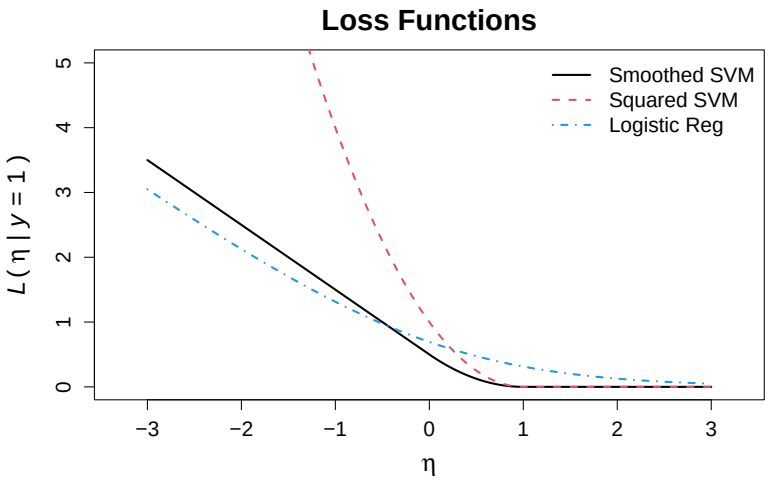

Figure 1. The three loss functions used for binary classification of $y_i \in \{-1, 1\}$. Created using the `visualize.loss()` function in the **grpnet** R package (Helwig, 2026).

2.3 Penalty Functions

The penalty function $\mathcal{P}_{\lambda,\alpha}$ resembles an elastic net penalty (Zou & Hastie, 2005) that linearly combines an L1 and L2 penalty component, such as

$$\mathcal{P}_{\lambda,\alpha}(\beta) = \alpha\mathcal{P}_{1,\lambda}(\beta) + \frac{1-\alpha}{2}\mathcal{P}_{2,\lambda}(\beta) \tag{6}$$

where $\mathcal{P}_{1,\lambda}$ is one of three L1 penalty functions, the least absolute shrinkage and selection operator (LASSO) penalty (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001), or the minimax concave penalty (MCP) (C.-H. Zhang, 2010) (see Figure 2), and $\mathcal{P}_{2,\lambda}(\beta_j) = \lambda\beta_j^2$ is the classic L2 (ridge) penalty (Hoerl & Kennard, 1970). The LASSO penalty is

$$\mathcal{P}_{1,\lambda}^{(\text{lasso})}(\beta) = \lambda|\beta| \tag{7}$$

which is the scaled absolute value of the coefficient. The SCAD penalty has the form

$$\mathcal{P}_{1,\lambda}^{(\text{scad})}(\beta) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda \\ \frac{\gamma\lambda|\beta| - (\beta^2+\lambda^2)/2}{\gamma-1} & \text{if } \lambda < |\beta| \leq \gamma\lambda \\ (\gamma+1)\lambda^2/2 & \text{otherwise} \end{cases} \tag{8}$$

where $\gamma > 2$ is an additional tuning parameter. Finally, the MCP has the form

$$\mathcal{P}_{1,\lambda}^{(\text{mcp})}(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\gamma} & \text{if } |\beta| \leq \gamma\lambda \\ \gamma\lambda^2/2 & \text{otherwise} \end{cases} \tag{9}$$

where $\gamma > 1$ is an additional tuning parameter. For both the SCAD and MCP, we tuned $\gamma \in \{3, 4, 5\}$ throughout this paper, given that values in this range have been shown to perform well (Breheny & Huang, 2011, 2015; Huang et al., 2012). Compared to the LASSO, the SCAD and MCP are designed to impose less bias on large coefficients.

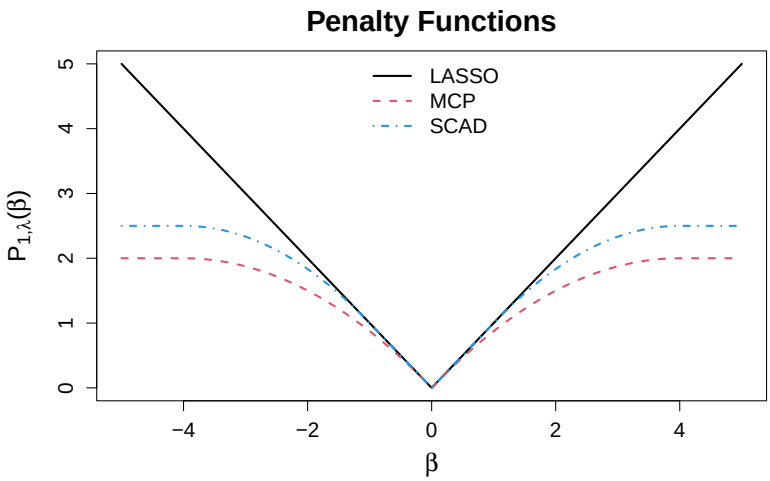

Figure 2. The three L1 penalties used for variable selection with $\lambda = 1$. Created using the `visualize.penalty()` function in **grpnet** R package (Helwig, 2026).

3. Data and Analyses

3.1 Electroencephalography Dataset

The EEG dataset used in this paper was originally collected in Henri Begleiter's neurodynamics lab (Begleiter, 1995) and obtained from the UCI machine learning repository (Kelly et al., 2025). Specifically, we use the dataset named "EEG Database" on the UCI machine learning repository. The dataset contains multiple trials of event related potential EEG recordings from $n = 122$ subjects ($n_a = 77$ alcoholics and $n_c = 45$ controls) who participated in a visual stimulus task. The study used a standard 61-channel EEG cap (plus nose and ears) with CZ serving as the reference channel. The task involved showing participants images from the Snodgrass and Vanderwart stimuli set (Snodgrass & Vanderwart, 1980). For each trial, the subjects were shown an initial image (S1) followed by a second image that was either matching (S2 match) or not matching (S2 nomatch). The subjects were asked to press a mouse key in one hand if the images matched and press a mouse key in the other hand if the images did not match. Each subject participated in 120 trials of the experiment, and only the artifact free data are included in the UCI database. For more details on the experimental setup, see the original article that describes the EEG task (X. L. Zhang et al., 1995).

The original dataset contains EEG recordings from 61 scalp electrodes located at standard sites (Standard Electrode Position Nomenclature, American Electroencephalographic Association 1990). However, for convenience of comparisons to previous analyses, we use a version of the dataset published on GitHub (Mohammed et al., 2020), which contains recordings from 57 electrodes with known spatial coordinates (see Figure 3, left). Each event related potential (ERP) trial was recorded at 256 Hz for one second following the presentation of a visual stimulus, and each subject's data is averaged over all available ERP trials of the S1 condition. Similar to previous analyses of these data (e.g., Helwig, 2016; Helwig & Ma, 2016; Helwig & Snodgress, 2019; Mohammed et al., 2019, 2020), we focus on the initial visual stimulus (i.e., S1), which is not influenced by the subject's response. Previous analysis of these data have shown that alcoholics tend to have reduced P300 activation in occipital and parietal regions (Porjesz & Begleiter, 2003, 1990a, 1990b), i.e., reduced magnitude of the ERP wave about 300 ms after the stimulus (see Figure 3, right). Consequently, we expect the ERP signal in the occipito-parietal and centro-parietal areas of the brain to be influential in distinguishing alcoholics from controls in a temporal window around 300 ms after the stimulus.

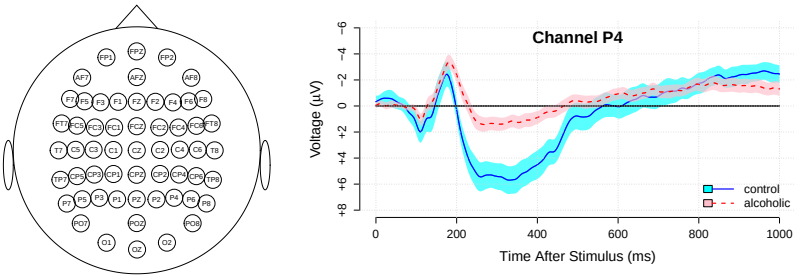

Figure 3. The EEG cap (left) and the average signal at P4 (right). Created using the `eegcap()` and `eegtime()` functions in the **eegkit** R package (Helwig, 2025a).

3.2 Penalized Regression/Classification Modeling

The penalized regression (PR) framework outlined in Section 2 was used to develop a classification rule that connects the EEG data to the group labels (i.e., alcoholic versus controls). Each subject's group label was numerically encoded as either –1 (alcoholic) or +1 (control). Each subject's EEG signal was recorded for 256 time points from 57 channels, which resulted in $p = 14592$ predictor variables available for selection (256 time points $\times$ 57 channels). Note that vectorizing the spatiotemporal data ignores the spatiotemporal structure of the input features, which is a strength (because no prior spatiotemporal assumptions are needed) and a weakness (because some information is lost). As a result, we rely on the penalties to help learn the appropriate spatiotemporal structure: the L1 penalties $\mathcal{P}_{1,\lambda}$ are employed to select which features (i.e., combinations of time points and channels) are useful for classifying the subjects, whereas the L2 penalty $\mathcal{P}_{2,\lambda}$ is used to shrink and stabilize features with high degrees of multicollinearity. All PR models were fit using the **grpnet** R package (Helwig, 2026), which implements an adaptively bounded gradient descent algorithm (Helwig, 2025b).

In addition to the PR framework outlined in Section 2, we also built a classification rule using random forest (RF) models (Breiman, 2001). Note that RF models are ensemble learning methods that aggregate predictions from many (typically hundreds) of decision trees, each of which is fit to a random sample of the data. RF models have the benefit of being extremely flexible and entirely nonparametric, i.e., they do not assume a linear model such as in Equation (1). As a result, RF models are often capable of building prediction rules that perform well for a wide variety of classification and regression tasks, which makes them one of the most popular approaches used in practical applications of machine learning. However, the generality and flexibility of RF models is also their primary drawback; in particular, RF solutions are often difficult to interpret. Furthermore, having a well-fitting RF model does not exclude the possibility that a much simpler (e.g., linear) model would fit the data as well or better. All RF models were fit using the **ranger** R package (Wright & Ziegler, 2017), which is an implementation of RF models for high-dimensional data.

3.3 Hyperparameter Tuning and Model Selection

We used 10 independent replications of a nested cross-validation procedure to tune the model hyperparameters (i.e., λ, α, γ) and select the optimal model (i.e., combination of loss and penalty). For each replication, we used the following procedure. First, stratified sampling was used to split the subjects' data into training and testing sets using a 70/30 split, i.e., 54 alcoholics and 32 controls were randomly sampled for the training set, and the remaining subjects (i.e., 23 alcoholics and 13 controls) were included in the test set. Next, using only the 86 subjects in the training data, 5-fold cross-validation was applied to find the optimal hyperparameters, i.e., those that minimize the average misclassification rate. Finally, the optimal model was used to form class predictions for both the training and testing sets, and the misclassification rates were recorded. Note that all aspects of the

model fitting and tuning were done using only the training data, i.e., the test data were only used to evaluate the chosen model, and the entire procedure (of sample splitting and hyperparameter tuning) was repeated 10 times to assess the stability of the solution.

For the RF models, there are a variety of different hyperparameters and/or options that could be considered. In most applications of RF modeling, the number of trees in the forest (`num.trees`) and the number of features used in each tree (`mtry`) have been shown to have the greatest influence on the quality of the resulting prediction rule. To make things comparable to the hyperparameter tuning and model selection process used for the PR framework, we treated the number of trees as the hyperparameter to tune via k-fold cross-validation (using values of $\{10, 25, 50, 100, 200, 500, 1000\}$), and we compared the results for three different rules used to select the number of features: $q = p^{1/2}$ (square-root), $q = p^{3/4}$ (three-fourths power), and $q = p$ (all features). Note that the square-root rule randomly selects $q = p^{1/2}$ features for each tree, which is the default in most RF softwares (including **ranger**). For each of the considered number of features, the previously discussed training/testing splitting and 5-fold cross-validation procedures were used to tune the hyperparameter (i.e., number of trees) and select the optimal model (i.e., number of features).

3.4 Variable Importance Calculations

For each solution, the importances of the $p = 14592$ variables (256 time points $\times$ 57 channels) were calculated using cosine diagnostics that are frequently used in multiple and generalized nonparametric regression models (Gu, 2013; Helwig, 2020). Let $\mathbf{X}$ denote the $n \times p$ design matrix where the i-th row contains the vectorized version of the i-th subject's EEG data, and let $\boldsymbol{\eta}_{j(\lambda,\alpha)} = \mathbf{X}_j \beta_{j(\lambda,\alpha)}$ denote the j-th term's estimated effect, where $\mathbf{X}_j$ is the j-th column of $\mathbf{X}$ and $\beta_{j(\lambda,\alpha)}$ is the estimator for β_j. The variable importance (VI) for the j-th term is defined as

$$\phi_j = \frac{\boldsymbol{\eta}_{j(\lambda,\alpha)}^{\top} \mathbf{C} \boldsymbol{\eta}_{(\lambda,\alpha)}}{\boldsymbol{\eta}_{(\lambda,\alpha)}^{\top} \mathbf{C} \boldsymbol{\eta}_{(\lambda,\alpha)}} \tag{10}$$

where $\mathbf{C} = (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^{\top})$ is a centering matrix and $\boldsymbol{\eta}_{(\lambda,\alpha)} = \sum_{j=1}^{p} \boldsymbol{\eta}_{j(\lambda,\alpha)}$ is the total effect summed across all p predictors. Note that the VI indices satisfy $\sum_{j=1}^{p} \phi_j = 1$, but the ϕ_j are not required to be non-negative. When all of the ϕ_j are non-negative, the VI indices give the proportion of the variation in the linear predictor that is explained by the j-th term. Note that larger relative values of ϕ_j indicate that the j-th term is more important for forming the predictions, whereas $\phi_j \leq 0$ implies that the j-th term is not needed given the other terms in the model. For interpretational purposes, we scale the VI indices by 100, so that values range from about 0% to 100% explained variation.

For the RF models, the "impurity" importance measure (with the default Gini index) is used to assess the VI of each selected feature. Note that this approach quantifies VI based on the total decrease in the node impurity (as measured by the Gini index) across all data splits, which is then averaged over all of the trees used in the RF. This measure is known to be biased in the sense that it tends to favor continuous measures over binary features; however, given that all of our features are continuous, we are not concerned with such bias issues. Another potential option would be to use the "permutation" approach for assessing VI; however, for high-dimensional predictors, the permutation-based approach requires much more computational work and is not expected to perform much differently. In our opinion, regardless of which approach is used to assess VI in RF models, the primary drawbacks are that (1) importance is often aggregated across trees that contain different features sets, and (2) there is no underlying model with which the importance is being calculated with respect to—so interpreting the meaning of VI in RF is often a challenge. For comparability to the ϕ_j values, the RF impurity importance measures are scaled to sum to 100 across all p features.

4. Results

4.1 Misclassification Error

The top row of Figure 4 displays box plots of the misclassification error in the training (left) and testing (right) sets, with the middle panel showing the tuning error, i.e., the 5-fold cross-validation error rate. More specifically, the top-middle panel shows the average misclassification rates obtained from the five held-out samples within the training set, and this is the panel that is used to select the optimal hyperparameters (λ, α, γ for PR or `num.trees` for RF) and model (loss and penalty for PR or `mtry` for RF). The training (left) and testing (right) panels in the top row are obtained by using the optimal hyperparameters and model (selected from middle panel) to classify the subjects in the training and testing sets, respectively. Thus, the 70% of subjects in the training data are used to construct the left and middle panels, whereas the 30% of subjects in the testing data are only used for the right panel. Focusing on the top-middle panel, all of the PR and RF models produce similar tuning errors (i.e., median misclassification rates around 22-23%) across all of the hyperparameters and modeling choices, which suggests that all models should cross-validate similarly well. When these tuned models are applied to independent test data (top-right), the misclassification errors are similar across for the PR and RF models with median misclassification rates around 19-22%.

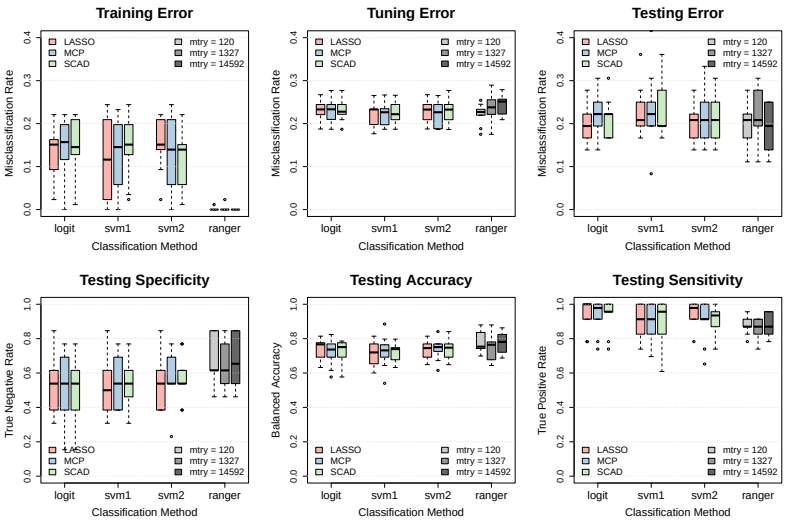

Figure 4. Classification performance for each method in training, tuning, and testing sets.

The bottom row of Figure 4 contains the specificity (left), balanced accuracy (center), and sensitivity (right) for the test data using the optimal hyperparameters/model selected from the tuning error (top-middle panel). In the bottom row, note that specificity is the probability of correctly identifying a control, and sensitivity is the probability of correctly identifying an alcoholic. Although all methods have similar balanced accuracy (median of 72–78%), the PR and RF methods display noteworthy differences in their specificity and sensitivity. Specifically, the PR models have higher sensitivity (i.e., true positive rate for identifying alcoholism), whereas the RF models have higher specificity (i.e., true negative rate for identifying controls). Thus, the bottom row of Figre 4 reveals that the PR models may be more useful if the goal is to identify EEG biomarkers that are useful for distinguishing alcoholics from controls, whereas the RF models may be more useful for developing a classifier that makes fewer false positives (i.e., fewer incorrect classifications of controls as alcoholics).

4.2 Variable Selection

As a reminder, the two primary goals of our analyses are to (1) identify which combinations of time points and channels are most useful for distinguishing the alcoholic and control subjects, and (2) understand the nature in which the two groups' EEG activity differs. To accomplish the first goal, we will spend most of our energy interpreting the VI indices that were introduced in the previous section. However, we begin by examining which variables (i.e., combinations of time points and channels) were selected for use in the prediction rules. Note that for both the PR and RF models, the concept of "variable selection" (VS) is actualized by dichotomizing the VI indices such that a variable is considered selected (i.e., active) if $|\phi_j| > 0$ and is not selected (inactive) otherwise. Thus, VS is a crude (dichotomized) version of VI—but it is important to note that VS is not the same thing as VI. As we will graphically demonstrate, it is possible for a variable to have a high VS index even if the variable is not necessarily important from a VI perspective.

Figure 5 displays heat maps that show the proportion of times (across the 10 replications of the 70/30 sample splitting) that a given variable was selected by each PR and RF model. Note that all of the PR models show a similar heat map, where VS indices are highest for time points in the 200–400 ms range. Furthermore, for the PR models, there tend to be "clusters" of VS that occur for certain channels (e.g., rows 21–27, 45–47, 50–51, etc.), which correspond to electrodes in the central and parietal–occipital regions. Thus, when using the PR models, there tends to be noteworthy consistency in the temporal and spatial natures of the selected variables. In contrast, when looking at the VS for the RF models, there is much less spatiotemporal consistency in the VS results. When using the variable selection rules of $q = p^{1/2}$ and $q = p$, there is a weak tendency for the VS indices to be higher in the 200–400 ms time windows, as well as a weak tendency for VS of certain channels. However, compared to the PR results, the VS results for the RF models appear more sporadic.

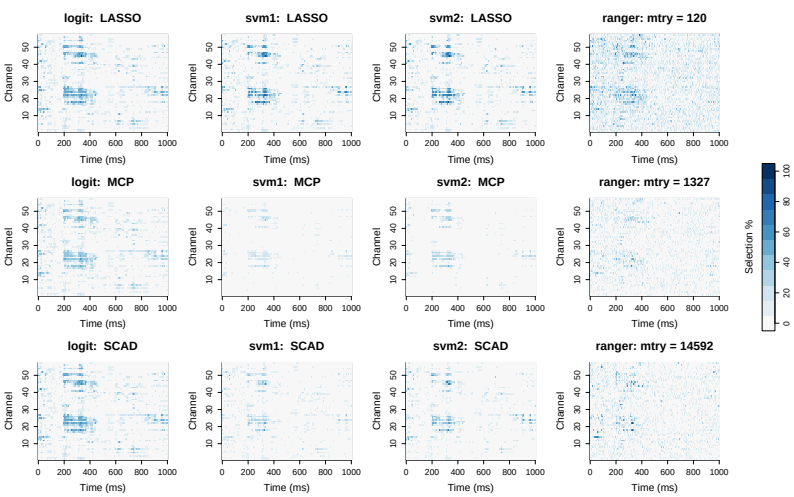

Figure 5. Variable selection percentages by time and channel for each method.

4.3 Variable Importance

Figure 6 displays heat maps of the average (across the 10 replications of the 70/30 sample splitting) VI indices for each variable using each method. At first glance, it may appear that Figure 6 contains nothing but whitespace within each panel—indicating that no variables are important. However, upon careful inspection, it is apparent that some large VI indices exist, but they are sparse with respect to both space and time. For all of the PR models, the important activity is spatiotemporally

localized around time point 86 (≈ 333 ms after stimulus) and channels 22 (P4) and 45 (CP4). In contrast, for the RF models, there are not any clear spikes in the VI indices for any combination of time and/or channel. As a reminder, the RF models use a different approach (i.e., node impurity) to measure VI compare the the PR methods, but the VI indices from all methods are plotted on a common scale (i.e., 0–100%) to facilitate visualize comparisons of which variables are relatively important within each model. Although the VI indices from the PR and RF methods look similarly sparse in Figure 6, noteworthy differences will become apparent when looking at the VI indices using alternative visualization strategies.

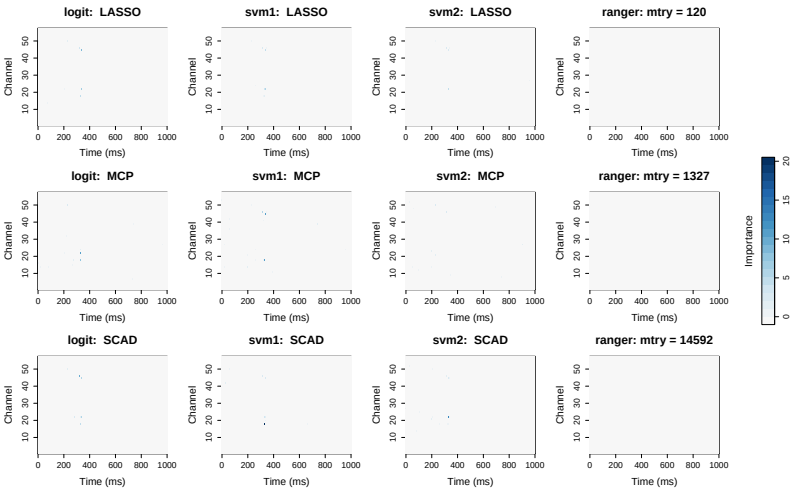

Figure 6. Variable importance indices by time and channel for each method.

For an alternative look at the spatiotemporal sparsity of the VI indices, Figure 7 displays a line plot of VI (y-xais) by time (x-axis) with a separate line drawn for each channel. The line plots reveal that, for the PR models: (1) all of the large VI indices occur in the 85–87 time point window, and (2) the LASSO penalty tends to give approximately equal importance to the CP4 (cyan line) and P4 (magenta line) channels, whereas the MCP and SCAD penalties tend to give more importance to P4. In contrast, for the RF models, there are not any noteworthy spikes in VI for particular time points and/or channels. Instead of spatiotemporally localizing the important activity, the RF models tend to assign relatively small importance to a large number of time/channel combinations. Furthermore, when using the RF models, the VI indices tend to be rather sensitive to the number of features used in each tree, given that the VI profiles for each selection rule differ notably from one another.

Finally, Figure 8 displays the VI indices at time point 86 with the importance mapped to colors for each channel on the EEG cap. Unlike the other two VI figures, note that Figure 8 is only displaying the VI for a single time point—with the goal of clarifying the spatial nature of the important variables. This figure makes it obvious that the PR methods consistently determine that channel P4 (primary effect) and channel CP4 (secondary effect) are the most important locations for distinguishing the controls from the alcoholics. In particular, the MCP penalty quite clearly indicates that there is a strong P4 effect at time point 86 (with a minor effect of CP4). The other two L1 penalties (i.e., LASSO and SCAD) suggest that P4 is the dominant effect, but that CP4 has a noteworthy effect also. In contrast, the RF results do not show any clear spatial pattern in the VI indices at time point 86 (or other time points). Thus, Figure 8 further reveals that the VI indices from the RF models display a rather different pattern than those from the PR models.

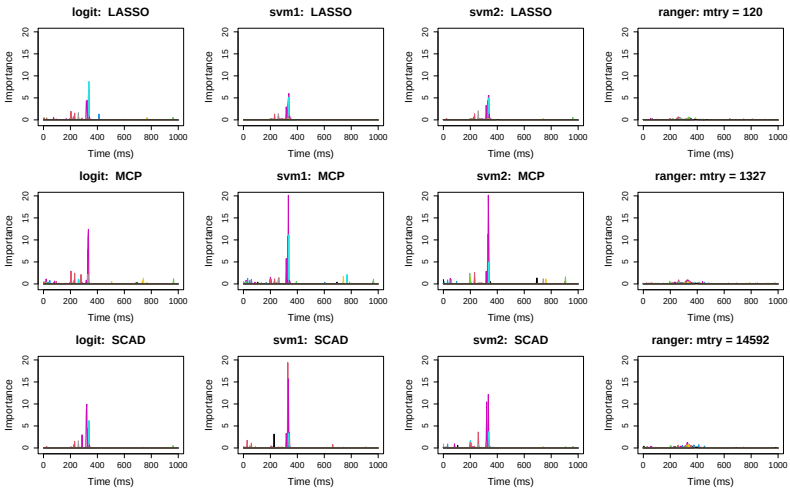

Figure 7. Variable importance indices by time for each channel and method.

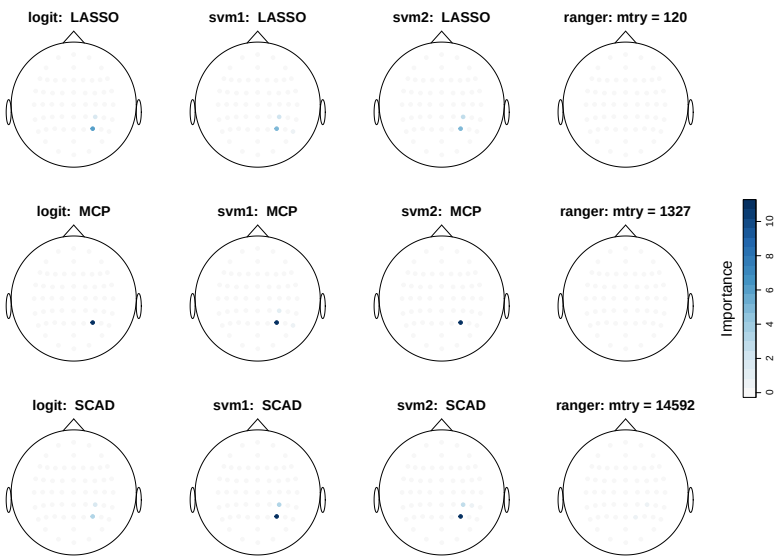

Figure 8. Variable importance indices by channel at time point 86 ($\approx$ 333 ms).

4.4 Interpreting Effects

A noteworthy benefit of the proposed PR framework is that it offers transparent insights into the mechanisms that determine the prediction rules. In particular, by estimating clearly defined model parameters (i.e., the β_j coefficients) that have a straightforward effects on the model predictions, the proposed PR framework makes it possible to easily explore the conditional effect of any model term. For example, Figure 9 displays the estimated linear predictors, i.e., $\boldsymbol{\eta}_{j(\lambda,\alpha)} = \mathbf{X}_j\beta_{j(\lambda,\alpha)}$, for channels P4 (top) and CP4 (bottom) at time point 86 ($\approx$ 333 ms). Note that the predictions are evaluated at an equidistant sequence of voltages that span the range of the observed EEG activity at the corresponding channels. The conditional effect plots in Figure 9 reveal a clear pattern: having a positive voltage at time point 86 in channels P4 and CP4 corresponds to a higher probability of being

classified as a control. These results complement the average signals displayed in Figure 3, and help clarify which spatiotemporal locations are most useful for distinguishing controls from alcoholics.

For the RF models, Figure 9 displays a partial dependence plot (PDP) in an attempt to visualize the effects of the variables that were determined important from the PR methods. However, this approach did not produce very useful interpretations, given that the PDPs for the selected variables showed very little change in the estimated probabilities (< 0.005) across the entire range of predictor values. This is not surprising given that these variables are much less relatively important in the RF models compared to the PR models. Note that, in the PR models, these effects are some of the only important effects in the model, so we see a large change in the model predictions as these effects are varied. In contract, for the RF models, these effects are one of a large number of small effects that combine to make the model predictions, so there is little change in the model predictions as we vary this one input. Furthermore, it is worth noting that these PDPs require extra computational work, whereas the PR effects plots essentially "come for free" as a part of the model fitting process.

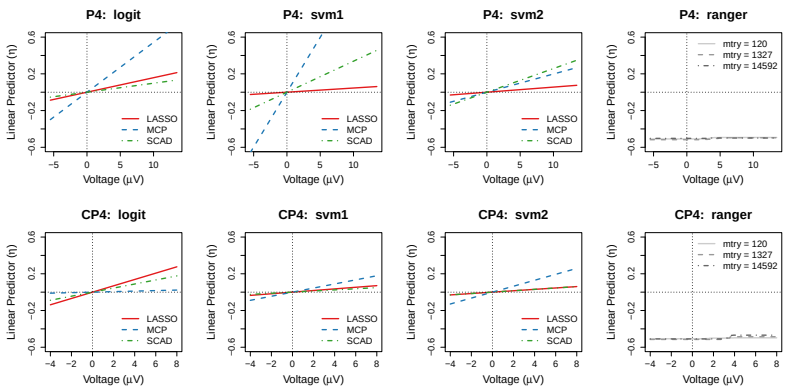

Figure 9. Estimated effects for channels P4 and CP4 at time point 86 ($\approx$ 333 ms).

5. Discussion

5.1 Summary of Findings

In this paper, we demonstrated how PR models can be used for binary classification of event-related potential EEG data collected from alcoholics and controls. Our proposed approach involves vectorizing spatiotemporal EEG data into a high-dimensional predictor vector, and then applying a elastic net estimator (Helwig, 2025b, 2026) to identify which combinations of time points and channels are most useful for distinguishing alcoholics from controls. Using an open-source EEG dataset, we explored the potential of our proposed approach using all combinations of three binary loss functions (logistic regression, smoothed SVM, and squared SVM) and three L1 penalties (LASSO, MCP, and SCAD). We employed multiple replications of a nested cross-validation procedure to explore the sensitivity of our classification and importance quantification results. Furthermore, we compared our proposed PR approach to a popular RF implementation (Wright & Ziegler, 2017) that is frequently used for binary classification with high-dimensional predictors.

Our results revealed that the proposed PR modeling framework produces classification rules that (A) perform as well as RF models when classifying new data (see Figure 4), (B) possess spatiotemporal sparsity in the activations and importances (see Figures 5–8), and (C) provide parsimonious interpretations of the prediction mechanisms (see Figure 9). In particular, across all combinations of loss and penalty functions, the PR models produced solutions where the spatiotemporal activations were in central occipital-parietal areas during the P300 time window, which is expected from

previous analyses of these data (Atiya & Al-Ani, 2009; Helwig, 2016; Helwig & Ma, 2016; Helwig & Snodgress, 2019; Ingber, 1997, 1998; Kuncheva & Diez, 2013; Mohammed et al., 2020; Zhu et al., 2014). In contrast, the RF models produced solutions where the activations and importances lacked spatiotemporal consistency—and were more sensitive to the number of (randomly) selected predictors. Furthermore, the RF models lacked clear insights into the prediction mechanisms, which hindered interpretation. Thus, compared to the RF methods, the proposed PF framework has potential to provide more interpretable machine learning solutions—without sacrificing prediction accuracy.

5.2 Limitations and Future Directions

In this paper, we analyzed the average EEG response across multiple trials, which has been the focus in many past analyses of these data (see previous references). Note that averaging across multiple trials assumes that multiple responses from the same subject are noisy measurements of the same canonical ERP trajectory. If this assumption is reasonable, then averaging the data across trials will produce a better (i.e., less noisy) estimate of the mean ERP than any individual trial. However, in practice, it may be the case that the dynamics and/or variability of the ERPs across trials within subjects contains information that could be used to improve our ability to classify alcoholism from the EEG data. Consequently, future work could consider analyzing the trial-level ERP data (which are available from the raw data source), in order to explore how trial-level characteristics could improve the classification rule, e.g., by including variability across trials for each time/channel as additional features in the classification rule.

Our approach vectorizes the matrix of multi-channel ERP data (256 time points $\times$ 57 channels) into a high-dimensional feature vector (of size $p = 14592$), which does not utilize any information about the spatiotemporal structure. As mentioned in Section 3.2, this could be considered a benefit (because smoothness assumptions are not needed) as well as a drawback (because information is lost). Given that our approach does not utilize spatiotemporal information, it is quite remarkable that the proposed PR method tends to select variables that have a clear spatiotemporal structure. However, it may be the case that incorporating the spatiotemporal information could improve the classification results. In many EEG studies, temporal information is incorporated as a preprocessing step in which the data are filtered at particular frequencies, which could be treated as a hyperparaeter in future applications of our approach. Furthermore, future studies could consider binning the spatiotemporal data and using a group selection rule to select spatiotemporal bins.

5.3 Concluding Remarks

This paper proposed a PR framework for binary classification of high-dimensional EEG data and compared the proposed method to RF binary classifiers. As we demonstrated in this paper, different classification methods can produce similar misclassification results but differ notably in their sensitivity (and specificity) and interpretation. In this paper, we found that the PR models tended to provide greater sensitivity (and reduced specificity) and improved interpretations compared to the RF models—even though both methods produced similar overall classification performance. Although our demonstration focused on EEG data, the approach could be easily applied to other high-dimensional datasets in the psychological and educational sciences. In particular, we expect the proposed approach to be useful for classification of psychologically relevant outcomes (e.g., disease status, symptom presence, milestone met, etc.) from a variety of types biomedical, psychological, and behavioral features. When applying this method, we recommend using a comparative approach (such as in this paper), where results are compared for different combinations of loss functions and penalty options. Comparing the results using different PR modeling options, as well as a different modeling framework such as RF models, helps establish the robustness of the solution and provides a more holistic understanding of the important relationships in the data.

Reproducibility Statement Online supplementary materials contain data and code for replication.

Funding Statement None

Competing Interests None

References

Atiya, A. F., & Al-Ani, A. (2009). A penalized likelihood based pattern classification algorithm. *Pattern Recognition*, *42*, 2684–2694. https://doi.org/10.1016/j.patcog.2009.04.016

Begleiter, H. (1995). EEG Database [DOI: https://doi.org/10.24432/C5TS3D].

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression with applications to biological feature selection. *Annals of Applied Statistics*, *5*(1), 232–253. https://doi.org/10.1214/10-AOAS388

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*, 173–187. https://doi.org/10.1007/s11222-013-9424-2

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 211–218. https://doi.org/10.1007/BF02288367

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Gu, C. (2013). *Smoothing spline ANOVA models* (Second). Springer-Verlag. https://doi.org/10.1007/978-1-4614-5369-7

Helwig, N. E. (2016). Efficient estimation of variance components in nonparametric mixed-effects models with large samples. *Statistics and Computing*, *26*, 1319–1336. https://doi.org/10.1007/s11222-015-9610-5

Helwig, N. E. (2020). Multiple and generalized nonparametric regression. In P. Atkinson, S. Delamont, A. Cernat, J. W. Sakshaug, & R. A. Williams (Eds.), *SAGE Research Methods Foundations*. SAGE. https://doi.org/10.4135/9781526421036885885

Helwig, N. E. (2025a). *Eegkit: Toolkit for electroencephalography data* [R package version 1.0-5]. http://CRAN.R-project.org/package=eegkit

Helwig, N. E. (2025b). Versatile descent algorithms for group regularization and variable selection in generalized linear models. *Journal of Computational and Graphical Statistics*, *34*(1), 239–252. https://doi.org/10.1080/10618600.2024.2362232

Helwig, N. E. (2026). *grpnet: Group elastic net regularized GLMs and GAMs* [R package version 1.1]. https://doi.org/10.32614/CRAN.package.grpnet

Helwig, N. E., & Ma, P. (2016). Smoothing spline ANOVA for super-large samples: Scalable computation via rounding parameters. *Statistics and Its Interface*, *9*, 433–444. https://doi.org/10.4310/SII.2016.v9.n4.a3

Helwig, N. E., & Snodgress, M. A. (2019). Exploring individual and group differences in latent brain networks using cross-validated simultaneous component analysis. *NeuroImage*, *201*, 116019. https://doi.org/10.1016/j.neuroimage.2019.116019

Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441, 498–520.

Huang, J., Breheny, P., & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, *27*(4), 481–499. https://doi.org/10.1214/12-STS392

Ingber, L. (1997). Statistical mechanics of neocortical interactions: Canonical momenta indicators of electroencephalography. *Physical Review E*, *55*, 4578–4593. https://doi.org/10.1103/PhysRevE.55.4578

Ingber, L. (1998). Statistical mechanics of neocortical interactions: Training and testing canonical momenta indicators of EEG. *Mathematical Computer Modelling*, *27*, 33–64. https://doi.org/10.1016/S0895-7177(97)00265-3

Kelly, M., Longjohn, R., & Nottingham, K. (2025). The UCI Machine Learning Repository. https://archive.ics.uci.edu

Kuncheva, L. I., & Diez, J. J. R. (2013). Interval feature extraction for classification of event-related potentials (erp) in eeg data analysis. *Progress in Artificial Intelligence*, *2*, 65–72. https://doi.org/10.1007/s13748-012-0037-3

Mohammed, S., Dey, D. K., & Zhang, Y. (2019). Bayesian variable selection using spike-and-slab priors with application to high dimensional electroencephalography data by local modelling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *68*(5), 1305–1326. https://doi.org/10.1111/rssc.12369

Mohammed, S., Dey, D. K., & Zhang, Y. (2020). Classification of high-dimensional electroencephalography data with location selection using structured spike-and-slab prior. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *13*(5), 465–481. https://doi.org/https://doi.org/10.1002/sam.11477

Mwangi, B., Tian, T., & Soares, J. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*, 229–244. https://doi.org/10.1007/s12021-013-9204-3

Porjesz, B., & Begleiter, H. (2003). Alcoholism and human electrophysiology. *Alcohol Research and Health*, *27*, 153–160. https://pmc.ncbi.nlm.nih.gov/articles/PMC6668890/

Porjesz, B., & Begleiter, H. (1990a). Event-related potentials for individuals at risk for alcoholism. *Alcohol*, *7*, 465–469. https://doi.org/10.1016/0741-8329(90)90033-9

Porjesz, B., & Begleiter, H. (1990b). Neuroelectric processes in individuals at risk for alcoholism. *Alcohol & Alcoholism*, *25*, 251–256. https://doi.org/10.1093/oxfordjournals.alcalc.a044998

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for the naming agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Tibshirani, R. (1996). Regression and shrinkage via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288. https://doi.org/10.2307/2346178

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, *25*, 1129–1141. https://doi.org/10.1007/s11222-014-9498-5

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. https://doi.org/10.1214/09-AOS729

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Proceedings of the Twenty-First International Conference on Machine Learning*, 116. https://doi.org/10.1145/1015330.1015332

Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W., & Litke, A. (1995). Event related potentials during object recognition tasks. *Brain Research Bulletin*, *38*, 531–538. https://doi.org/10.1016/0361-9230(95)02023-5

Zhu, G., Li, Y., Wen, P., & Wang, S. (2014). Analysis of alcoholic eeg signals based on horizontal visibility graph entropy. *Brain Informatics*, *1*, 19–25. https://doi.org/10.1007/s40708-014-0003-x

Zou, H. (2019). Classification with high dimensional features. *WIREs Computational Statistics*, *11*(1), e1453. https://doi.org/https://doi.org/10.1002/wics.1453

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x