CORE-Coma: Deep Learning Framework for Coma Prognosis from Auditory Event-Related Potentials

Elham Bagheri[1,4], Paniz Tavakoli[1], Adianes Herrera-Diaz[1], Rober Boshra[1], Richard Kolesar[2], Alison Fox-Robichaud[3], John F. Connolly[1], James Reilly[4*]

[1] ARiEAL Research Centre, McMaster University, Hamilton, ON L8S 4L8, Canada
[2] Department of Anesthesia, McMaster University, Hamilton, ON L8S 4L8, Canada
[3] Department of Medicine, McMaster University, Hamilton, ON L8S 4L8, Canada
[4] Department of Electrical & Computer Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada
[*] reillyj@mcmaster.ca

Abstract

Accurate prognosis of coma emergence is difficult because bedside behavioral scales can fail to detect residual consciousness. Auditory oddball event-related potentials (ERPs) offer a physiological readout, but single-component markers (e.g., MMN or P3) have limited sensitivity and generalizability. We present CORE-Coma, a deep learning framework for full-waveform ERP analysis, trained exclusively on healthy controls and evaluated zero-shot in coma patients. We analyzed ERPs from 39 healthy controls and 8 coma patients in the intensive care unit (ICU), segmenting EEG recordings into ∼5-minute sub-blocks to capture temporal fluctuations. We define two complementary, model-derived metrics: a time-resolved ERP Separability Score (ESS) and a subject-level Global ERP Separability Index (GESI). Controls showed near-ceiling standard–deviant separability (ROC AUC=0.99), while separability was reduced in coma (ROC AUC=0.68). CORE-Coma identified all patients who emerged from coma (3/3; sensitivity 100%) and 4/5 patients who did not emerge (specificity 80%), yielding accuracy=87.5% (7/8). ESS revealed temporal fluctuations (waxing–waning) of responsiveness in coma at ∼5-minute resolution, absent in controls. SHAP explanations localized influential features, including frontocentral electrodes and time windows consistent with canonical oddball components: 100–150 ms (N1/MMN) and 270–370 ms (P3a/P3b). By combining bedside-feasible scalp EEG with time-resolved and subject-level metrics, CORE-Coma offers an etiology-agnostic approach to coma prognosis. Prospective multicenter studies are needed to validate performance and support clinical deployment.

Introduction

Coma is a deep state of prolonged unconsciousness with a complete absence of wakefulness and volitional behavior. Detecting and characterizing consciousness in comatose patients is challenging due to the lack of reliable verbal and motor output, yet doing so is essential for goal setting and family counseling, rehabilitation planning, and resource allocation. Current bedside assessments based on behavior, such as the Glasgow Coma Scale (GCS) (Teasdale and Jennett 1974) and the Coma Recovery Scale–Revised (CRS–R) (Giacino, Kalmar, and Whyte 2004), can misclassify patients, with misdiagnosis rates approaching 40% (Schnakers 2020; Wannez et al. 2017; Bodien et al. 2020; Donnelly and McCulloch 2014; Turgeon et al. 2011).

Data-driven approaches aim to enhance prognosis by uncovering latent structures in neurophysiology. Initial studies utilized classical machine learning techniques, such as linear discriminant analysis (LDA), to identify mismatch negativity (MMN) as a predictor for recovery, but sensitivity remained low despite MMN's prognostic relevance (Fischer et al. 1999; Naccache et al. 2004). Recent EEG research employs supervised models, including random forests for reactivity/continuity and ERP components (Claassen et al. 2019), as well as convolutional neural networks (CNNs) for analyzing continuous EEG and auditory paradigms (Tjepkema-Cloostermans et al. 2019; Aellen et al. 2023). However, the reliance on narrow hand-crafted markers, limited interpretability, and etiology-specific cohorts hinders generalization. Beyond EEG, neuroimaging pipelines using support vector machines (SVMs) with independent component analysis (ICA) on resting-state functional MRI (fMRI) have identified connectivity signatures of disorders of consciousness (Zhang et al. 2022), and graph-based ML has extended this line of work (Di Gregorio et al. 2022). Multimodal models combining EEG and fMRI or behavioral data with multi-layer perceptrons (MLPs) show promise (Amiri et al. 2023; Rohaut et al. 2024), but intensive care unit (ICU) deployment is hindered by cost, throughput, and single-subject reproducibility challenges (Marek et al. 2022). These constraints motivate EEG-first solutions that are interpretable, temporally resolved, and scalable at the bedside.

The auditory oddball paradigm, repetitive standard tones interspersed with infrequent deviants, is well-suited to comatose patients because it elicits ERPs without active attention or motor output (Duncan et al. 2009). Canonical components include N1 (∼100 ms) and MMN (∼150 ms) reflecting sensory memory and change detection (Näätänen, Gaillard, and Mäntysalo 1978; Näätänen et al. 2007; Näätänen, Kujala, and Light 2019), and later P3a/P3b (∼300 ms) indexing novelty and higher-order attentional processes (Blain-Moraes et al. 2016; Tavakoli et al. 2019). Critically, several of these components depend on residual consciousness,

making them attractive physiological readouts.

MMN presence correlates with emergence and recovery (Morlet and Fischer 2014; Fischer et al. 1999; Armanfard et al. 2018). Fischer et al. (1999) reported $>$ 91% return to consciousness with MMN and $>$ 90% non-awakening without MMN, yet only $\sim$ 30% of emergent patients expressed MMN (low sensitivity). Subsequent studies confirmed high specificity and positive predictive value (PPV) (Fischer et al. 2004) but persistent sensitivity limits ($\sim$ 56% at 1 month (Naccache et al. 2004), $\sim$ 32% at 12 months (Luaute et al. 2005)). Thus, approaches that use the full ERP waveform and its temporal dynamics, rather than a single component, may yield more sensitive biomarkers.

We hypothesize that standard–deviant distinguishability in auditory ERPs reflects residual neural responsiveness and can inform prognosis. To address the limited sensitivity of single-component markers, we adopt control-only training with zero-shot evaluation (source-only domain generalization) on coma patients, modeling the entire ERP waveform rather than hand-crafted features. Our proposed framework, the Control-trained Responsiveness Estimator for Coma (CORE-Coma), quantifies this distinguishability across channels and deviant types via two metrics: the time-resolved ERP Separability Score (ESS) and the subject-level Global ERP Separability Index (GESI). To capture temporal dynamics, recordings are segmented into non-overlapping $\sim$5-minute sub-blocks for tracking responsiveness. For interpretability, we use SHAP to attribute contributions across time, channels, and stimulus type. Unlike prior ERP prognosis centered on MMN/P3 or hand-crafted EEG features, CORE-Coma learns full-waveform representations from controls and produces time-resolved and subject-level metrics suitable for bedside use.

Methods

EEG Data and Preprocessing

The study was approved by the Hamilton Integrated Research Ethics Board (HiREB), no. 4840, and conducted in accordance with the Declaration of Helsinki. This research is part of an initiative aiming to develop a point-of-care system for automated coma prognosis (Connolly et al. 2019). We used a dataset consisting of 39 control subjects and 8 coma patients in the ICU. EEG data were recorded from 64 channels for controls and 8 to 64 channels for patients using the BioSemi ActiveTwo system. A reduced number of electrodes was used when the presence of surgical wounds, skull defects, or surgical drains (e.g., extraventricular drain) interfered with electrode placement. Bilateral mastoids served as a reference. Interelectrode impedances were kept below $10\,\mathrm{k}\Omega$. EEG activity was band-pass filtered online between 0.01–100 Hz. The physiological data were digitized continuously at a 512 Hz sampling rate.

For patients, a 24-hour ERP recording session was conducted immediately following informed consent by the substitute decision maker at the patient's bedside in the ICU at Hamilton General Hospital, a regional neurotrauma center. For controls, ERP sessions of up to 12 hours were conducted at McMaster University's Language, Memory, and Brain (LMB) Lab. Informed consent from each subject was provided before the collection of data.

Auditory stimuli were presented binaurally using Etymotic ER-1 insert earphones and Presentation software (Neurobehavioral Systems). Given the dynamic nature of brain states in coma, extended recording periods were used to capture temporal fluctuations in brain state. Stimuli consisted of standard tones (0.8 probability of occurrence, 800 Hz, 80 dB SPL, 75 ms duration, 5 ms rise–fall time), and three deviant types: (1) duration deviants (Dur; 0.14 probability, 30 ms duration), (2) the subject's own name (SON; 0.03 probability), and (3) environmental novel sounds (Env; 0.03 probability; e.g., dog bark, doorbell). In a subset of controls, novel sounds consisted only of a dog bark; in the other subset, a different novel sound was presented on each trial (e.g., 60 in total). Independent-samples t-tests revealed no significant differences in MMN or P3a amplitudes across these subsets ($p > 0.05$), so data from all controls were combined. SON was synthesized using the ReadSpeaker software to simulate a female native speaker of American English in a neutral voice. Stimuli were presented pseudo-randomly with the constraint that each deviant or novel stimulus was preceded by at least two standard tones. EEG was recorded in blocks of 2000 stimuli, with a stimulus onset asynchrony (SOA) of 800 ms for standards and 1220 ms for deviants, yielding block durations of approximately 25 min.

Table 1 summarizes the clinical and demographic characteristics of the coma patients, including the number of recording blocks collected for each patient. Glasgow Coma Scale (GCS) scores were obtained from patient records and used to summarize responsiveness. The GCS is based on eye-opening, motor, and verbal responses, with scores ranging from 3 (unresponsive) to 15 (fully responsive). The GCS scores at the time of recording are reported in this table. Clinical outcomes were assessed 6–7 weeks post-injury. Emergence from coma was defined according to the guidelines presented in (Young 2009). We refer to patients by anonymized IDs P01–P08 throughout.

Data were processed using Brain Products Analyzer 2 (Brain Products GmbH). Continuous EEG was band-pass filtered between 2–20 Hz (24 dB/octave slope). For controls, an infomax independent component analysis (ICA) (Chaumon, Bishop, and Busch 2015; Makeig et al. 1996) was used to identify eye movement artifacts that were statistically independent of the EEG activity and then removed from the EEG traces. Visually noisy channels were replaced via spherical spline interpolation of the surrounding electrode sites (Perrin et al. 1989). The continuous data were reconstructed into discrete single-trial 700-ms ERPs, beginning 100 ms before stimulus onset. The average activity in the 100 ms pre-stimulus period served as a zero-voltage baseline. Epochs in which EEG activity exceeded $\pm 100\,\mu\mathrm{V}$ relative to baseline were excluded. There was little variation in the rejection of trials across stimuli.

After exporting single-trial ERP epochs from Analyzer 2, subsequent analyses were implemented in Python using MNE, scikit-learn, and TensorFlow/Keras. For modeling, we used the 0–500 ms post-stimulus window (256 samples at 512 Hz).

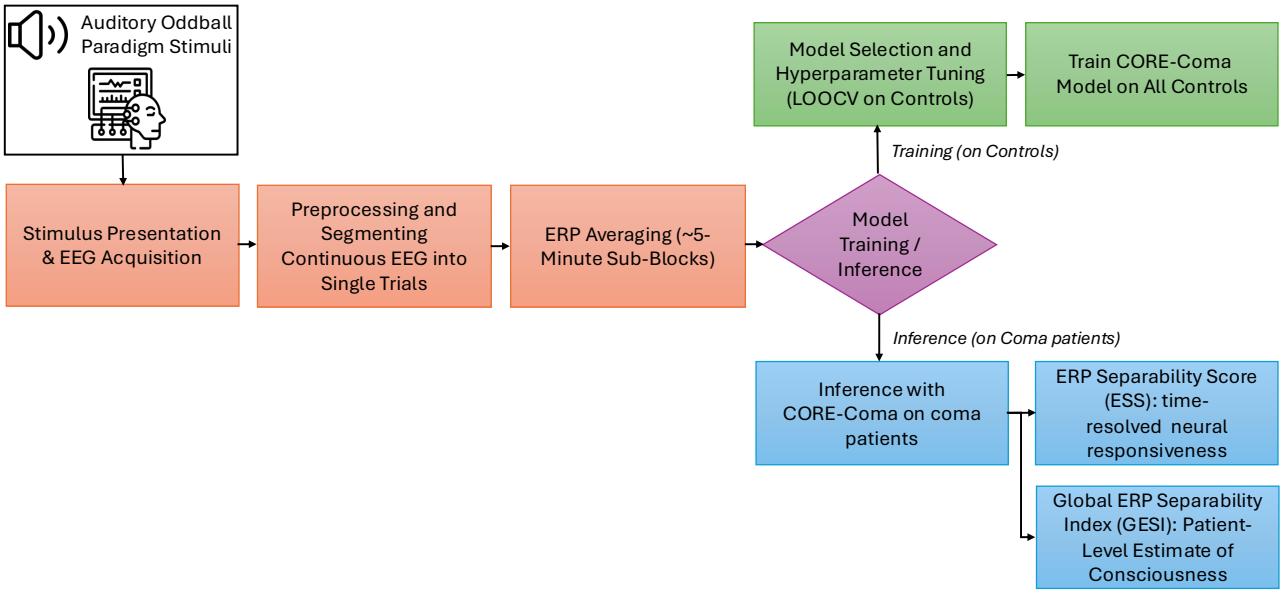

Figure 1: Workflow of CORE-Coma. Auditory oddball ERPs are preprocessed and averaged in non-overlapping ∼5-minute sub-blocks. A multi-branch 1D CNN trained on healthy controls outputs a time-resolved ERP Separability Score (ESS) per sub-block; aggregating across sub-blocks yields the subject-level Global ERP Separability Index (GESI). Model selection uses leave-one-subject-out (LOOCV) on controls; coma data were held out for inference only.

Data Preparation Each block was divided into non-overlapping sub-blocks of approximately 5 minutes. Traditional ERP studies average across the entire block to improve signal-to-noise ratio (SNR). In our approach, averaging within shorter sub-blocks captures temporal fluctuations in brain activity that can be obscured by long-interval averaging, which is particularly relevant in dynamically changing ICU brain states. This retains the SNR benefits of conventional ERP averaging (Luck 2014) while providing substantially finer temporal resolution and enabling time-resolved tracking of neural responsiveness.

To ensure stable averages, sub-block durations were adjusted so that at least 10 deviants of each type occurred within each sub-block. Single-trial ERP responses of the same deviant type and channel within a sub-block were averaged, yielding one standard and three deviant averaged waveforms per sub-block per channel. As shown in Table 1, the number of recorded blocks varied over patients due to clinical procedures and/or excess noise induced into the recordings as a consequence of gel drying or patient movement. Due to the random occurrence of the deviant stimuli, the number of sub-blocks per block varies slightly around a nominal value of five. The number of available channels varied from 8 to 64. We restricted analyses to the eight channels common across all subjects: Fz, F3, Cz, C3, C4, Pz, P3, and P4. These electrodes are salient for standard–deviant distinguishability: MMN typically appears strongly over fronto-central regions (Duncan et al. 2009); N1 is reliably detected at Fz and Cz (Luck 2014); and P3a/P3b often show strong parietal amplitudes, especially at Pz (Polich 2007).

Model Design

Overview Our framework learns standard–deviant distinguishability from the entire ERP waveform using a multi-branch 1D convolutional neural network (CNN) followed by a multi-layer perceptron (MLP) aggregation head. Unless otherwise noted, the hyperparameters below were selected via nested LOOCV on controls.

Inputs Each sub-block yields averaged waveforms for the eight channels (Fz, F3, Cz, C3, C4, Pz, P3, P4) and four stimulus types (one standard, three deviants). Each waveform spans 500 ms and is sampled at 512 Hz (256 samples).

Parallel CNN Branches There are 24 parallel branches in total (8 channels $\times$ 3 deviant types). For each branch, the channel's deviant waveform is paired with the matched standard waveform from the same sub-block; the branch is trained to discriminate deviant from standard for that channel. Thus, each CNN learns to distinguish its deviant from the corresponding standard on the same channel.

CNN Architecture Each CNN comprises three 1D convolutional layers with 32, 16, and 8 filters, respectively, with kernel size 3 and stride 1. Each convolution is followed by a rectified linear unit (ReLU) activation and max pooling with pool size 2 and stride 2. The resulting feature maps are flattened and passed to a fully connected layer with 8 units (ReLU), followed by a 1-unit sigmoid output. The branch output is a scalar in $[0, 1]$ that reflects control-like standard–deviant separability for that channel–deviant pair.

Patient ID	Sex	Age	Blocks	Etiology	GCS	Outcome
P01	F	69	6	Anoxia	Missing	Death
P02	M	56	9	Trauma	4	Death
P03	M	21	9	Trauma	5	Emergence
P04	M	29	16	Trauma	4	Emergence
P05	M	18	1	Anoxia	3	Death
P06	F	42	7	Neurosurgery	5	Death
P07	F	52	7	Neurosurgery	5	Emergence
P08	F	43	4	Trauma	4	UWS

Table 1: Demographic and clinical information for coma patients. Outcome refers to 6–7 weeks post-injury; "emergence" was defined according to (Young 2009). UWS: unresponsive wakefulness syndrome. GCS: Glasgow Coma Scale score at recording. Blocks: number of recording blocks per patient. The GCS score for patient P01 was not available from the medical record.

MLP Aggregation Head and Scoring The 24 branch outputs are concatenated and passed to an MLP regressor comprising a 32-unit ReLU hidden layer and a 1-unit sigmoid output layer. For each sub-block, the model outputs a continuous score in $[0, 1]$. Both the CNN branches and the MLP were trained with labels: standard $= 0$, deviant $= 1$.

Training and Model Development All training was performed exclusively on healthy control data; no coma data or outcome labels were used for training, validation, or model selection. This design uses control-only training so the network learns characteristic control patterns of standard–deviant separability. At test time, the control-trained model is applied to coma ERPs. Figure 1 illustrates the workflow. This constitutes zero-shot domain generalization, that is, source-only training with an unchanged label space (standard versus deviant), a distribution shift between source (controls) and target (coma), and no target-domain adaptation or tuning. In other words, we perform out-of-distribution (OOD) generalization.

Model Selection Model architecture and hyperparameters were selected via nested cross-validation with subject-wise leave-one-subject-out (LOOCV) as the outer loop on the control cohort. Within each outer fold, we reserved 10% of the training data as a validation split for early stopping and a grid search over architecture and training parameters, including, but not limited to, filters per convolutional layer, number of hidden units/layers, kernel size, convolution stride, max-pooling size/stride, learning rate, and batch size. Splits were subject-disjoint to prevent leakage. The held-out control subject served for evaluation; we averaged performance across folds and selected the configuration with the best LOOCV mean. Training ran up to 500 epochs with early stopping (patience 25) on the validation loss (Prechelt 2012; Zhang et al. 2021). Optimization used Adam with mean squared error (MSE) loss (Kingma and Ba 2014). The learning rate was tuned in the inner-loop grid search and selected as 10^{-3}. After model selection, we retrained the chosen configuration on all control data and used it for zero-shot inference on coma patients. Throughout, ERPs were normalized using training-set statistics, and the same transform was applied to validation and test data.

Metrics: ESS and GESI

We use distinguishability to describe the physiological notion of how different standard and deviant ERPs are; operationally, we quantify this with the ML notion of separability between their averaged waveforms, and define two metrics accordingly: the time-resolved ERP Separability Score (ESS) and the subject-level Global ERP Separability Index (GESI). ESS is a time-resolved measure of neural responsiveness, computed for each sub-block. GESI summarizes separability across all sub-blocks for a subject to provide a global prognosis. We treat deviants as the positive class (1) and standards as the negative class (0) throughout.

For each sub-block, averaged standard and deviant waveforms are processed through the 24 CNN branches. Each CNN produces a scalar in $[0, 1]$, with larger values indicating more control-like standard–deviant separability for that channel–deviant pair. The 24 outputs are concatenated and passed to the MLP to yield a decision score in $[0, 1]$ for each input. We refer to this score as ESS when computed on deviant (positive-class) waveforms; matched standards are used solely as negatives for ranking and ROC AUC. An ESS near 1 indicates control-like separability; values near 0 indicate minimal separability. We hypothesize that ESS serves as a time-resolved proxy for neural responsiveness.

GESI is computed per subject as the area under the receiver operating characteristic curve (ROC AUC) using the model's decision scores for deviant (ESS) and matched standard waveforms pooled across all sub-blocks for that subject (deviant$= 1$, standard$= 0$). It summarizes how consistently the model separates standard and deviant responses across time. GESI captures temporal dynamics of separability, providing a robust, threshold-independent summary of neural responsiveness. Higher GESI values indicate greater similarity to control ERP patterns and, therefore, stronger neural responsiveness, which we hypothesize is associated with a higher likelihood of emergence; lower values reflect decreased responsiveness and may correspond to a lower chance of recovery.

Decision Rule Because GESI is the ROC AUC in $[0, 1]$, we use 0.5 as the chance-level reference. Values > 0.5 indicate above-chance separability (more control-like re-

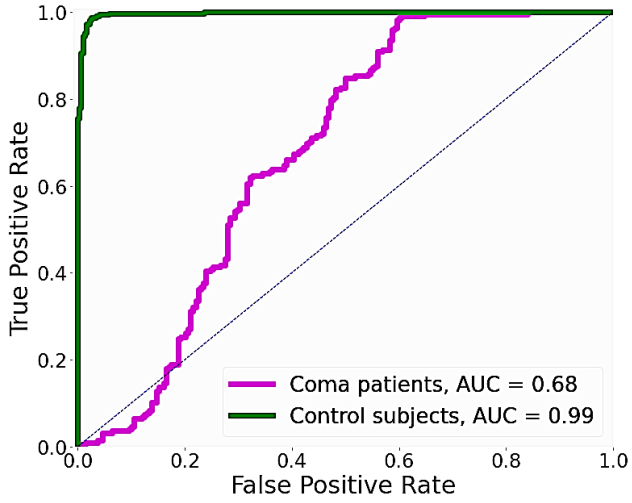

Figure 2: ROC curves for standard–deviant separability at the cohort level in controls and coma patients. Separability is near-ceiling for controls, but reduced for coma.

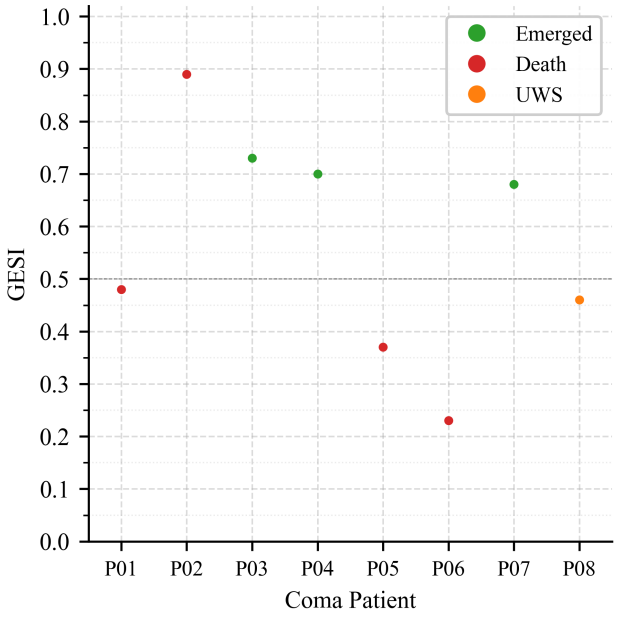

Figure 3: Global ERP Separability Index (GESI) for coma patients. Higher GESI values align with emergence (P02 later died despite an above-threshold score; see Discussion). The dashed line marks the pre-specified decision rule at chance (GESI $= 0.5$). Controls (not shown) are near 1.

sponsiveness); values < 0.5 indicate a systematically inverted ranking (standards tending to outrank deviants), with stronger evidence against control-like responsiveness the farther below 0.5; values near 0.5 reflect weak or ambiguous separability. We fixed 0.5 *a priori* as the operating point, with no outcome-driven tuning.

Zero-Shot Inference

At test time, we apply the control-trained model to coma ERPs with no target-domain data or tuning, consistent with zero-shot domain generalization. We compute ESS for each sub-block and GESI for each subject. Inference on controls is in-distribution; inference on coma is out-of-distribution due to domain shift. This zero-shot protocol can be applied to other clinical cohorts.

SHAP-based Explanations

We use SHAP (Shapley Additive Explanations) to provide post hoc explanations and to support interpretability and transparency (Lundberg and Lee 2017). Prior applications of SHAP to ERP have demonstrated feasibility (Boshra et al. 2019). SHAP values quantify each feature's contribution to a prediction; here, features span ERP time points, channels, and deviant types. In our setting, large-magnitude SHAP values indicate greater influence on the model output. We computed SHAP values using a background distribution drawn from controls. We summarize attributions as mean absolute SHAP ($|\mathrm{SHAP}|$) to quantify influence irrespective of polarity. While SHAP attributions are not causal and time-series feature dependence can affect values, they provide stable, local importance maps that help localize temporal and spatial drivers of the model's decisions in our data.

Results

Standard–Deviant Separability We assess distinguishability using the model's decision scores in $[0, 1]$ for each averaged waveform. ROC curves were computed by pooling scores across all sub-blocks and subjects within each cohort (labels: deviant$= 1$, standard$= 0$); higher AUC indicates stronger separability. For controls, AUC is derived from subject-wise LOOCV; for coma, AUC reflects zero-shot inference from the final model trained on all controls. Across controls, AUC was 0.99; across coma patients, it was 0.68, indicating markedly reduced separability in disordered consciousness. The ROC curves are shown in Figure 2.

Because standards were averaged from more trials than deviants, potentially increasing SNR, we tested whether this drove separability. In controls, we randomly subsampled standard trials to match the deviant count for each channel and deviant pair, averaged them to form matched-standard waveforms, and recomputed AUC. The AUC remained 0.98 (vs 0.99 with all standards), indicating that separability reflects genuine waveform differences rather than SNR.

Coma Outcome Prediction As described in Methods, GESI is a threshold-independent, subject-level summary of overall separability. Figure 3 shows GESI for each patient with emergence labels. Patients with GESI > 0.5 predominantly emerged (3/4), whereas those with GESI ≤ 0.5 did not (4/4). P08 entered unresponsive wakefulness syndrome (UWS), which is not a fully conscious state; for binary outcome analysis, we treated UWS as non-emergent. Controls (not shown) consistently yielded GESI near 1, in line with the control ROC AUC of 0.99. GESI matched the clinical outcome for seven of eight patients; P02 later died despite a GESI greater than 0.5. Using the pre-specified deci-

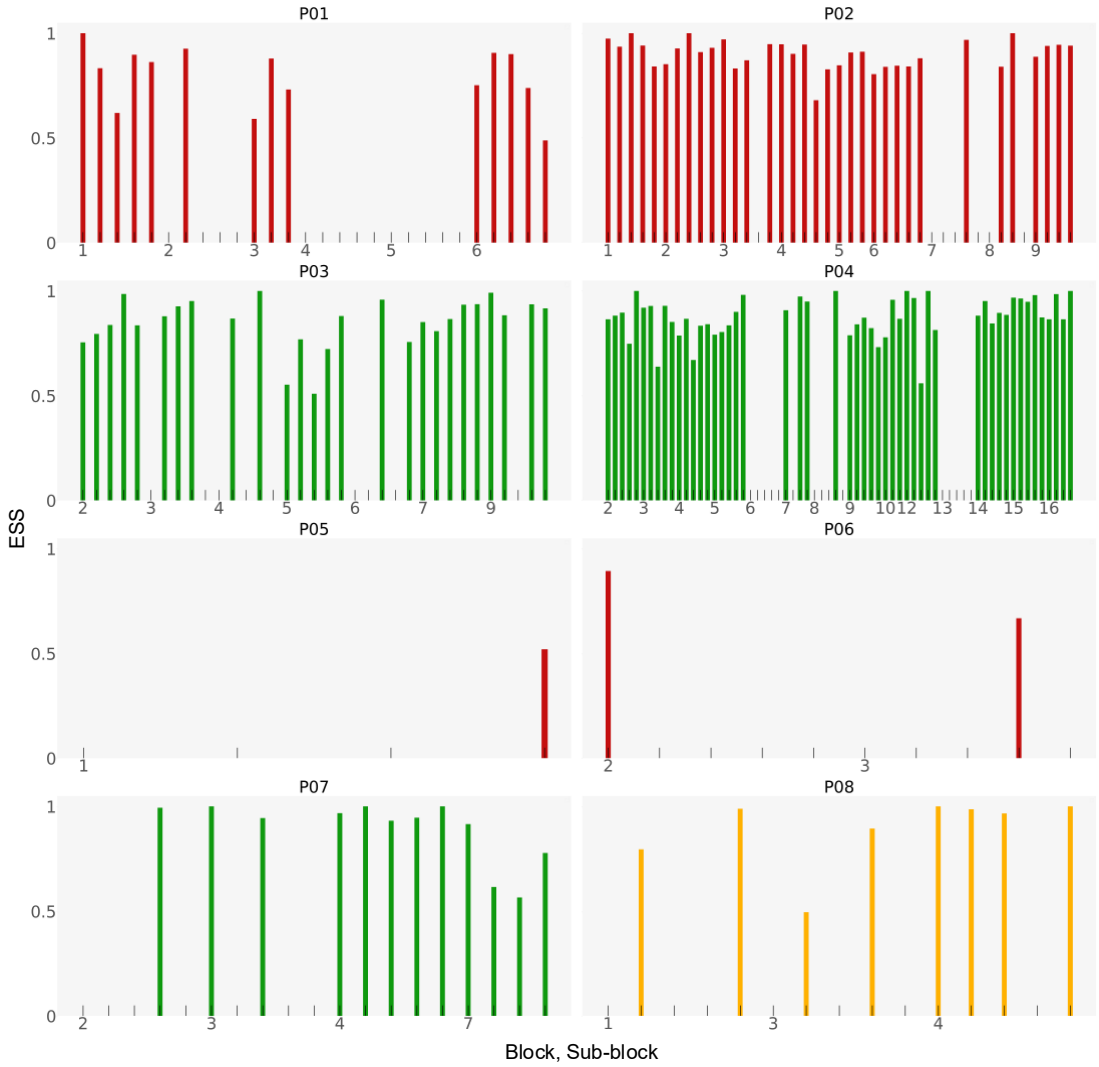

Figure 4: Time-resolved ERP Separability Scores (ESS) for coma patients. Bar colors encode outcomes (green: emerged; red: non-emergent; orange: unresponsive wakefulness syndrome, UWS). Gaps (omitted bars) mark sub-blocks in which the model failed to rank the deviants above the matched standard, indicating intervals of diminished neural responsiveness. Controls (not shown) remain near 1 across time, consistent with the absence of waxing–waning in healthy consciousness.

sion rule at chance (GESI $= 0.5$), the model identified all emergent patients (3/3; sensitivity 100%) and four of five non-emergent patients (specificity 80%), yielding an overall accuracy of 87.5% (7/8). This contrasts with prior MMN-focused work (Morlet and Fischer 2014; Fischer et al. 1999), which reported high specificity but lower sensitivity.

Time-Resolved ESS Trajectories Figure 4 shows ESS over time; each bar denotes a sub-block. Gaps (omitted bars) mark sub-blocks in which separability failed, and the model output for the deviants was lower than the paired standard, indicating intervals of diminished neural responsiveness. All eight patients exhibit irregular waxing–waning at ∼5-minute resolution. Emergent patients show denser and more frequent high-ESS intervals, whereas non-emergent patients remain consistently low. This aligns with prior observations of temporal responsiveness fluctuations (Armanfard et al. 2018; Herrera-Diaz et al. 2023) and suggests that the density of high-ESS intervals may serve as a prognostic marker. The proportion of sub-blocks with elevated ESS provides a compact proxy for moment-to-moment neural responsiveness.

SHAP-based Explanations Figure 5 summarizes mean absolute SHAP contributions by channel and deviant type, showing frontocentral dominance (e.g., Fz, Cz, C3) and higher contributions from the subject's own name (SON) and environmental (Env) deviants relative to duration (Dur) deviants. Complementing this, Figure 6 shows temporal SHAP profiles, highlighting contributions around 100–150 ms (likely associated with N1/MMN) and smaller effects near 270–370 ms (likely associated with P3a/P3b) in controls, with attenuation in coma. A modest late contri-

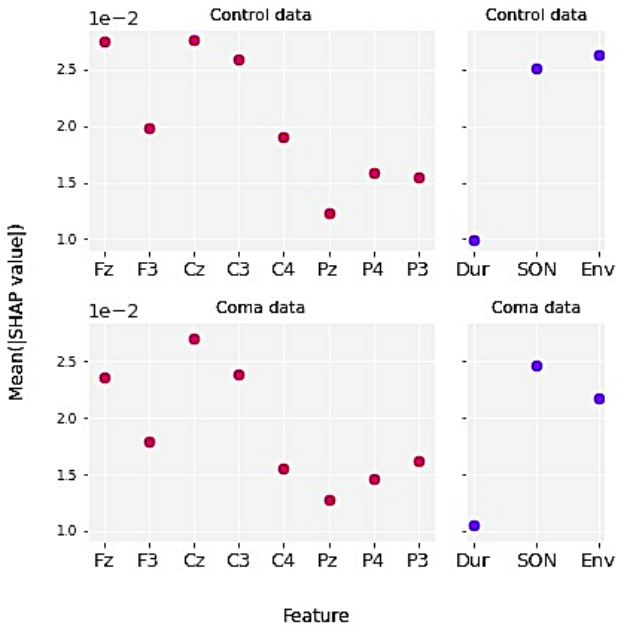

Figure 5: SHAP summary of spatial and stimulus-type contributions. Mean absolute SHAP values aggregated by EEG channel and deviant type show frontocentral dominance (Fz, Cz, C3) and larger contributions from the subject's own name (SON) and environmental (Env) deviants compared with duration (Dur) deviants.

bution was also detectable around 450–500 ms in controls, compatible with a reorienting negativity (RON); this effect was weak or absent in coma. A subtle early rise near 20–40 ms was sometimes evident in controls but inconsistent in coma, and should be interpreted cautiously.

Discussion

Summary of Findings We introduced CORE-Coma, a deep learning framework with control-only training and zero-shot evaluation on coma patients. This framework models the entire event-related potential (ERP) waveform to quantify the distinguishability between standard and deviant stimuli, and it predicts emergence from a comatose state. The approach generates two complementary outputs: a time-resolved ERP Separability Score (ESS) for approximately 5-minute sub-blocks and a subject-level Global ERP Separability Index (GESI). In our study (39 controls and 8 coma patients), CORE-Coma correctly identified all emergent patients (3/3) and 4/5 non-emergent patients at the prespecified decision rule at chance ($\mathrm{GESI} = 0.5$), yielding 7/8 correct. Additionally, the ESS trajectories showed irregular patterns of waxing and waning responsiveness in the coma patients, a pattern that was not observed in the control subjects. These findings support the core hypothesis that reduced consciousness manifests as decreased distinguishability between standard and deviant auditory ERPs, and that monitoring this signal at bedside can provide useful prognostic information.

Neurophysiological Interpretability The attribution patterns are consistent with canonical oddball physiology: a frontocentral emphasis and stronger weighting of salience-rich deviants (the subject's own name and environmental sounds) relative to the duration deviant. The dominant early window ($\sim$100–150 ms) maps onto N1/MMN generators, with additional contributions around $\sim$270–370 ms compatible with the P3a/P3b complex. Occasional later influence near $\sim$450–500 ms is consistent with a reorienting negativity (RON) response (Justo-Guill'en et al. 2019; Escera and Corral 2007), and small very-early effects ($\sim$20–40 ms) may reflect subcortical or earliest cortical processing. Taken together, these correspondences suggest the model relies on physiologically meaningful signals and motivate the 0–500 ms analysis window: it captures the most informative early processes (largely within the first $\sim$300 ms) while also allowing later novelty/attention effects and potential post-injury latency shifts to contribute. We emphasize that SHAP offers associative explanations; it is hypothesis-generating rather than causal.

Comparison with Prior Work ERP-based prognostic models often focus on single components (e.g., MMN or P3), yielding high specificity but modest sensitivity. Some methods use hand-crafted EEG features or target specific etiologies (e.g., post cardiac arrest). In contrast, CORE-Coma learns full-waveform representations from healthy controls (control-only training) and applies them zero-shot across heterogeneous coma etiologies. Sub-block analysis increases temporal resolution to reveal fluctuations in responsiveness, complementing block-averaged ERP and continuous EEG. These design choices delivered high sensitivity in our cohort while maintaining interpretability.

Clinical Relevance CORE-Coma uses bedside-feasible scalp EEG with eight common channels to yield (i) ESS for $\sim$5-minute monitoring of neural responsiveness and (ii) GESI as a threshold-independent, subject-level summary. These metrics complement behavioral scales (e.g., GCS), which can misclassify residual consciousness: ESS can prompt targeted probing during waxing intervals, while GESI provides a stable basis for interdisciplinary discussions and goals-of-care. Control-only training reduces reliance on scarce labeled coma datasets and supports ICU deployment. The fully automated pipeline avoids manual ERP component identification (e.g., visual MMN), reducing operator dependence and improving bedside scalability. While validation in larger, multicenter cohorts is needed, the framework is designed to generalize because it relies on control-only training rather than etiology-specific patient labels.

Relationship to GCS at Recording Of the eight patients, three later emerged (P03, P04, P07). All three showed GESI > 0.5 despite low GCS at recording (P04: GCS = 4; P03, P07: GCS = 5). Of the five non-emergent cases, four had GESI ≤ 0.5 (P01: GCS missing; P05: GCS = 3; P06: GCS = 5; P08: GCS = 4). P02 was the one exception: a high GESI despite not surviving (traumatic etiology: motor vehicle collision). We do not infer a proximate cause of death here. These results indicate that GESI reflects residual neural re-

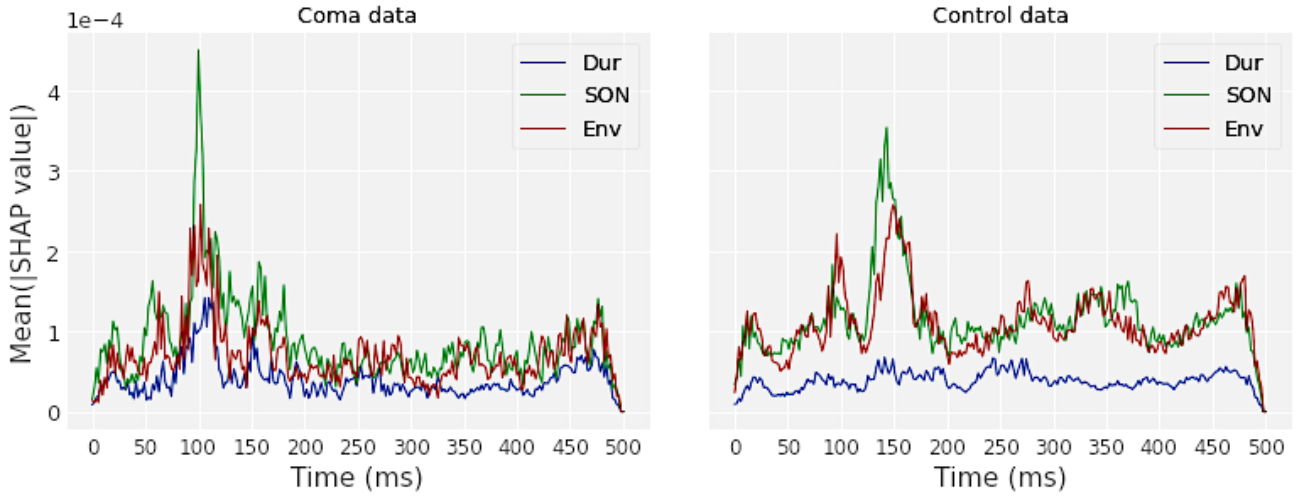

Figure 6: Temporal SHAP profiles averaged across channels for each deviant. Controls exhibit prominent contributions around 100–150 ms (likely associated with N1/MMN) and smaller effects near 270–370 ms (likely associated with P3a/P3b); coma patients show attenuated temporal structure.

sponsiveness, whereas eventual survival also depends on the subsequent clinical course. These comparisons suggest that GESI provides information complementary to contemporaneous behavioral scores. GESI is intended as a physiological index of neural responsiveness to complement, not replace, clinical judgment and behavioral scales.

Limitations and Future Work This was a single-center ICU cohort (8 patients). Around-the-bed 24-hour ERP acquisition, clinical interruptions (imaging, surgery, sedation adjustments), and consent logistics constrained enrollment. Prospective multicenter studies with larger cohorts are needed to test generalizability, calibrate decision thresholds and probability estimates, and assess robustness across clinical variability (sedation, etiology, channel availability). Patient P05 contributed only one block because the patient died during the recording interval, which highlights the practical challenges of ICU data collection. Patient P02 had a high GESI yet did not survive; the etiology was a traumatic motor vehicle collision. This indicates that GESI indexes neural responsiveness rather than mortality risk, and that outcomes can diverge due to medical and systemic factors outside EEG-based responsiveness.

Future work should incorporate additional clinical variables to contextualize metrics. We focused on eight common electrodes; more channels may improve performance, although wounds, skull defects, and drains can limit montage size in the ICU. Because responsiveness fluctuates over time, sessions should be long enough to capture multiple cycles; shorter recordings can bias estimates. Studies should define a waxing-density metric (for example, the proportion of sub-blocks with ESS $\geq \tau$), perform calibration, and quantify incremental prognostic value. Finally, real-time monitoring and extensions to related disorders of consciousness are promising directions.

Conclusions

We introduced CORE-Coma, a deep learning framework trained on healthy controls and evaluated in a zero-shot domain generalization setting on coma patients. CORE-Coma models full-waveform auditory ERPs and quantifies standard–deviant distinguishability, yielding two complementary metrics: the time-resolved ERP Separability Score (ESS) and the subject-level Global ERP Separability Index (GESI). In our cohort, CORE-Coma correctly identified all emergent patients (3/3) and 4/5 non-emergent patients (overall accuracy 87.5%); ESS trajectories revealed waxing–waning responsiveness in coma that was not seen in controls. Model explanations via SHAP were consistent with canonical oddball physiology. Because training uses only controls, the framework is designed to generalize across etiologies, reduce reliance on scarce labeled patient data, and fit bedside workflows. The same zero-shot domain generalization approach can be applied to other clinical cohorts and related disorders of consciousness. Next steps include prospective multicenter validation with larger cohorts, calibration of clinically meaningful decision thresholds and probability estimates, workflow integration for real-time use, and extensions to related disorders of consciousness.

Acknowledgments

The authors thank Dr. Cindy Hamielec for support with data collection at Hamilton General Hospital, Ontario, and Ms. Chia-Yu Lin for assistance in preparing the original grant application and administrative support for data collection. We also acknowledge Hope Morrison for administrative assistance, and Richard Mah and Kiersten Mangold for contributions to data collection. We thank the Language, Memory, and Brain (LMB) Lab and the ARiEAL Research Centre at McMaster University for their support.

References

Aellen, F. M.; Alnes, S. L.; Loosli, F.; Rossetti, A. O.; Zubler, F.; De Lucia, M.; and Tzovara, A. 2023. Auditory stimulation and deep learning predict awakening from coma after cardiac arrest. *Brain*, 146(2): 778–788.

Amiri, M.; Fisher, P. M.; Raimondo, F.; Sidaros, A.; Cacic Hribljan, M.; Othman, M. H.; Zibrandtsen, I.; Albrechtsen, S. S.; Bergdal, O.; Hansen, A. E.; et al. 2023. Multimodal prediction of residual consciousness in the intensive care unit: the CONNECT-ME study. *Brain*, 146(1): 50–64.

Armanfard, N.; Komeili, M.; Reilly, J. P.; and Connolly, J. F. 2018. A machine learning framework for automatic and continuous MMN detection with preliminary results for coma outcome prediction. *IEEE journal of biomedical and health informatics*, 23(4): 1794–1804.

Blain-Moraes, S.; Boshra, R.; Ma, H. K.; Mah, R.; Ruiter, K.; Avidan, M.; Connolly, J. F.; and Mashour, G. A. 2016. Normal brain response to propofol in advance of recovery from unresponsive wakefulness syndrome. *Frontiers in human neuroscience*, 10: 248.

Bodien, Y.; Barra, A.; Temkin, N.; Barber, J.; Foreman, B.; Robertson, C.; Vassar, M.; Manley, G.; Giacino, J.; and Edlow, B. 2020. Diagnosing Level of Consciousness: The Limits of the Glasgow Coma Scale Total Score. *Archives of Physical Medicine and Rehabilitation*, 101(11): e7.

Boshra, R.; Ruiter, K. I.; DeMatteo, C.; Reilly, J. P.; and Connolly, J. F. 2019. Neurophysiological correlates of concussion: deep learning for clinical assessment. *Scientific reports*, 9(1): 17341.

Chaumon, M.; Bishop, D. V.; and Busch, N. A. 2015. A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of neuroscience methods*, 250: 47–63.

Claassen, J.; Doyle, K.; Matory, A.; Couch, C.; Burger, K. M.; Velazquez, A.; Okonkwo, J. U.; King, J.-R.; Park, S.; Agarwal, S.; et al. 2019. Detection of brain activation in unresponsive patients with acute brain injury. *New England Journal of Medicine*, 380(26): 2497–2505.

Connolly, J. F.; Reilly, J. P.; Fox-Robichaud, A.; Britz, P.; Blain-Moraes, S.; Sonnadara, R.; Hamielec, C.; Herrera-D'ıaz, A.; and Boshra, R. 2019. Development of a point of care system for automated coma prognosis: a prospective cohort study protocol. *BMJ open*, 9(7): e029621.

Di Gregorio, F.; La Porta, F.; Petrone, V.; Battaglia, S.; Orlandi, S.; Ippolito, G.; Romei, V.; Piperno, R.; and Lullini, G. 2022. Accuracy of EEG biomarkers in the detection of clinical outcome in disorders of consciousness after severe acquired brain injury: preliminary results of a pilot study using a machine learning approach. *Biomedicines*, 10(8): 1897.

Donnelly, E.; and McCulloch, K. 2014. Measurement characteristics and clinical utility of the Coma Recovery Scale-Revised among individuals with acquired brain injury. *Archives of Physical Medicine and Rehabilitation*, 95(7): 1417–1418.

Duncan, C. C.; Barry, R. J.; Connolly, J. F.; Fischer, C.; Michie, P. T.; N"a"at"anen, R.; Polich, J.; Reinvang, I.; and Van Petten, C. 2009. Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology*, 120(11): 1883–1908.

Escera, C.; and Corral, M. 2007. Role of mismatch negativity and novelty-P3 in involuntary auditory attention. *Journal of psychophysiology*, 21(3-4): 251–264.

Fischer, C.; Luaute, J.; Adeleine, P.; and Morlet, D. 2004. Predictive value of sensory and cognitive evoked potentials for awakening from coma. *Neurology*, 63(4): 669–673.

Fischer, C.; Morlet, D.; Bouchet, P.; Luaute, J.; Jourdan, C.; and Salord, F. 1999. Mismatch negativity and late auditory evoked potentials in comatose patients. *Clinical neurophysiology*, 110(9): 1601–1610.

Giacino, J. T.; Kalmar, K.; and Whyte, J. 2004. The JFK Coma Recovery Scale-Revised: measurement characteristics and diagnostic utility. *Archives of physical medicine and rehabilitation*, 85(12): 2020–2029.

Herrera-Diaz, A.; Boshra, R.; Tavakoli, P.; Lin, C.-Y. A.; Pajankar, N.; Bagheri, E.; Kolesar, R.; Fox-Robichaud, A.; Hamielec, C.; Reilly, J. P.; et al. 2023. Tracking auditory mismatch negativity responses during full conscious state and coma. *Frontiers in Neurology*, 14: 1111691.

Justo-Guill'en, E.; Ricardo-Garcell, J.; Rodr'ıguez-Camacho, M.; Rodr'ıguez-Agudelo, Y.; de Larrea-Mancera, E. S. L.; and Sol'ıs-Vivanco, R. 2019. Auditory mismatch detection, distraction, and attentional reorientation (MMN-P3a-RON) in neurological and psychiatric disorders: a review. *International Journal of Psychophysiology*, 146: 85–100.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Luaute, J.; Fischer, C.; Adeleine, P.; Morlet, D.; Tell, L.; and Boisson, D. 2005. Late auditory and event-related potentials can be useful to predict good functional outcome after coma. *Archives of physical medicine and rehabilitation*, 86(5): 917–923.

Luck, S. J. 2014. *An introduction to the event-related potential technique*. MIT press.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Makeig, S.; Jung, T.-P.; Ghahremani, D.; and Sejnowski, T. J. 1996. Independent component analysis of simulated ERP data. *Institute for Neural Computation, University of California: technical report INC-9606*.

Marek, S.; Tervo-Clemmens, B.; Calabro, F. J.; Montez, D. F.; Kay, B. P.; Hatoum, A. S.; Donohue, M. R.; Foran, W.; Miller, R. L.; Hendrickson, T. J.; et al. 2022. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902): 654–660.

Morlet, D.; and Fischer, C. 2014. MMN and novelty P3 in coma and other altered states of consciousness: a review. *Brain topography*, 27(4): 467–479.

Naccache, L.; Puybasset, L.; Gaillard, R.; Serve, E.; and Willer, J.-C. 2004. Auditory mismatch negativity is a good predictor of awakening in comatose patients: a fast and reliable procedure. *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology*, 116(4): 988–989.

Näätänen, R.; Gaillard, A. W.; and Mäntysalo, S. 1978. Early selective-attention effect on evoked potential reinterpreted. *Acta psychologica*, 42(4): 313–329.

Näätänen, R.; Kujala, T.; and Light, G. 2019. *Mismatch negativity: a window to the brain*. Oxford University Press.

Näätänen, R.; Paavilainen, P.; Rinne, T.; and Alho, K. 2007. The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology*, 118(12): 2544–2590.

Perrin, F.; Pernier, J.; Bertrand, O.; and Echallier, J. 1989. Spherical splines for scalp potential and current density mapping. *Electroencephalography and clinical neurophysiology*, 72(2): 184–187.

Polich, J. 2007. Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, 118(10): 2128–2148.

Prechelt, L. 2012. *Early Stopping — But When?*, 53–67. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-35289-8.

Rohaut, B.; Calligaris, C.; Hermann, B.; Perez, P.; Faugeras, F.; Raimondo, F.; King, J.-R.; Engemann, D.; Marois, C.; Le Guennec, L.; et al. 2024. Multimodal assessment improves neuroprognosis performance in clinically unresponsive critical-care patients with brain injury. *Nature Medicine*, 1–7.

Schnakers, C. 2020. Update on diagnosis in disorders of consciousness. *Expert Review of Neurotherapeutics*, 20(10): 997–1004.

Tavakoli, P.; Dale, A.; Boafo, A.; and Campbell, K. 2019. Evidence of P3a during sleep, a process associated with intrusions into consciousness in the waking state. *Frontiers in Neuroscience*, 1028.

Teasdale, G.; and Jennett, B. 1974. Assessment of coma and impaired consciousness: a practical scale. *The Lancet*, 304(7872): 81–84.

Tjepkema-Cloostermans, M. C.; da Silva Lourenço, C.; Ruijter, B. J.; Tromp, S. C.; Drost, G.; Kornips, F. H.; Beishuizen, A.; Bosch, F. H.; Hofmeijer, J.; and van Putten, M. J. 2019. Outcome prediction in postanoxic coma with deep learning. *Critical care medicine*, 47(10): 1424–1432.

Turgeon, A. F.; Lauzier, F.; Simard, J.-F.; Scales, D. C.; Burns, K. E.; Moore, L.; Zygun, D. A.; Bernard, F.; Meade, M. O.; Dung, T. C.; et al. 2011. Mortality associated with withdrawal of life-sustaining therapy for patients with severe traumatic brain injury: a Canadian multicentre cohort study. *Cmaj*, 183(14): 1581–1588.

Wannez, S.; Heine, L.; Thonnard, M.; Gosseries, O.; Laureys, S.; and Collaborators, C. S. G. 2017. The repetition of behavioral assessments in diagnosis of disorders of consciousness. *Annals of neurology*, 81(6): 883–889.

Young, G. B. 2009. Coma. *Annals of the New York Academy of Sciences*, 1157(1): 32–47.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.

Zhang, J.; Zhang, E.; Yuan, C.; Zhang, H.; Wang, X.; Yan, F.; Pei, Y.; Li, Y.; Wei, M.; Yang, Z.; et al. 2022. Abnormal default mode network could be a potential prognostic marker in patients with disorders of consciousness. *Clinical Neurology and Neurosurgery*, 107294.