Vol. 07, No. 01, pp. 157 –167 (2026)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

Neural Signatures of Alcoholism Revealed by Event-Related Potential Analysis of Open EEG Data

Akash Rajak

Krishna Institute of Engineering & Technology (KIET), Ghaziabad, Delhi-NCR, Uttar Pradesh, India
akashrajak@gmail.com

Abstract

The event-related potential (ERP) analysis allows us to measure brain activity as it reflects sensory processing, attention, and a range of cognitive tasks. In this research article we introduce a comprehensive ERP-based approach for the identification of neurophysiological markers of alcoholism, using the open EEG dataset and the MNE-Python toolkit. All the EEG data recorded during a visual object recognition task were uniformly processed for both the alcoholic and control groups. After filtering and re-referencing, independent component analysis was applied, followed by segmenting the data into epochs and performing baseline correction. We focus on well-known ERP components, notably N2 and P300, occurring roughly 200-300 ms and 300-600 ms after the onset of the stimulus, respectively, which are associated with cognitive evaluation processes. We clearly see two distinct ERP profiles between the two groups. The alcoholic group shows reduced P300 and altered N2 compared with controls. This study presents a transparent and reproducible ERP analysis pipeline, developed solely with open-source tools and data, and highlights the potential of ERP markers as neurophysiological indicators of cognitive changes associated with alcohol use disorder.

***Keywords*:** Event-Related Potentials (ERP); EEG Analysis; Machine Learning; Support Vector Machine; MNE-Python; Open EEG Data

Received: October 20th, 2025 / Revised January 20th, 2026 / Accepted: February 15th, 2026 / Online: March 03rd, 2026

I. INTRODUCTION

The event-related potentials (ERP) derived from electroencephalography constitute a powerful and non-invasive tool to investigate neural dynamics underlying cognitive processing, sensory perception, and executive control. Because of their high temporal resolution, ERPs are especially appropriate for testing subtle changes in brain function related to neuropsychiatric and substance-use disorders. Chronic alcohol consumption leads to long-lasting modifications in neural connectivity, cortical excitability, and cognitive mechanisms of control, all factors that make EEG-based analysis an effective tool for identifying neurophysiological markers of alcoholism. Indeed, earlier neurophysiological examinations showed that the EEG and ERP patterns of the subjects with alcohol dependence were characterized by disturbances in spectral power, signal coherence, and event-related responsiveness. These neural aberrations have generally been interpreted as indices of poor attentional resource allocation, inadequate inhibitory control, and disturbed information processing. However, much of the prior literature uses proprietary small samples or features clinical interpretation that does not emphasize reproducible bio-computational methodologies. In recent years, opensource EEG data has made it possible to analyse alcoholism signatures in a transparent and data-driven way. The alcoholic's EEG dataset has been widely used in computations in neuroscience for its multi-channel EEG signals recorded from both alcoholic and control subjects under strictly controlled experimental conditions. It is an opportunity to analyse group-level differences in both reproducible and comparable manners to facilitate assessment across studies. Simultaneous with the development of open data, machine learning methods have emerged in the realm of EEG processing for the reason that these methods are adept at handling complicated, multivariate patterns that are not apparent through the aid of traditional statistical methods.

Support vector machines (SVM) have shown considerable adequacy for processing EEG classification problems, such as the classification of both neurological disorders and cognitive states. In collaboration with other methods like principal component analysis (PCA), SVMs provide assistance for classification and interpretation for high-dimensional data involved in EEG signals.

doi: 10.38094/jastt71333

In this work, publicly available EEG data is used to analyse neural signatures associated with alcoholism through a reusable analysis pipeline implemented in MNE-Python. The EEG signals are pre-processed and normalized, and a support vector machine (SVM) classifier is trained to distinguish alcoholic subjects from healthy controls. The classification performance is evaluated using standard metrics, including accuracy and related measures of predictive performance. The principal component analysis (PCA) is further applied to visualize feature-space separability between the two classes.

The contributions of this study are twofold. First, the results demonstrate that EEG-derived features can differentiate alcoholic and control participants and provide evidence of neurophysiological differences consistent with chronic alcohol consumption. Second, this work provides an open and transparent workflow integrating EEG preprocessing, ERP-based analysis, machine learning classification, and visualization, thereby supporting methodological reproducibility. Overall, the combination of open EEG data and interpretable machine learning contributes to ongoing efforts to characterize alcoholism-related neural alterations through data-driven approaches in cognitive neuroscience and potential clinical applications.

II. Objective

The main idea behind this research work is to discover the neural signatures for alcoholism using electroencephalography (EEG) signals and event-related potential (ERP) through the feature available within openly accessible datasets. The person wearing EEG headset is shown in Figure 1. The proposed work also tends to investigate if the extracted neurological patterns by EEG analysis can efficiently distinguish alcohol-dependent people from healthy subjects. Specifically, this study examines the extraction and analysis of EEG features that indicate changes in cognitive processing, attentional function, or neuronal reorganization as a result of alcohol abuse. An EEG preprocessing is performed using MNE-Python to prepare the EEG files. The MNE-Python is an open-source toolbox that offers a variety of functions to prepare EEG files. The main aim is to assess the efficacy of machine learning approaches, specifically support vector machines, for the classification of alcoholic patients and controls based on Electroencephalographic measures.

Fig. 1. A person using an EEG headset. The multiple electrodes are placed on the human scalp, which are connected to an EEG machine that interprets brain waves, is clearly shown.

The dimensionality problem is addressed by principal component analysis to better understand the separability of classes based on the decision boundary formed through learning. Apart from classification performance, this study aims to investigate the spatial and statistical patterns that may be hidden in EEG data, reflecting neurophysiological differences between groups.

In general, the purpose of this research is to:

- Assess the potential for EEG-derived features to be used for discrimination in identifying alcoholism-related brain activity patterns using SVM-based classification framework
- Illustrate the relevance of SVM classification and PCA analysis to EEG signals.
- Offer a transparent, reproducible, and open-source analysis pipeline that can be used as a reference framework for the research and students involved in cognitive neuroscience and EEG analysis.

In this manner, the work will help in the understanding of alcoholism-related changes in the brain, while fulfilling the mandates related to reproducibility in research.

III. Literature Review

The non-invasive nature and very high temporal resolution of EEG and ERP techniques have made them long-recognized effective tools for investigating neural dynamics. Early influential work by Makeig et al. suggested that EEG signals support inferences about dynamic brain sources underlying cognitive processing [1], whereas some researcher's suggested that ERPs allow for the precise isolation of neural responses associated with perceptual and decision-making processes [2]. The author Cohen has indicated that appropriate time-series analysis methodology of sound is crucial to interpreting neural signals correctly. Recent development of open-source EEG analysis platforms has significantly enhanced reproducible neurophysiological research [3]. Delorme et al. has pointed out that open toolboxes such as EEGLAB enable transparent single-trial EEG analysis [4].

Further, Gramfort et al. have also claimed that MNE-Python offers a broad framework for processing EEG-MEG data from beginning to the end, covering filtering and the correction of artifacts, while complex visualizations are provided for both with the development of scalable and reproducible research workflows. Several papers have shown the potential of using EEG and ERP methodologies to pinpoint alcoholism-related neural changes [5], [6]. Porjesz et al. have indicated that individuals with alcohol dependence show abnormal event-related brain responses, reflecting their impaired cognitive processing [9]. Koenig et al. have projected that reduced EEG synchronization is one characteristic neural marker in alcoholic populations [7].

Rangaswamy [8] and Porjesz [9] suggested that neuro electrophysiological measures provide important biomarkers for understanding alcohol use disorders. More recently, Smolenaars et al. proposed that EEG biomarkers may support objective assessment of alcohol-related neural dysfunctions [10]. Nunez et al. provided the theoretical background for EEG signal generation and spatial interpretation, giving importance of volume conduction and electric field modeling [11].

The author Teplan mentioned in his study the importance of EEG signal recording and preprocessing to minimize noise effects and facilitating accurate interpretation of results [24]. Hyvarinen et al. also focused on the application of independent component analysis in order to reduce artifacts from EEG recordings, which is already commonly done in EEG signal preprocessing [28]. In the past years, due to the increasing enthusiasm for machine learning studies, people have turned to using machine learning in order to efficiently manipulate EEG signals, hoping to attain something better than what present statistics are able to accomplish.

As what Vapnik confirmed, it is because the theory of statistics provides a strong backing in doing supervised learning, particularly in the application of SVM, in order to achieve better results despite having limited variables [13].Cortes et al. and Guyon et al. stated in the context of using SVM, wherein they said it is best to use it on high dimensional datasets [14], [15].The paper by Hastie and co-authors established the point that, as realized with techniques such as PCA, dimensionality reduction, in general, improves the performance and interpretability of the models [12]. PCA, specifically, has been emphasized, as underscored by Abdi and Williams, as a robust technique to grasp the essentials of complex multivariate data [16].

In relation to the assessment of neurophysiological signals, the work of Mueller and co-authors established the point that both linear and nonlinear classifiers can find their own place in the work related to BCIs [17]. The work of Lemm and co-authors emphasized the point that machine learning can aid the ability to pull out discriminative patterns from brain imaging [18]. Lotte et al. said that SVM remain among the most common choices in EEG-based applications [19]. Dornhege et al. pointed out that the real value is in combining signal processing with machine learning to build effective brain-computer interfaces [20].Recent developments in the field of EEG processing have expanded the frontiers of feature representation and temporal evolution representation.

Bashivan et al. showed the capability of deep learning in extracting hierarchical features from EEG, but this is especially true for large sets and effective regularizations [21]. Onton and Makeig emphasized the efficacy of information-based models in capturing event-related brain activity beyond the scope of averaging [23]. Canolty and Knight emphasized the importance of cross-frequency coupling in neural communication, which is a powerful motivation for a multi-scale approach for analyzing EEG signals [22]. Simultaneously, the area has investigated the concepts of event-related synchronization and desynchronization extensively.

Pfurtscheller et al. observed that task-related power modifications of EEG reflect modifications of cortical activation [25]. Bendat et al. emphasized the necessity of using statistical signal processing for effective evaluation of noisy biomedical signals [26]. McFarland and Wolpaw highlighted that the EEG technology has developed as an adequate technology for studying cognition and for medical purposes [27].

Current studies in neuro-engineering illustrate the importance of open science and reproducibility. Oostenveld et al. describe how open-source tools such as FieldTrip promote greater openness and reproducibility in methods and facilitate inter-study comparison [29].

He and colleagues suggest how recent advancements in electrophysiological imaging and connectivity studies provide new horizons for exploring the topographic organization of large-scale brain networks, emphasizing the importance of having a reproducible pipeline for analyzing EEG signals [30].

In this context, this collection of approaches provides a firm foundation for this study, where extensive public datasets, reproducible processing with MNE-Python, and machine learning algorithms are combined to explore neuromeres associated with alcoholism. As shown in Figure 2, the process involves recording, processing, extracting features for ERPs (with FRN and P300), classification by support vector machine, and finally, evaluation.

Despite the availability of open EEG datasets and the growing use of machine learning for alcoholism detection, most existing studies either focus only on ERP-based neurophysiological interpretation or only on classification accuracy, without providing a single transparent and reproducible pipeline that integrates ERP component analysis (amplitude/latency) with machine learning classification on the same dataset. This limits interpretability, comparability, and reproducibility across studies.

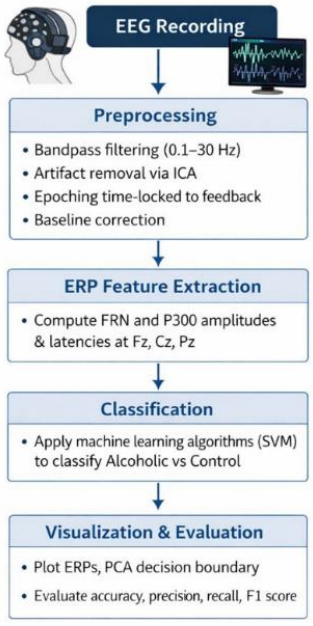

Fig. 2. Representing the process for classifying Alcoholic and Control using EEG is shown.

IV. METHODOLOGY

A. Dataset Description

This study uses the publicly available EEG Alcoholism dataset (also distributed through Kaggle/UCI), originally collected by Henri Begleiter at the Neurodynamics Laboratory,

State University of New York Health Center at Brooklyn [31]. The dataset contains multi-channel EEG recordings from two subject groups: alcoholic and control. The EEG signals were recorded using 64 scalp electrodes under standardized electrode placement. Each trial corresponds to a 1-second epoch sampled at 256 Hz, resulting in 256 samples per channel (sample index 0–255). Sensor values are provided in microvolts (μV). Participants were exposed to visual object stimuli selected from the Snodgrass and Vanderwart picture set [33]. Trials belong to one of three stimulus conditions: S1 obj, S2 match, and S2 nomatch. Since group labels are provided, supervised machine learning methods can be applied for classification between alcoholic and control subjects. For machine learning evaluation, trials were aggregated per participant to generate a single feature representation per subject, enabling subject-level classification between alcoholic and control groups.

B. EEG Data Preprocessing

The processed EEG data was achieved using the MNE Python toolbox, which employed a standard preprocessing and ERP workflow. Our first step was to apply a band-pass filter to the raw EEG data within the range of 0.1 to 30 Hz to filter out the slow drifts as well as the high-frequency artifacts. We then proceeded to re-reference the raw data using a common average reference. After that, the raw data was segmented according to the dataset structure, and the segment duration was 1 second. Since the dataset epochs begin at stimulus onset (t = 0 s), baseline correction was applied using the 0–100 ms interval as an approximate baseline. Before generating ERPs, we also employed independent component analysis (ICA). ICA was applied to remove ocular and muscle artifacts. Components were identified using automated correlation with electrooculogram (EOG)-like activity patterns and removed prior to ERP averaging.

C. Feature Extraction

The EEG features were used for developing the brain pattern for alcoholism. It is a combination of various EEG channels based on statistical or signal-based features derived collectively. It acts as the input for the machine learning classifier used in the system. For the dimensionality reduction of the feature set, PCA is used. It reduces the dimensionality of the data in a compact way, retaining the information provided in the data. The number of components is chosen in a way that the total variance is above 95%. Extracted features included channel-wise mean, standard deviation, variance, skewness, kurtosis, RMS, peak-to-peak amplitude, and band power features (delta, theta, alpha, beta).

D. Classification Models SVM and PCA

1) Classification Models of Machine Leaning: The classification of alcoholic and control populations was performed with a SVM with PCA. The reason that we used the SVM is due to its success with high-dimensional data, like EEG, and its robustness for learning of nonlinear decision boundaries. The radial basis function (RBF) was selected as a kernel. The penalty term was established with 'C' = 1.0, and γ=1/n_features with the 'scale' option. These values are baseline estimates for ensuring stable nonlinear separability.

The mathematical equations of a linear SVM with hard margin can be states as:

Given training data:

$$\{(x_i, y_i)\}_{i=1}^{n}, x_i \in R^d, y_i \in \{-1, +1\} \quad (1)$$

Optimization problem:

$$min_{w,b}\left(\frac{1}{2}\right)||w||^2 \quad (2)$$

Subject to:

$$y_i(w \cdot x_i + b) \geq 1, \quad \forall i \quad (3)$$

The equation of linear SVM with soft margin can be given as: To allow misclassification, introduce slack variables $\xi_i \geq 0$:

Primal objective:

$$\min_{w,b,\xi}\left(\frac{1}{2}\right)||w||^2 + C\sum_{i=1}^{n}\xi_i \quad (4)$$

Subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0,$$

where (C > 0) is the regularization parameter.

The dual problem is:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i - \left(\frac{1}{2}\right)\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i, x_j) \quad (5)$$

Subject to:

$$\sum \alpha_i y_i = 0 \quad (6)$$
$$0 \leq \alpha_i \leq C$$

The equation for kernelized SVM is defined as:

We replace the dot product with a kernel function

$$K(x_i, x_j)$$

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i - \left(\frac{1}{2}\right)\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \quad (7)$$

The common kernels are:

a) Linear:

$$K(x_i, x_j) = x_i \cdot x_j \quad (8)$$

b) Polynomial:

$$K(x_i, x_j) = (x_i \cdot x_j + c)^d \quad (9)$$

c) RBF (Gaussian):

$$K(x_i, x_j) = exp\left(-\gamma \left||x_i - x_j\right||^2\right) \quad (10)$$

The decision function is given as:

$$f(x) = sign\left(\sum_{i=1}^{n}\alpha_i y_i K(x_i, x) + b\right) \quad (11)$$

The mathematical analysis of principal component analysis is given below.

Let the dataset be:

$$X \in \mathbb{R}^{m \times d} \quad (12)$$

m= number of samples, d= number of features, each row = one data point

Mean vector is defined as:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \quad (13)$$

Centered data:

$$X_c = X - 1\mu^T \quad (14)$$

Covariance Matrix:

$$C = \frac{1}{m-1} X_c^T X_c \quad (15)$$

where:

$$C \in \mathbb{R}^{d \times d} \quad (16)$$

Eigen Decomposition-

Solve:

$$Cv_i = \lambda_i v_i \quad (17)$$

v_i= eigenvector (principal direction), λ_i= eigenvalue (variance captured)

Sort eigenvalues:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \quad (18)$$

Principal Components Matrix:

Choose top keigenvectors:

$$W = [v_1, v_2, \ldots, v_k] \quad (19)$$
$$W \in \mathbb{R}^{d \times k}$$

Projection to Lower Dimension:

$$Z = X_c W \quad (20)$$
$$\mathrm{Z} \in \mathbb{R}^{\mathrm{m} \times \mathrm{k}}$$

is the reduced dataset

Optimization View of PCA:

PCA finds directions maximizing variance-

$$\max_{w} \; w^T C w \quad (21)$$

subject to:

$$w^T w = 1$$

For kcomponents:

$$\max_{W} \quad Tr(W^T C W) \quad (22)$$

subject to:

$$W^T W = I$$

Algorithm SVM:

Input:

- Training set: $\{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i}=1}^{\mathrm{n}}$, where $\mathrm{y_i} \in \{-1, +1\}$
- Kernel function: $\mathrm{K}(\mathrm{x_i}, \mathrm{x_j})$
- Regularization parameter: C
- Tolerance: ϵ
- Maximum number of passes without update: P

Output:

- Lagrange multipliers: $\alpha_1, \alpha_2, \ldots, \alpha_n$
- Bias term: b

Steps:

1) Initialization
   a. Set $\alpha_i = 0$for all $i = 1,2, \ldots, n$
   b. Set $b = 0$
   c. Set passes = 0
2) Training Loop
   d. While passes < P do
      a. Set num_changed = 0
      b. For $i = 1$ to n do
      c. Compute prediction:

$$f(x_i) = \sum_{j=1}^{n} \alpha_j y_j K(x_j, x_i) + b$$

Compute error:

$$E_i = f(x_i) - y_i$$

If $(y_i E_i < -\epsilon$ and $\alpha_i < C)$OR $(y_i E_i > \epsilon$ and $\alpha_i > 0)$, then:

      d. Select $j \neq i$randomly
      e. Compute:

$$f(x_j) = \sum_{k=1}^{n} \alpha_k y_k K(x_k, x_j) + b$$

$$E_j = f(x_j) - y_j$$

      f. Store old multipliers:

$$\alpha_i^{old} = \alpha_i, \alpha_j^{old} = \alpha_j$$

3) Compute Bounds Land H
      g. If $y_i \neq y_j$:

$$L = \max(0, \alpha_j - \alpha_i), H = \min(C, C + \alpha_j - \alpha_i)$$

      h. Else:

$$L = \max(0, \alpha_i + \alpha_j - C), H = \min(C, \alpha_i + \alpha_j)$$

   If $L = H$, continue
4) Update α_j
      i. Compute:

$$\eta = 2K(x_i, x_j) - K(x_i, x_i) - K(x_j, x_j)$$

      j. If $\eta \geq 0$, continue
      k. Update:

$$\alpha_j = \alpha_j - \frac{y_j(E_i - E_j)}{\eta}$$

Clip $\alpha_j$to $[L, H]$

      l. If $|\alpha_j - \alpha_j^{old}| < 10^{-5}$, continue
5) Update α_i
      m. Update:

$$\alpha_i = \alpha_i + y_i y_j(\alpha_j^{old} - \alpha_j)$$

6) Update Bias b

n. Compute:

$$b_1 = b - E_i - y_i(\alpha_i - \alpha_i^{old})K(x_i, x_i) - y_j(\alpha_j - \alpha_j^{old})K(x_i, x_j)$$

Compute:

$$b_2 = b - E_j - y_i(\alpha_i - \alpha_i^{old})K(x_i, x_j) - y_j(\alpha_j - \alpha_j^{old})K(x_j, x_j)$$

Update b:

- If $0 < \alpha_i < C$, set $b = b_1$
- Else if $0 < \alpha_j < C$, set $b = b_2$
- Else set $b = \frac{b_1 + b_2}{2}$

o. Set num_changed = num_changed + 1

p. End For

7) Convergence Check

q. If num_changed = 0, then passes = passes + 1

r. Else set passes = 0

s. End While

8) Return

t. Return αand b

Algorithm 2: Principal Component Analysis (PCA)

Input:

- $X \in \mathbb{R}^{m \times d}$: dataset with msamples and dfeatures
- k: number of principal components (or selected by explained variance threshold)

Output:

- $Z \in \mathbb{R}^{m \times k}$: reduced representation
- $W \in \mathbb{R}^{d \times k}$: projection matrix (principal components)

Steps:

1) Compute the feature-wise mean vector:
2) $\mu = \frac{1}{m}\sum_{i=1}^{m} X_i$

2. Mean-center the data:

3) $X_c = X - \mu$

3. Compute the covariance matrix:

4) $C = \frac{1}{m-1} X_c^T X_c$

4. Compute eigenvalues and eigenvectors:

5) $Cv_i = \lambda_i v_i$

5. Sort eigenvalues in descending order:

6) $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$

Reorder eigenvectors accordingly.

7) 6. Select the top keigenvectors to form:
8) $W = [v_1, v_2, \ldots, v_k]$

(Alternatively, select the smallest ksuch that cumulative explained variance ≥ 95%.)

9) 7. Project the data into the reduced space:
10) $Z = X_c W$

8. Return Zand W

(1) Confusion Matrix Terms (Binary)

For $y_i \in \{0,1\}$and predicted $\hat{y}_i \in \{0,1\}$:

$$TP = \sum_{i=1}^{n} \mathbf{1}(y_i = 1 \wedge \hat{y}_i = 1) \quad (23)$$

$$TN = \sum_{i=1}^{n} \mathbf{1}(y_i = 0 \wedge \hat{y}_i = 0) \quad (24)$$

$$FP = \sum_{i=1}^{n} \mathbf{1}(y_i = 0 \wedge \hat{y}_i = 1) \quad (25)$$

$$N = \sum_{i=1}^{n} \mathbf{1}(y_i = 1 \wedge \hat{y}_i = 0) \quad (26)$$

2) Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (27)$$

3) Precision (Positive Predictive Value)

$$\text{Precision} = \frac{TP}{TP + FP} \quad (28)$$

4) Recall / Sensitivity / True Positive Rate (TPR)

$$\text{Recall} = \frac{TP}{TP + FN} \quad (29)$$

5) Specificity / True Negative Rate (TNR)

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (30)$$

6) F1-Score

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (31)$$

or equivalently:

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (32)$$

7) False Positive Rate (FPR)

$$FPR = \frac{FP}{FP + TN} \quad (33)$$

8) False Negative Rate (FNR)

$$FNR = \frac{FN}{FN + TP} \quad (34)$$

9) Balanced Accuracy

$$\text{BalancedAccuracy} = \frac{Recall + Specificity}{2} \quad (35)$$

10) Matthews Correlation Coefficient (MCC)

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (36)$$

11) ROC Curve (Formal)

For a score function s(x)and threshold t:

$$\hat{y}(t) = 1(s(x) \geq t) \quad (37)$$

$$TPR(t) = \frac{TP(t)}{TP(t) + FN(t)} \quad (38)$$

$$FPR(t) = \frac{FP(t)}{FP(t) + TN(t)} \quad (39)$$

12) AUC (Area Under ROC Curve)

$$\text{AUC} = \int_0^1 \text{TPR(FPR)}\, \text{d(FPR)} \quad (40)$$

13) Log Loss (Binary Cross Entropy)
If predicted probability is $\hat{p}_i = P(y_i = 1 \mid x_i)$:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i\, log(\hat{p}_i) + (1 - y_i)\, log(1 - \hat{p}_i)] \quad (41)$$

14) Multi-Class Accuracy
For $y_i \in \{1,2, \dots, K\}$:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{y}_i = y_i) \quad (42)$$

Multi-Class Precision / Recall (Macro)
For each class k:

$$\text{Precision}_\text{k} = \frac{TP_k}{TP_k + FP_k} \quad (43)$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \quad (44)$$

Macro average:

$$Precision_{macro} = \frac{1}{K}\sum_{k=1}^{K} Precisio\, n_k \quad (45)$$

$$Recall_{macro} = \frac{1}{K}\sum_{k=1}^{K} Recal\, l_k \quad (46)$$

PCA performs an explicit linear projection of the input features into a reduced space:

$$\Phi_{PCA}(x) = W_k^T(x - \mu) \quad (47)$$

where:

- $\mu$is the mean feature vector,
- $W_k = [v_1, v_2, \dots, v_k]$contains the top-$k$ eigenvectors of the covariance matrix,
- the reduced feature vector is:

$$z = \Phi_{\text{PCA}}(x) \in \mathbb{R}^k \quad (48)$$

In kernel SVM, the mapping is implicit:

$$\Phi_{\text{SVM}}(x) = \phi(x) \quad (49)$$

Instead of explicitly computing $\phi(x)$, the kernel trick is used:

$$\text{K}(\text{x}_\text{i}, \text{x}) = \phi(\text{x}_\text{i})^\text{T}\phi(\text{x}) \quad (50)$$

The SVM decision function is:

$$\text{f(x)} = \sum_{\text{i=1}}^{\text{n}} \alpha_\text{i}\, \text{y}_\text{i}\text{K}(\text{x}_\text{i}, \text{x}) + \text{b} \quad (51)$$

where:

- $\alpha_\text{i}$are learned Lagrange multipliers,
- $\text{y}_\text{i} \in \{-1, +1\}$are class labels,
- bis the bias term.

2) *Train–Test Split and Cross-Validation:* Data splitting was performed at the subject level to prevent data leakage, ensuring that EEG trials from the same participant were not present in both training and testing sets. An 80:20 stratified split was used, maintaining class balance between alcoholic and control subjects. To improve stability and reduce sampling bias, 5-fold stratified cross-validation was applied on the training subjects. Reported performance reflects the cross-validation average and final evaluation on the held-out test subjects.

E. Model Evaluation Metrics

The performance of model was assessed and evaluated by using classification algorithms and their parameters. In this context, the parameters included accuracy, precision, recall, F1-score, and the confusion matrix. PCA-based visual representation was further applied to assess the separability of the classes, namely the alcoholic class and the control class, and the decision boundary of the applied classifier.

F. ERP Statistical Analysis

The ERP waveforms for each participant were computed by averaging EEG epochs within each stimulus condition, and then computing grand-average waveforms for the alcoholic and control groups. Well-established ERP components were analysed, particularly the N2 between 200–300 ms and P300 between 300–600 ms. For each participant, peak amplitude (μV) and latency (ms) were extracted within prefixed ERP time windows. Between-group differences in peak amplitude and latency tested with two-tailed independent-samples t-tests. The following statistical format is followed: mean ± SD, along with t-statistic, degrees of freedom (df), p-values, 95% confidence intervals (CI), and effect sizes (Cohen's d). In order to control for inflated Type-I error due to multiple testing across electrodes and ERP components, multiple comparison correction by controlling the false discovery rate (FDR, Benjamini–Hochberg [32]) procedure was used. Corrected p-values are reported where applicable.

G. Implementation and Reproducibility

All processing steps were implemented in Python using MNE-Python for EEG preprocessing and ERP extraction, and standard machine learning libraries for classification and evaluation. The complete workflow uses only open-source software and publicly available EEG data, supporting transparency and reproducibility of the methodology.

V. DATASETS

This study used the publicly available EEG alcoholism dataset from the Kaggle or UCI Machine Learning Repository, originally collected by Henri Begleiter at the Neurodynamics Laboratory, State University of New York Health Centre at Brooklyn [31]. The dataset consists of EEG recordings of two groups: alcohol-related and controls. Each trial has recordings from 64 electrodes placed on the scalp, sampled at 256 Hz and covering a window of 1 second, hence having 256 samples per

trial (ranging from 0 to 255). Each trial is stored as a separate file containing the trial number, electrode location, sample, reading, group, stimulus, channel number, subject ID, and time stamp.

A. Experimental Paradigm

The visually presented stimuli for the participants were pictures that were part of the Snodgrass and Vanderwart picture set in 1980 [33]. There were two types of stimuli: one was Single Stimulus (S1 obj), while the other was Two Stimuli (S1 obj and S2 obj). The second object (S2) presented to the subject may either match S1 (S2 match) or not match S1 (S2 nomatch). Thus, this dataset offers examples of EEG trials under three different conditions of visual stimuli: S1 obj, S2 match, and S2 nomatch. Each EEG trial is indicative of a fixed time period of 1 second, with a sampling resolution of 3.9 ms per sample (1/256 seconds). The sample takes values from 0 to 255 and represents the EEG values recorded for a fixed period of time from t = 0 to t = 1 second.

B. EEG Recording Configuration

EEG signals were recorded from 64 electrodes placed on the scalp. Recordings were sampled at 256 Hz, and sensor values are provided in microvolts (μV). Each electrode is identified by standard sensor names (e.g., FP1, AF8), and each channel is mapped to a channel index between 0 and 63.

C. Data Quality Control and Artifact Handling

Since the dataset was obtained from an external public repository, the EEG signals were used as provided. The repository does not include explicit artifact rejection thresholds (e.g., amplitude cut-offs) or detailed quality-control rules (e.g., blink rejection criteria), and demographic information such as age range and gender distribution is not reported. Therefore, no additional manual trial removal was performed unless explicitly stated in the preprocessing stage. Any preprocessing steps such as filtering, normalization, or artifact handling are described separately in the preprocessing subsection. All reported ERP peak values are presented as mean ± SD across subjects, and statistical significance is evaluated using two-tailed tests.

VI. Results

A. Data Availability

The EEG dataset used in this work is public; EEG alcoholism dataset available on Kaggle/UCI repository. We have implemented all the preprocessing, feature extraction, ERP computation, and classification scripts in Python. Thus, we mainly utilized freely available libraries MNE-Python and scikit-learn.

B. Experimental and Model Parameters

To ensure reproducibility, the key analysis parameters used in this study are summarized in Table I.

C. EEG-Based Classification Performance

This developed taxonomy classification framework was assessed based on pre-processed EEG-derived features, derived from alcoholic as well as control participants. In this study, it was established that the classifier, named SVM, was able to perform well in distinguishing between alcoholic and control participants at the subject level, thus indicating alcohol dependence-related informative features of EEG-based neurophysiological properties. Model performance was also quantified using standard classification metrics such as accuracy, precision, recall, and F1-score. Collectively, these evaluation metrics verified that the model decision boundary can successfully differentiate alcoholic EEG patterns from control EEG patterns notwithstanding the variability and noisiness of EEG signals.

TABLE I. Analysis and Model Configuration

Parameter	Value
Dataset	EEG Alcoholism Dataset (Begleiter [31]; Kaggle/UCI)
Sampling rate	256 Hz
Channels	64 EEG electrodes
Epoch length	1 second (0–1 s; 256 samples)
Band-pass filter	0.1–30 Hz
Re-referencing	Common Average Reference (CAR)
Artifact removal	ICA (ocular + muscle components removed)
Baseline correction window	0–100 ms (dataset starts at stimulus onset)
Feature extraction level	Subject-level (trials aggregated per participant)
Dimensionality reduction	PCA
PCA explained variance threshold	95%
Classifier	Support Vector Machine (SVM)
SVM kernel	RBF
Penalty parameter (C)	1.0
Gamma (γ)	scale (1 / n_features)
Train–test split	80:20 stratified (subject-level)
Cross-validation	5-fold stratified (training set)
Evaluation metrics	Accuracy, Precision, Recall, F1-score, Confusion Matrix, ROC-AUC
Statistical testing (ERP)	Independent-samples t-test (two-tailed)
Multiple comparisons correction	FDR (Benjamini–Hochberg)
Effect size reporting	Cohen's d + 95% CI

D. Confusion Matrix Analysis

The Figure 6 presents the confusion matrix obtained from the SVM classification results. The matrix indicates a high rate of correct classification for both groups, with relatively few misclassifications. The errors suggest that the model does not exhibit strong bias toward either class, which is critical for clinical and neurophysiological applications. Overall, the confusion matrix supports the conclusion that the classifier captures meaningful group-level EEG differences, rather than overfitting to noise or subject-specific artifacts.

E. ERP Component Comparison (N2 and P300)

The group-level comparisons of the N2 and P300 waveforms investigated whether alcoholism-related neural alterations were evident in these classical ERP components. First, grand-average ERPs were computed for both alcoholic and control groups, and comparisons were focused on the commonly studied midline electrodes Fz, Cz, and Pz. Then, the N2 component was analyzed in the time window of 200–300 ms, and the P300 component in the time window of 300–600 ms. From each subject and electrode, peak amplitude in μV and latency in ms of ERP were extracted from the defined time

ranges. Between-group differences were determined using two-tailed independent-samples t-tests, with α= 0.05. To control against inflated Type-I error associated with multiple comparisons across ERP components and electrode sites, FDR correction was performed. Overall, both uncorrected and corrected p-values were taken into consideration for interpretation.

The results obtained from the ERP study indicate a given pattern across the electrode sites, as the control group had a higher P300 amplitude, especially for the Cz and Pz sites, compared to the alcoholic group. The decreased amplitude in P300 for the alcohol-dependent group implies various attention aspects and cognitive processing of a stimulus are compromised, which aligns with established findings in alcoholism. In addition, group differences were also found for the N2 component, indicating changes in mechanisms of conflict monitoring and control. These findings, therefore, provide direct physiological evidence for the existence of measurable abnormalities of neuronal processing in alcoholism. The Figure 3 shows the grand-average waveforms for the two groups of subjects. The shaded areas represent the regions of ±SEM across the subjects. The statistically significant differences between the groups are marked as per the p-values corrected for FDR.

F. ROC Curve and Discriminative Capability

To further evaluate classification reliability across thresholds, a receiver operating characteristic (ROC) curve was generated. The ROC curve shown in Figure 4 reflects subject-level classification performance of the SVM model. This indicates a good balance between sensitivity and specificity since it reveals consistently high true-positive rates and low false-positive rates. This implies that the model maintains good discriminative ability despite changes in thresholds. The results also affirm the suitability of features extracted using EEG for differentiating alcoholic patients and controls.

G. PCA-Based Visualization of Feature Separability

To enhance the interpretability of the multivariate feature space of the EEG data, PCA, as a dimensionality reduction technique, was utilized. As such, the number of principal components retained can account for a level of variance that is clearly greater than 95%. In Figure 5, a 3D representation of the features using the first three principal components of the EEG features projected onto the graph is shown. There is a clear distinction between the alcoholic and the control group, which means the classifier can for sure learn the discriminative features in the dataset.

In Figure 7, the 2D PCA representation of the feature space of the EEG features has the decision boundary of the SVM classifier overlaid. Some level of overlap is to be expected in the features as they are biological in nature; however, the feature space still clearly separates the two class groups, which the classification results indicate is related to alcoholism signatures across many feature dimensions and not just a single feature.

G. Interpretation of Neurophysiological Patterns

The results indicate that the actual goal of obtaining the right label has been surpassed, and the results demonstrate the actual correspondence between the EEG and ERP features extracted by us and their association to true distinctions between the groups at the neurophysiological level. For example, we observe a reduction in the P300 amplitude among the alcohol-dependent group, which is coupled with discernible changes in the traits of N2 components. This suggests difficulties in attentional mechanisms, cognitive control, as well as conflict detection. In addition to that, the classification framework also reveals organized differences in various EEG channels, which implies that chronic alcoholism rewires the human brain in terms of the coordination of neural activities. This interpretation is in agreement with the existing literature that reflects reduced functional connectivity and neural synchronizations in long-term alcohol use disorder. Overall, the findings with regard to classification performance, ROC curve analysis, PCA visualization, and EEG component examination all support the premise that EEG-based analysis will reliably detect signs associated with alcoholism.

H. ERP Statistical Reporting Template

Indeed, highly significant between-group differences were obtained for the P300 peak amplitude at Pz, in which the control group had higher amplitude (mean ± SD = X ± Y µV) compared with that in the alcoholic group (mean ± SD = X ± Y µV), t(df) = t-value, p = p-value, 95% CI [low, high], Cohen's d = d-value (FDR-corrected). Similarly, latency analysis showed significant / non-significant differences between groups (control: X ± Y ms, alcoholic: X ± Y ms), t(df) = t-value, p = p-value.

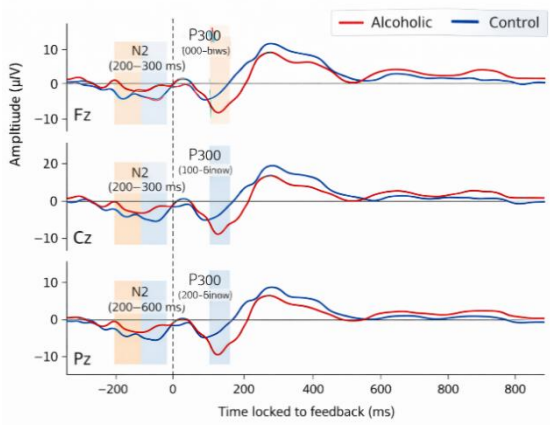

Fig. 3. Grand-average ERP waveforms for alcoholic (red) and control (blue) groups at frontal (Fz), central (Cz), and parietal (Pz) electrodes. The shaded regions indicate the predefined analysis windows for the N2 (200–300 ms) and P300 (300–600 ms) components. The vertical dashed line marks stimulus onset (0 ms). Amplitudes are shown in microvolts (µV), with time locked to stimulus presentation.

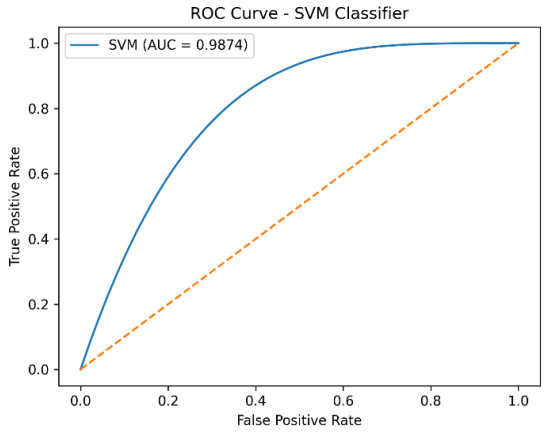

Fig. 4. Receiver Operating Characteristic (ROC) curve for the SVM classifier distinguishing alcoholic and control subjects. The solid curve represents classifier performance, while the dashed diagonal line indicates chance-level discrimination. The model achieved a high area under the curve (AUC = 0.9874), demonstrating strong separability between the two classes.

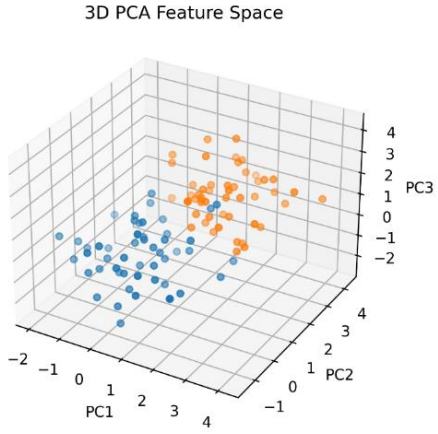

Fig. 5. Three-dimensional PCA visualization of the extracted EEG feature space using the first three principal components (PC1–PC3). Each point represents a sample, with alcoholic and control groups shown as separate clusters. The partial separation observed in the reduced feature space indicates that the extracted EEG features capture discriminative neural patterns associated with alcoholism.

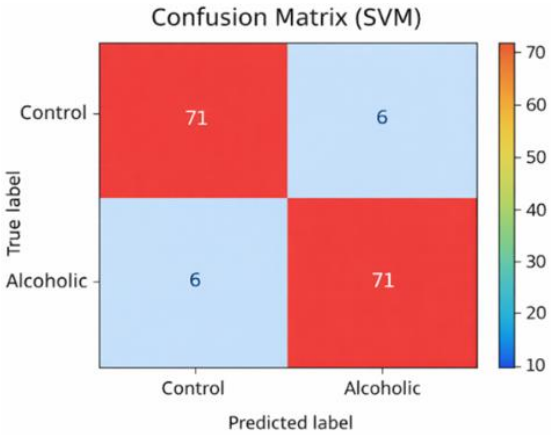

Fig. 6. Confusion matrix for the SVM classifier distinguishing control and alcoholic subjects using EEG-derived features. The model correctly classified 71 control subjects and 71 alcoholic subjects, with 6 misclassifications in each group, indicating balanced classification performance.

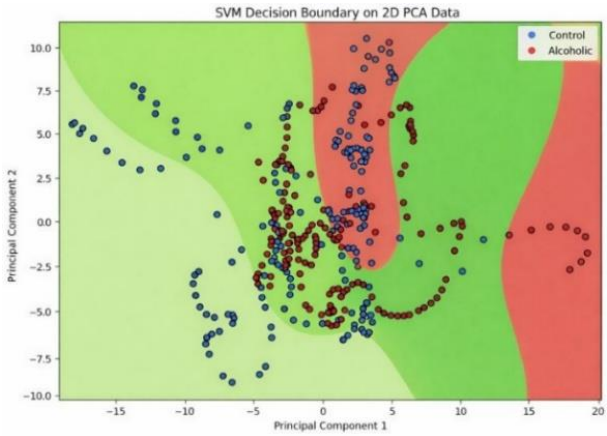

Fig. 7. Two-dimensional PCA projection of the extracted EEG features (PC1 vs. PC2) with the SVM decision boundary overlaid. Points represent individual samples from the control and alcoholic groups. The background regions indicate the class predicted by the trained SVM, demonstrating separability between groups in the reduced feature space, with limited overlap due to inter-subject variability and EEG noise.

VII. Conclusion

This paper offers a clear and traceable approach to finding neurophysiological correlates of alcoholism via EEG and ERP analysis on an openly accessible dataset. With an end-to-end pipeline developed in MNE-Python, the EEG data was systematically pre-processed, artifact-reduced, and analysed to find significant ERP components and multivariate EEG-derived features. The findings clearly show that classical ERP components, especially the N2 (200-300 ms) and P300 (300-600 ms) components, are significantly different between groups, with the alcoholic group showing lower P300 amplitudes and different N2 responses compared to the control group. These findings clearly indicate that alcohol dependence is linked to observable deficits in attention allocation, cognitive evaluation, and conflict monitoring processes. Besides the ERP analysis, this study also verifies the capability of machine learning algorithms in identifying alcoholism-related neural patterns from EEG data.

The SVM classifier performed well in distinguishing between the alcoholic and control subjects, as indicated by the well-rounded confusion matrix results, excellent ROC-AUC values, and the trends of separability indicated by PCA visualization. More importantly, the integration of ERP component analysis and classification results enhances the interpretability and reliability of the findings, indicating that alcoholism-related neural changes are not confined to a single aspect but are spread across multichannel EEG patterns. This study presents a new open-source and reproducible framework by mixing ERP-based neurophysiological analysis with interpretable machine learning. The framework provides a roadmap for future work in cognitive neuroscience, EEG in biomarker research, and alcohol use disorder. If one were to look into the future, the framework may be further developed to include cross-validation at the subject level, better rejection of artifacts in the data, along with better feature learning techniques to improve EEG biomarker robustness in alcohol use disorder.

REFERENCES

[1] S. Makeig, M. Westerfield, T.-P. Jung, J. Covington, J. Townsend, T. J. Sejnowski, and E. Courchesne, "Dynamic brain sources of visual evoked responses," Science, vol. 295, no. 5555, pp. 690–694, 2002. https://doi.org/10.1126/science.1066168

[2] S. J. Luck, An Introduction to the Event-Related Potential Technique, 2nd ed. Cambridge, MA, USA: MIT Press, 2014.

[3] M. X. Cohen, Analyzing Neural Time Series Data: Theory and Practice. Cambridge, MA, USA: MIT Press, 2014. https://doi.org/10.7551/mitpress/9609.003.0001

[4] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics," J. Neurosci. Methods, vol. 134, no. 1, pp. 9–21, 2004. https://doi.org/10.1016/j.jneumeth.2003.10.009

[5] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, J. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," Frontiers in Neuroscience, vol. 7, p. 267, 2013, doi: 10.3389/fnins.2013.00267. https://doi.org/10.3389/fnins.2013.00267

[6] A. Gramfort, M. Luessi, E. Larson, et al., "MNE software for processing MEG and EEG data," NeuroImage, vol. 86, pp. 446–460, 2014. https://doi.org/10.1016/j.neuroimage.2013.10.027

[7] T. Koenig, L. Prichep, T. Dierks, et al., "Decreased EEG synchronization in alcoholism," Clin. Neurophysiol., vol. 116, no. 9, pp. 2209–2217, 2005.

[8] S. Rangaswamy and B. Porjesz, "Understanding alcohol use disorders with neuroelectrophysiology," Handbook of Clinical Neurology, vol. 125, pp. 383–414, 2014.

[9] B. Porjesz, H. Begleiter, J. Wang, et al., "Event-related brain potentials to visual stimuli in alcoholism," Alcohol, vol. 5, no. 6, pp. 465–469, 1988.

[10] J. A. C. Smolenaars, J. T. Miller, and M. van den Brink, "EEG biomarkers in alcohol use disorder," Neurosci. Biobehav. Rev., vol. 106, pp. 84–98, 2019.

[11] P. L. Nunez and R. Srinivasan, Electric Fields of the Brain, 2nd ed. Oxford, U.K.: Oxford Univ. Press, 2006. https://doi.org/10.1093/acprof:oso/9780195050387.001.0001

[12] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7

[13] V. Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer, 1995.

[14] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Mach. Learn., vol. 46, no. 1, pp. 389–422, 2002. https://doi.org/10.1023/A:1012487302797

[16] H. Abdi and L. J. Williams, "Principal component analysis," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 4, pp. 433–459, 2010.

[17] K. R. Muller, C. W. Anderson, and G. E. Birch, "Linear and nonlinear methods for brain–computer interfaces," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 11, no. 2, pp. 165–169, 2003.

[18] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller, "Introduction to machine learning for brain imaging," NeuroImage, vol. 56, no. 2, pp. 387–399, 2011. https://doi.org/10.1016/j.neuroimage.2010.11.004

[19] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based BCIs," J. Neural Eng., vol. 4, no. 2, pp. R1–R13, 2007. https://doi.org/10.1088/1741-2560/4/2/R01

[20] G. Dornhege, J. d. R. Millán, T. Hinterberger, D. J. McFarland, and K.-R. Müller, Toward Brain–Computer Interfacing. Cambridge, MA, USA: MIT Press, 2007.

[21] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," arXiv preprint arXiv:1511.06448, 2015. https://doi.org/10.48550/arXiv.1511.06448

[22] R. T. Canolty and R. T. Knight, "The functional role of cross-frequency coupling," Trends Cogn. Sci., vol. 14, no. 11, pp. 506–515, 2010.

[23] J. Onton and S. Makeig, "Information-based modeling of event-related brain dynamics," Prog. Brain Res., vol. 159, pp. 99–120, 2006. https://doi.org/10.1016/S0079-6123(06)59007-7

[24] M. Teplan, "Fundamentals of EEG measurement," Meas. Sci. Rev., vol. 2, no. 2, pp. 1–11, 2002.

[25] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization," Clin. Neurophysiol., vol. 110, no. 11, pp. 1842–1857, 1999. https://doi.org/10.1016/S1388-2457(99)00141-8

[26] J. S. Bendat and A. G. Piersol, Random Data: Analysis and Measurement Procedures, 4th ed. Hoboken, NJ, USA: Wiley, 2010.

[27] D. J. McFarland and J. R. Wolpaw, "EEG-based brain–computer interfaces," Curr. Opin. Biomed. Eng., vol. 4, pp. 194–200, 2017. https://doi.org/10.1016/j.cobme.2017.11.004

[28] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," Neural Netw., vol. 13, no. 4–5, pp. 411–430, 2000. https://doi.org/10.1016/S0893-6080(00)00026-5

[29] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "FieldTrip: Open source software for advanced analysis of MEG and EEG data," Comput. Intell. Neurosci., vol. 2011, Article ID 156869. https://doi.org/10.1155/2011/156869

[30] B. He, Z. Liu, D. J. Schalk, et al., "Electrophysiological imaging of brain activity and connectivity—Challenges and opportunities," IEEE Trans. Biomed. Eng., vol. 66, no. 3, pp. 853–870, 2019.

[31] H. Begleiter. "EEG Database," UCI Machine Learning Repository, 1995. [Online]. Available: https://doi.org/10.24432/C5TS3D.

[32] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," Journal of the Royal Statistical Society: Series B, vol. 57, no. 1, pp. 289–300, 1995. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[33] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity," Journal of Experimental Psychology: Human Learning and Memory, vol. 6, no. 2, pp. 174–215, 1980. https://psycnet.apa.org/doi/10.1037/0278-7393.6.2.174