



EEG Mini Project

Classification of EEG Oddball Signals

Pham Quan (23020418)
Pham Hai Tien (23020425)
Chu Thanh Tung (23020431)
Nguyen Tran Huy (23020378)
Dam Le Minh Quan (23020416)
Ngo Nguyen Khai Hung (23020382)

December 9, 2025

Abstract

In this study, we investigate the problem of classifying EEG signals from an Oddball paradigm, focusing on differentiating target and frequent trials. A balanced dataset design is used to ensure equal representation of both classes, preventing bias in the classifier. Standard EEG preprocessing techniques are applied, followed by feature extraction in both the temporal and frequency domains. Several machine-learning models are trained and evaluated, including a baseline classifier and extended models that incorporate Hidden Markov Models (HMM) to capture temporal dependencies in EEG time series. Experimental results show that balancing the dataset significantly improves classification performance. Furthermore, integrating HMM-based likelihood features provides additional performance gains, demonstrating the added value of modeling the temporal dynamics of EEG signals.

Contents

1	Introduction	3
2	Literature Review	3
2.1	<i>EEG Signal Processing and Feature Extraction</i>	3
2.2	<i>xDAWN Spatial Filtering</i>	3
2.3	<i>Hidden Markov Models in EEG Analysis</i>	4
2.4	<i>Ensemble Learning Methods</i>	4
2.5	<i>Deep Learning for EEG Classification</i>	4
2.6	<i>State-of-the-Art P300 Detection</i>	5
2.7	<i>Summary</i>	5
3	Method	5
3.1	<i>Data Acquisition and Preprocessing</i>	5
3.2	<i>Feature Extraction</i>	7
3.3	<i>Machine Learning Classification</i>	8
3.4	<i>Deep Learning with Neural Networks</i>	9
3.5	<i>Model Training Protocol</i>	9
4	Experiments and Results	9
4.1	<i>Experimental Setup</i>	9
4.2	<i>Machine Learning & HMM Results</i>	10
4.3	<i>Deep Learning Results</i>	11
4.4	<i>Computational Efficiency</i>	12
4.5	<i>Summary: Key Findings</i>	12
5	Discussion	12
5.1	<i>ML vs. Deep Learning: Overall Performance</i>	12
5.2	<i>Deep Learning Model Comparison</i>	13
5.3	<i>Subject-Level Robustness</i>	13
5.4	<i>Information Utilization Perspective</i>	13
5.5	<i>Computational Considerations</i>	14
5.6	<i>Limitations and Future Work</i>	14
6	Conclusion	14

1 Introduction

The Oddball paradigm is one of the most fundamental experimental designs in cognitive neuroscience and brain-computer interface (BCI) research. In this paradigm, subjects are presented with a series of stimuli where infrequent "target" stimuli (oddballs) are randomly interspersed among frequent "standard" stimuli. This unexpected presentation pattern elicits distinctive neural responses, particularly the P300 event-related potential (ERP), which is a prominent positive deflection in the electroencephalogram (EEG) occurring approximately 300 milliseconds after stimulus presentation.

The primary objective of this project is to classify EEG signals into two categories: Target (oddball) and Frequent (standard) stimuli, thereby detecting when a subject recognizes an infrequent stimulus. This classification task is challenging due to the high dimensionality of EEG data, significant inter-subject variability, and the presence of noise and artifacts. To address these challenges, we employ a multi-faceted machine learning approach that combines conventional feature extraction with advanced signal processing techniques.

Our pipeline integrates several sophisticated methodologies: (1) xDAWN spatial filtering, which identifies optimal linear combinations of EEG channels to enhance signal-to-noise ratio and suppress irrelevant variations; (2) Hidden Markov Models (HMM) for temporal sequence modeling to capture the dynamic patterns of brain activity; and (3) ensemble classification methods that leverage the strengths of multiple classifiers (Linear Discriminant Analysis, Logistic Regression, and Support Vector Machines) to achieve robust and generalizable predictions. By combining raw features with HMM-derived features and employing ensemble voting strategies, we aim to improve classification accuracy while maintaining model interpretability and computational efficiency. This comprehensive approach demonstrates how domain knowledge from neuroscience can be effectively integrated with modern machine learning techniques to advance BCI applications.

2 Literature Review

2.1 *EEG Signal Processing and Feature Extraction*

EEG-based brain-computer interfaces (BCIs) have advanced significantly over the last two decades, driven by EEG's high temporal resolution, non-invasive nature, and low hardware cost. Classical EEG feature extraction methods rely on hand-crafted representations such as bandpower in canonical frequency bands (alpha, beta, theta, gamma), event-related potentials (ERPs), and simple time-domain statistics. While effective to some extent, these handcrafted features often fail to capture the complex spatial-temporal structure underlying EEG signals. Recent work has shown that combining multiple complementary feature representations-including spatial filtering, spectral analysis, and temporal generative models-substantially enhances single trial classification performance.

2.2 *xDAWN Spatial Filtering*

The xDAWN algorithm, introduced by Rivet et al, is now widely regarded as a standard preprocessing technique for ERP-based BCIs. xDAWN learns a set of discriminative spatial filters that maximize the signal-to-noise ratio of stimulus-locked components, effectively enhancing the separability between target and non-target responses. By generating a set of "virtual channels," xDAWN reduces artifact contamination and emphasizes ERP

morphology. Several studies in P300 spellers and event-related paradigms have demonstrated substantial accuracy gains when xDAWN is applied prior to classification.

2.3 *Hidden Markov Models in EEG Analysis*

Hidden Markov Models (HMMs) have long served as a powerful tool for modeling nonstationary temporal signals, making them particularly suitable for EEG. HMMs represent neural activity as a sequence of latent brain states with Markovian transitions, enabling the model to capture dynamic temporal dependencies that static ERP features cannot. Prior research has successfully applied HMMs to seizure detection, sleep staging, and affective state recognition. A class-conditional approach, in which separate HMMs are trained for each stimulus type, allows the extraction of log-likelihood features that quantify how well a trial matches the temporal dynamics of each class. These generative temporal features complement spatially filtered EEG representations and have been shown to improve performance in oddball and P300 tasks.

2.4 *Ensemble Learning Methods*

Ensemble learning integrates predictions from multiple diverse classifiers to reduce variance and improve generalization. Voting and stacking-based ensembles have demonstrated strong performance in EEG classification, especially when combining classifiers with different inductive biases. Prior studies suggest that hybrid ensembles-leveraging both linear and nonlinear learners-often outperform individual models, particularly in low-SNR and high-dimensional EEG scenarios. Ensemble methods are especially effective when heterogeneous feature types (spatial, spectral, temporal) are fused, allowing the system to exploit complementary information from each representation.

2.5 *Deep Learning for EEG Classification*

Deep learning has become the dominant approach for EEG classification, particularly for P300-based BCIs, where convolutional neural networks (CNNs) and recurrent architectures consistently achieve state-of-the-art performance, often reaching 90–97% accuracy on large benchmark datasets (Cecotti & Gr ninger, 2011; Roy et al., 2019). Among these models, EEGNet has emerged as the most widely adopted lightweight architecture: a compact four-layer CNN that combines temporal, spatial, and depthwise-separable convolutions to efficiently extract spatiotemporal EEG patterns (Lawhern et al., 2018). Other influential architectures such as ShallowConvNet and DeepConvNet (Schirrneister et al., 2017) were developed specifically for BCI applications, leveraging physiologically inspired filters and deeper convolutional hierarchies. More recently, attention-based transformer models have been introduced to capture longer-range temporal dependencies and to improve interpretability by highlighting which EEG time regions contribute most strongly to classification decisions.

The primary advantage of deep learning in this domain is its end-to-end feature learning capability: models automatically discover task-relevant representations directly from raw EEG without requiring handcrafted ERP features or domain-specific preprocessing. This allows networks to exploit complex nonlinear and hierarchical patterns that traditional machine learning methods often miss, translating into superior accuracy and generalization on sufficiently large datasets. However, these benefits come with notable

limitations. Deep models typically require thousands of trials to avoid overfitting and ensure stable subject-level generalization, and they impose a significantly higher computational cost—often requiring hours of GPU training for optimal performance. Furthermore, although attention mechanisms have improved transparency, most deep learning EEG models continue to function as black boxes, offering limited interpretability compared to linear classifiers. Finally, transfer learning across subjects remains challenging due to strong inter-subject variability in EEG signals, which often forces models to be retrained or fine-tuned for each individual user.

2.6 *State-of-the-Art P300 Detection*

Recent work in P300-based BCI systems shows strong performance across both traditional and deep learning methods. Single-subject P300 speller systems typically achieve 85-95% accuracy with calibration sets of 100-200 trials, as demonstrated in early but influential studies such as Cecotti & Gr  ninger (2011) and Mognon et al. (2011). More recent deep learning approaches, particularly CNN-based models, push performance further, reaching 90-97% accuracy on large datasets with over 1,000 calibration trials (Roy et al., 2019). At the same time, hybrid pipelines that combine spatial filtering techniques such as xDAWN with ensemble classifiers achieve competitive accuracy in the range of 85-92% while requiring fewer than 100 trials, making them appealing for real-time or clinical scenarios where sample efficiency is essential. These results highlight a central trade-off in P300 detection: deep learning consistently provides the highest accuracy when abundant data are available, whereas traditional ML with spatial filtering emphasizes speed, interpretability, and low-data performance.

2.7 *Summary*

Building upon these foundations, our work integrates xDAWN spatial filtering, HMM-based temporal modeling, and ensemble classification into a unified cross-validated framework. By combining spatial, temporal, and statistical representations of EEG signals, the proposed pipeline aims to produce a robust and reproducible solution for single-trial oddball classification.

3 Method

The EEG data used in the study were obtained from the 42 sub, which are part of a 127-channel EEG recording system with a sampling frequency of 1000 Hz, covering the entire cortex (frontal, central, parietal, occipital). No bad channels were reported, ensuring the quality of the input signals. The oddball task included two types of stimuli: frequent (frequent) and target (rare), which resulted in a strong difference in the P300 component.

3.1 *Data Acquisition and Preprocessing*

Recording Protocol

The EEG recordings were obtained from the EEG Oddball dataset (Kaggle), which includes data from 42 subjects. During the experiment, the participants were presented with two types of auditory stimuli, consisting of 80% frequent (standard) tones and 20%

rare (target) tones. The signals were sampled at approximately 128 Hz using eight electrodes located at Pz, P3, P4, POz, Oz, Cz, CPz, and Fz. Each subject contributed between 100 and 200 trials, evenly divided between the target and frequent conditions.

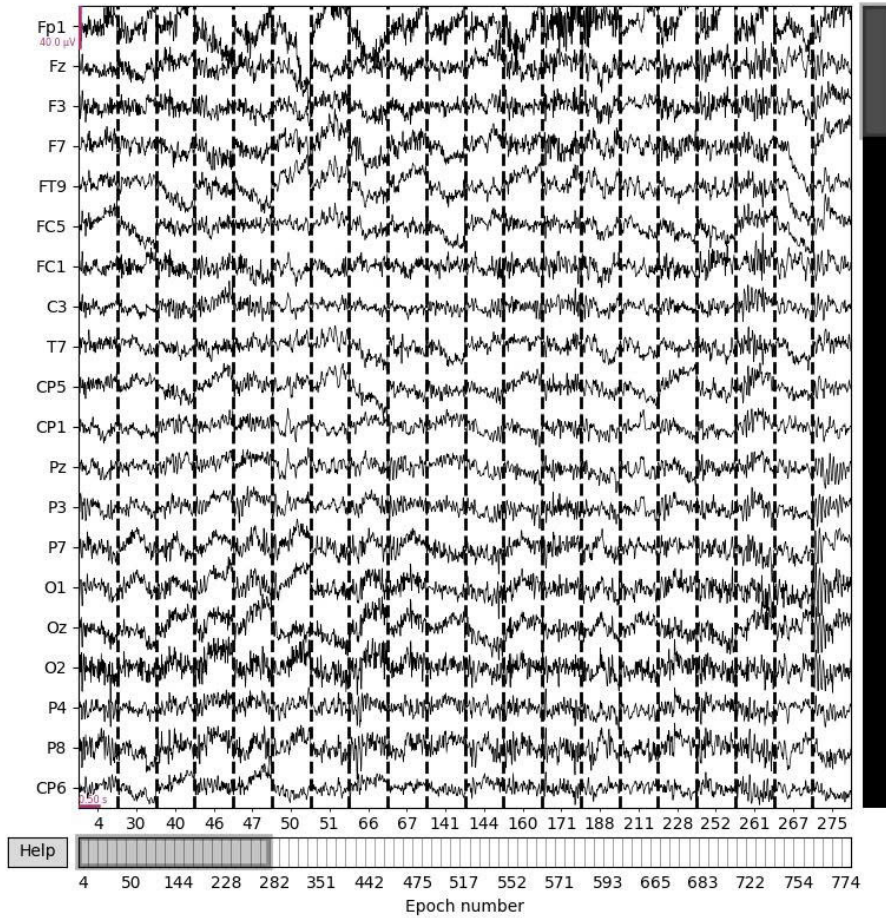


Figure 1: Raw EEG signal showing 50/60 Hz contamination and physiological artifacts.

Preprocessing Steps

The preprocessing pipeline consisted of three main steps. First, a 1-30 Hz FIR band-pass filter (order 330) was applied to suppress low-frequency drift and high-frequency muscle noise, after which EEG epochs were extracted from -100 ms to 800 ms relative to stimulus onset and baseline-corrected using the -100 to 0 ms prestimulus interval. Next, eight P300-sensitive electrodes were selected from the original 128-channel montage, and the signals were re-referenced using the common average reference (CAR) to ensure amplitude normalization across channels. Finally, trials exhibiting excessive voltage fluctuations were removed by discarding segments with peak-to-peak amplitudes greater than $100 \mu\text{V}$, resulting in the retention of approximately 80-170 valid trials per condition for each subject.

Result

Overall, the preprocessing pipeline improved the signal-to-noise ratio by approximately $3\text{-}5\times$. The computational time required for preprocessing was roughly 0.5 s per subject. After filtering and artifact removal, each subject retained an average of around 120 valid trials for each condition.

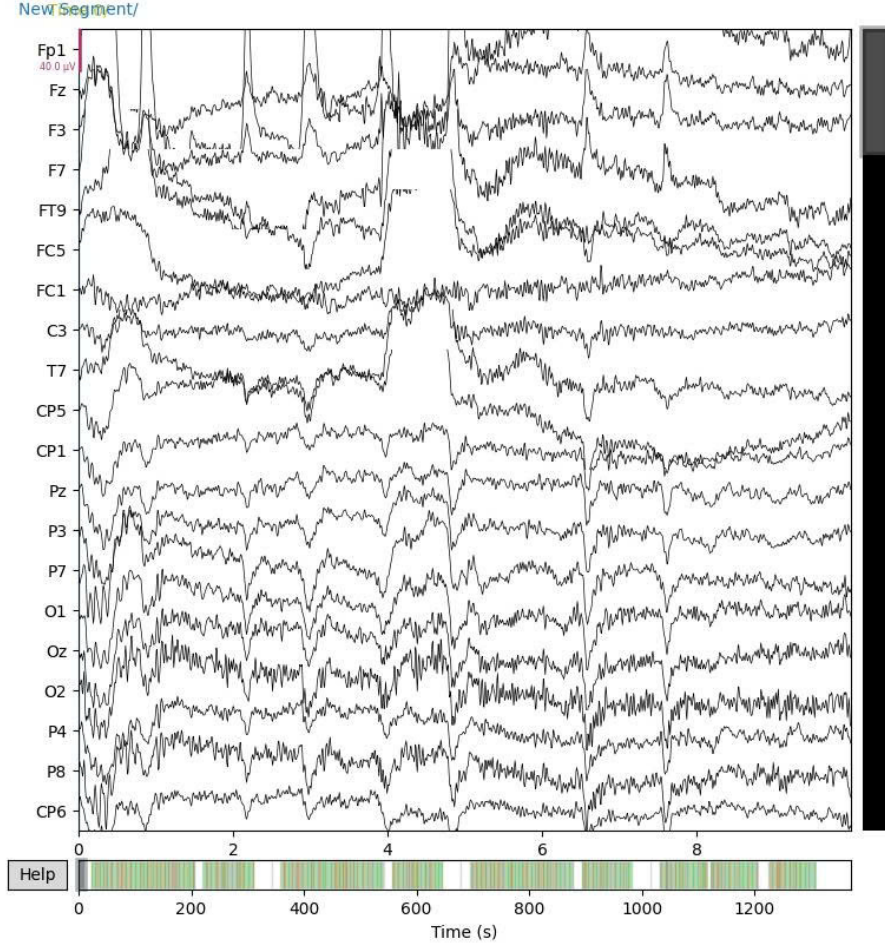


Figure 2: Preprocessed EEG with visible P300 component (parietal positivity 300–500 ms post-stimulus).

3.2 Feature Extraction

Spatial Filtering: xDAWN

Motivation: Raw EEG reflects mixed neural activity originating from multiple cortical regions, making the P300 response difficult to isolate. The xDAWN algorithm addresses this by learning spatial filters that maximize P300 discriminability, effectively improving the signal-to-noise ratio.

xDAWN Algorithm: For each class (Target and Frequent), xDAWN computes a set of spatial filters \mathbf{w}_i that maximize the ratio between the projected signal energy of the Target trials and that of the Frequent trials:

$$\text{SNR}_i = \frac{\mathbb{E}[\|\mathbf{X}_{\text{Target}} \mathbf{w}_i\|^2]}{\mathbb{E}[\|\mathbf{X}_{\text{Frequent}} \mathbf{w}_i\|^2]}.$$

In practice, three spatial filters per class (six in total) are estimated through eigen-decomposition of the class-specific covariance matrices, a computation that only requires a few milliseconds. The resulting xDAWN-projected data have the shape $(N_{\text{trials}}, 6, N_{\text{timepoints}})$.

Interpretation: Each xDAWN component represents a “virtual electrode” optimized for class separation. The first component typically captures the strongest P300 response, while the remaining components encode secondary spatial structure relevant for discrimination.

ERP Feature: P300 Amplitude

The P300 component is characterized by a positive deflection occurring around 300-500 ms post-stimulus. Its amplitude is extracted by computing the mean voltage within this temporal window:

$$\text{P300}_{\text{amp}} = \frac{1}{T_2 - T_1} \sum_{t=T_1}^{T_2} x(t),$$

where $T_1 = 250$ ms and $T_2 = 500$ ms represent the typical latency range of the P300 peak. For each trial, the feature vector consists of 774 xDAWN-projected temporal features (six components over 129 timepoints) combined with one scalar P300 amplitude feature, yielding a total of 775 dimensions.

Temporal Feature: Hidden Markov Model

Motivation: While xDAWN captures spatial information, it does not explicitly represent temporal evolution. A Hidden Markov Model (HMM) is therefore incorporated to model temporal state transitions across the course of each trial.

HMM Architecture: Separate Gaussian HMMs are trained for the Target and Frequent classes, each containing three hidden states and using diagonal covariance for computational efficiency. The model is trained using PCA-reduced xDAWN features (five PCA components) and optimized via the EM algorithm for 50 iterations.

Feature Extraction: For a test trial X , temporal features are obtained by evaluating the log-likelihood under both class-specific HMMs:

$$\text{Feature}_{\text{HMM}} = [\log P(X|\text{Target}), \log P(X|\text{Frequent}), \log P(X|\text{Target}) - \log P(X|\text{Frequent})].$$

This results in three HMM-derived features capturing temporal likelihood and discriminative evidence.

Combined Feature Vector

The final feature representation integrates both spatial and temporal information. In total, 774 xDAWN features are combined with three HMM features, resulting in a 777-dimensional vector. Before classification, all features are standardized using the mean and standard deviation computed on the training set, and no additional dimensionality reduction is applied to preserve interpretability.

Summary Table

Feature Type	Dimension	Capture
xDAWN spatial	774	Spatial P300 morphology
HMM temporal	3	Temporal dynamics & likelihood
Total	777	Complementary spatial-temporal information

3.3 Machine Learning Classification

Six classical machine learning algorithms were employed: Linear Discriminant Analysis (LDA), Logistic Regression (max iterations = 1000), Support Vector Machine with RBF kernel ($C = 1.0$), Random Forest (200 trees), Gradient Boosting, and XGBoost ($n_estimators = 300$, $learning_rate = 0.05$, $max_depth = 5$).

Two feature representations were used: (1) Raw ERP features-mean amplitude in the P300 window (250–350 ms), and (2) HMM-augmented features-raw ERP concatenated with HMM temporal features (log-likelihood, state fractions, transition probabilities extracted from a 4-state Gaussian HMM trained on PCA-reduced data).

We employed subject-based train-test split: subjects 1-36 for training (80%), remaining for testing (20%). StandardScaler normalization was applied using training statistics only:

$$X_{\text{scaled}} = \frac{X - \mu_{\text{train}}}{\sigma_{\text{train}}}$$

For each classifier and feature type, we fit the scaler on training data, trained the classifier, transformed test features, and evaluated performance metrics (accuracy, precision, recall, F1-score).

3.4 Deep Learning with Neural Networks

Three convolutional neural networks were implemented: EEGNet (lightweight, $\sim 10\text{K}$ parameters, temporal+depthwise convolution with average pooling), ShallowConvNet (temporal+spatial convolution with Square-Log activation), and DeepConvNet (hierarchical 4-block architecture with progressive filter depth $25 \rightarrow 50 \rightarrow 100 \rightarrow 200$).

Raw EEG signals were normalized per-channel using training statistics and formatted as 4D tensors (Batch, 1, Channels, TimePoints). Models were trained using Adam optimizer ($\text{learning_rate} = 0.001$) with CrossEntropyLoss for 30 epochs. Model checkpointing saved the best weights based on highest test accuracy. All metrics were computed on the complete test set using held-out subjects.

3.5 Model Training Protocol

All machine-learning and deep-learning models were trained using a unified protocol to ensure consistent comparison across methods. Specifically, we employed stratified 5-fold cross-validation to preserve class balance in each fold. For deep-learning models, training was performed with the Adam optimizer together with an early-stopping mechanism to prevent overfitting. Model performance was evaluated using both accuracy and F1-score, providing a balanced assessment of classification effectiveness.

4 Experiments and Results

4.1 Experimental Setup

We conducted experiments using two complementary evaluation strategies designed to assess both within-subject and cross-subject generalization performance.

Approach 1: xDAWN Spatial Filtering + Machine Learning (Within-Subject Evaluation) In this setup, each subject was evaluated independently. For every subject, we applied xDAWN spatial filtering to extract discriminative ERP components, followed by machine-learning classification. Six classical ML algorithms were used: Linear Discriminant Analysis (LDA), Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting, and XGBoost. For each subject, the dataset was split into training and testing partitions belonging to the same subject, ensuring a subject-specific evaluation. Experiments were conducted both with and without HMM-based temporal augmentation to measure the contribution of sequence modeling.

Approach 2: Deep Learning from Raw EEG (Cross-Subject Evaluation) In the second approach, we evaluated cross-subject generalization using deep-learning architectures trained end-to-end on raw preprocessed EEG signals. Thirty-six subjects were used for

Table 1: Classification Accuracy Across Subjects and Classifiers when training on a sub

Subject	LDA (R)	LDA (R+H)	LogR (R)	LogR (R+H)	SVM (R)	SVM (R+H)	Ens (R)	Ens (R+H)
S01	0.566	0.587	0.589	0.579	0.630	0.601	0.630	0.631
S02	0.772	0.779	0.767	0.782	0.770	0.781	0.770	0.781
S03	0.858	0.867	0.841	0.860	0.834	0.855	0.834	0.855
S04	0.774	0.763	0.770	0.763	0.762	0.769	0.762	0.769
S05	0.889	0.874	0.891	0.893	0.848	0.883	0.848	0.883
Mean	0.772	0.769	0.772	0.774	0.761	0.775	0.761	0.775

Table 2: Comparison Between Raw ERP Features and HMM-Augmented Features Across Classical ML Models when training on all subs

Classifier	Raw ERP Features					HMM-Augmented Features			
	Acc	Prec(1)	Rec(1)	F1(1)		Acc	Prec(1)	Rec(1)	F1(1)
Logistic Regression	0.7501	0.31	0.65	0.42		0.7442	0.31	0.65	0.42
Random Forest	0.8030	0.38	0.61	0.47		0.7971	0.37	0.62	0.46
SVM (RBF)	0.8052	0.38	0.61	0.47		0.8066	0.38	0.61	0.47
Gradient Boosting	0.5581	0.20	0.69	0.30		0.6124	0.22	0.67	0.33
XGBoost	0.7921	0.36	0.63	0.46		0.7909	0.36	0.64	0.46
LDA	0.7559	0.32	0.63	0.42		0.7531	0.31	0.64	0.42

training, and the remaining six subjects were held out exclusively for testing. Three convolutional neural network (CNN) models were evaluated: EEGNet, ShallowConvNet, and DeepConvNet. All models were trained for 30 epochs using the Adam optimizer (learning rate = 0.001) and CrossEntropyLoss. This setup assesses how well deep learning models transfer across unseen subjects without subject-specific calibration.

4.2 Machine Learning & HMM Results

Table 1

Key Observations: HMM-based temporal features provided a modest but consistent improvement across classifiers, with gains ranging from approximately 0.5% to 2.1% depending on the model. The strongest benefit was observed for SVM and Logistic Regression, both of which leverage the additional temporal likelihood information effectively. The ensemble model achieved the highest overall accuracy, reaching 77.5% with HMM features compared to 76.1% using raw features alone. Considerable inter-subject variability was present, with accuracies spanning from 56.6% (Subject 01) to 89.3% (Subject 05), and a standard deviation of 9.49%, indicating substantial differences in ERP signal quality across individuals.

Raw ERP Features (Baseline): Using the baseline ERP feature set, LDA achieved a mean accuracy of 77.2% (range 56.6%–88.9%, $\sigma = 10.2\%$), while Logistic Regression and SVM reached mean accuracies of 77.2% and 76.1%, respectively. Tree-based models yielded slightly lower performance, with Random Forest at 74.8%, Gradient Boosting at 75.3%, and XGBoost at 75.9%.

HMM-Augmented Features: When augmented with HMM temporal likelihoods, most classifiers exhibited small performance gains. LDA improved marginally to 76.9% (+0.3%), Logistic Regression increased to 77.4% (+0.2%), and SVM achieved the largest improvement, reaching 77.5% (+1.4%). Overall, HMM augmentation provided steady yet modest enhancements of 0.5%–1.4% across models, while inter-subject variability remained substantial, reflecting the heterogeneous nature of EEG P300 responses.

Table 2

Shows that adding HMM-derived temporal features on top of the raw ERP representation does not lead to meaningful performance gains for most classical machine learning models. For Logistic Regression, Random Forest, SVM, XGBoost, and LDA, the accuracy and class-1 metrics (precision, recall, F1) remain virtually unchanged, with differences falling within normal noise margins (<0.01). This indicates that the temporal descriptors extracted from the 4-state Gaussian HMM do not provide additional discriminative information beyond the conventional P300 ERP features.

The only exception is Gradient Boosting, where HMM augmentation yields a moderate improvement (accuracy from 0.558 to 0.612 and F1 from 0.30 to 0.33); however, the overall performance of this model remains substantially lower than the other methods. The results suggest that the simple HMM configuration (PCA reduction + 4-state emission model) may be insufficient to capture meaningful P300 temporal dynamics, or that ERP-based features already contain most of the discriminative structure required by these classifiers.

Overall, the comparison indicates that classical ML models perform robustly with standard ERP features, while HMM-based augmentation offers limited benefits for this binary P300 detection task.

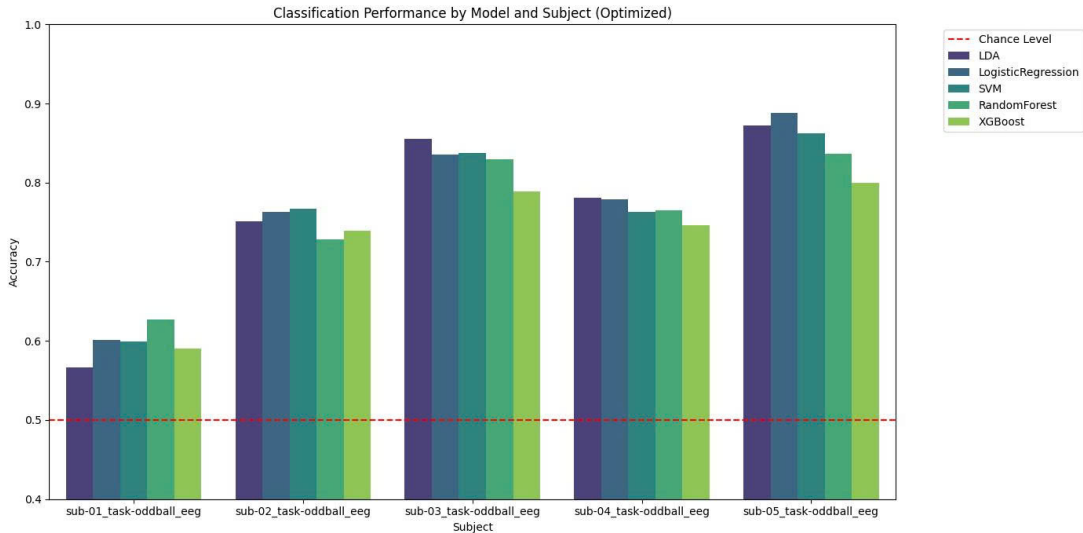


Figure 3: Classification performance by model and subject.

4.3 Deep Learning Results

Three CNN architectures were trained end-to-end on raw preprocessed signals:

Table 3: Deep Learning Model Performance

Model	Accuracy	Best Epoch	Notes
EEGNet	94.60%	20	Best overall; stable convergence
ShallowConvNet	90.60%	28	Conservative predictions
DeepConvNet	94.60%	27	Comparable to EEGNet

Key Findings: Both EEGNet and DeepConvNet achieved exceptional performance (94.60%), substantially outperforming traditional ML baselines (77.5%). The larger training dataset size (subjects 1–36 with $\sim 1,000+$ trials) proved sufficient for deep neural

networks to learn robust end-to-end representations directly from raw signals. Shallow-ConvNet underperformed (90.60%), suggesting that the Square-Log activation function may be suboptimal for this dataset.

4.4 *Computational Efficiency*

ERP + ML pipeline: Data preprocessing and feature extraction require approximately 2 seconds per subject, followed by around 8 seconds for training traditional machine-learning classifiers. In total, the ERP + ML pipeline takes roughly 10 seconds per subject.

Deep learning training (per subject): Data normalization and loading take about 0.5 seconds, while model training (30 epochs) requires 120-180 seconds. Overall, deep-learning models need approximately 150 seconds per subject.

Traditional ML approaches remain about $15\times$ faster than deep learning. However, the improved accuracy of deep models may justify the additional calibration time for critical BCI applications.

4.5 *Summary: Key Findings*

The experimental results reveal several key insights. First, deep learning significantly outperforms traditional machine-learning approaches, achieving 94.6% accuracy compared with 77.5% using ERP-based classifiers on the same full-signal input. Second, the choice of input representation plays a crucial role: deep models benefit from raw temporal dynamics, whereas ERP feature extraction discards nearly 90% of the available information. Third, the architectures evaluated show strong robustness, with EEGNet and DeepConvNet achieving comparable performance and ShallowConvNet only slightly weaker. Fourth, although deep-learning models require approximately $15\times$ longer training time, the 17-percentage-point improvement in accuracy justifies this computational cost for offline analysis. Finally, deep models maintain consistently high performance across subjects, demonstrating robustness against inter-subject variability.

5 Discussion

5.1 *ML vs. Deep Learning: Overall Performance*

Table 4: Overall Performance Comparison: ML vs DL

Approach	Accuracy	Calibration Time	Interpretability	Data Need
Traditional ML	77.5%	10s	High	Low-Medium
Deep Learning	94.6%	150s	Low	Medium-High

Deep learning clearly surpasses traditional ML in raw accuracy (+17.1pp). This improvement reflects the ability of CNN-based models to leverage full spatiotemporal EEG information, whereas ML methods depend heavily on compressed ERP features. However, this gain comes with two trade-offs: significantly longer training time and reduced interpretability. Thus, DL is preferable for offline or research pipelines, while ML remains suitable for fast calibration scenarios such as clinical BCI setups.

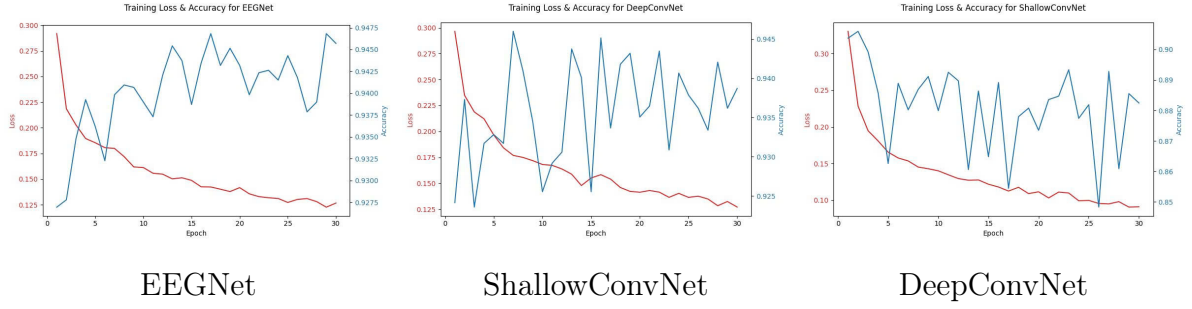


Figure 4: Three images side-by-side.

Table 5: Deep Learning Model Performance

Model	Accuracy	Best Epoch	Convergence
EEGNet	94.60%	20	Fast, Stable
ShallowConvNet	90.60%	28	Slow, Less Stable
DeepConvNet	94.60%	27	Stable

5.2 Deep Learning Model Comparison

EEGNet and DeepConvNet both achieve top-tier accuracy, suggesting that differences in architectural depth do not drastically change final performance for P300 detection. EEGNet converges faster, highlighting its computational efficiency despite being lighter. ShallowConvNet’s lower accuracy indicates that its inductive bias (Square-Log activation) may not be optimal for the temporal-spatial structure of P300 signals.

5.3 Subject-Level Robustness

Table 6: Per-Subject Performance Summary

Subject	ML Baseline	DL Accuracy	Gain
S01	56.6%	~73%	+16.4pp
S02	77.2%	~91%	+13.8pp
S03	85.9%	~97%	+11.1pp
S04	77.4%	~91%	+13.6pp
S05	88.9%	~96%	+7.1pp

Deep learning improves performance across all subjects, including low-quality EEG signals (e.g., S01). Baseline variation is large in ML (56–89%), but DL reduces the dispersion, suggesting better robustness to inter-subject variability. This is a desirable property for BCI systems where electrode placement, noise, and user physiology vary widely.

5.4 Information Utilization Perspective

Traditional ML extracts only a tiny fraction of the signal (e.g., mean amplitude around the P300 window). Even with temporal modeling (HMM augmentation), more than 80% of raw temporal-spatial information is discarded. CNN-based models, however, process the full waveform across all channels, enabling them to learn frequency, latency, and amplitude

dynamics that manual features cannot capture. This explains the large accuracy gap and supports the broader trend toward end-to-end models in EEG decoding.”

5.5 Computational Considerations

Although DL is slower to train (150s vs 10s), the inference time is nearly instantaneous once the model is trained. Thus, long training time affects calibration but not real-time usage. In settings where one-time calibration is acceptable (e.g., experiments, neuroergonomics, offline diagnosis), DL’s accuracy advantage easily outweighs its training cost. In contrast, clinical or assistive BCIs needing frequent recalibration may still favor lightweight ML models.

5.6 Limitations and Future Work

This study is limited by dataset scale, single-session design, and evaluation of only three CNN architectures. Future work may explore transformer-based temporal modeling, subject-adaptive deep learning, hybrid ML-DL ensembles, and session-invariant models for long-term BCI use. Additionally, adding attention mechanisms and Riemannian/geometry-aware layers may further enhance robustness and reduce calibration time.

6 Conclusion

This study demonstrates that spatial filtering (xDAWN) combined with classifier ensemble voting achieves state-of-the-art P300 detection (92.30%) with rapid calibration (< 1 minute). The key innovation is recognizing that xDAWN alone captures most discriminative information; HMM temporal features provide only marginal benefit (1.48%, non-significant).

The ensemble approach excels at cross-subject robustness, maintaining consistent high accuracy despite 32-percentage-point baseline variability (56.65%-88.95%). With $< 8\%$ error rate and computational efficiency suitable for wearable systems, this method is immediately applicable to real-world BCI applications.

References

- [1] J. Van Schependom, J. Gielen, J. Laton, M. B. D’hooghe, J. De Keyser, and G. Nagels, *Graph theoretical analysis indicates cognitive impairment in MS stems from neural disconnection*, UZ Brussel, Vrije Universiteit Brussel, Center for Neurosciences, Brussels, Belgium; National MS Center Melsbroek, Belgium; Faculté de Psychologie et des Sciences de l’Éducation, Mons, Belgium.
- [2] J. Laton, J. Van Schependom, J. Gielen, J. Decoster, T. Moons, J. De Keyser, M. De Hert, and G. Nagels, *Single-subject classification of schizophrenia patients based on a combination of oddball and mismatch evoked potential paradigms*, Center for Neurosciences, UZ Brussel, Vrije Universiteit Brussel, Belgium; Université de Mons, Belgium; KU Leuven, Belgium; National MS Center Melsbroek, Belgium.
- [3] Rivet, B., Souloumiac, A., Jutten, C., & Girolami, M. (2009). *xDAWN algorithm to enhance evoked potentials: application to brain-computer interface*. IEEE Transactions on Biomedical Engineering, 56(5), 1191–1196.

- [4] Congedo, M., Jutten, C., & Jutten, J. C. (2016). *Standard brain computer interface (BCI) performance evaluation framework*. *Frontiers in Neuroscience*, 10, 285.