Assessing EEG Classification Performance: Classical vs Deep Learning Approaches

Pavitha Dissanayake
Department of Computer Science and Engineering
University of Moratuwa
pavitha.21@cse.mrt.ac.lk

Dr. Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa
ruthaya@cse.mrt.ac.lk

***Abstract*—Brain–Computer Interfaces (BCIs) enable direct communication between the brain and external systems through the decoding of neural activity, typically recorded via Electroencephalography (EEG). Despite their potential, EEG decoding faces challenges such as low signal-to-noise ratio, inter-subject variability, and limited data availability. While deep learning promises automated feature extraction and superior representational power, its performance often degrades in small datasets due to overfitting. This study evaluates the comparative effectiveness of classical machine learning and deep learning pipelines on the EEG-ExPy benchmark across three paradigms, N170, P300, and SSVEP. We replicate baseline EEG-ExPy results, introduce fine-tuned classical pipelines, and implement several autoencoder-based deep models. Our findings reveal that deep architectures consistently underperform due to limited data and poor generalization, whereas classical covariance-based and tangent space models achieve stable, high performance, particularly within SSVEP tasks (mean AUC $\approx$ 0.95 single-subject, 0.86 all-subject). These results emphasize that, under data-scarce conditions, carefully tuned classical approaches remain more robust, interpretable, and data-efficient than deep learning models for EEG-based BCI decoding.**

***Index Terms*—Brain–Computer Interface (BCI), Electroencephalography (EEG), EEG-ExPy, Tangent Space, Covariance Matrices, Riemannian Geometry, Deep Learning, Autoencoders, SSVEP, P300, N170, Machine Learning**

I. Introduction

Brain–Computer Interfaces (BCIs) enable direct communication between the human brain and external devices by translating neural activity into executable signals. Among neuroimaging modalities, Electroencephalography (EEG) is widely preferred in many experimental setups because it is non-invasive, affordable, and offers high temporal resolution. Still, decoding reliable, meaningful signals from EEG is hard: low signal-to-noise, non-stationarity, session- and subject-to-subject variability, and often small datasets present serious barriers to applying more powerful models like deep learning.

Classical machine-learning approaches, such as covariance-based feature extraction (e.g., ERP covariance, TangentSpace, Xdawn) followed by linear classifiers (LDA, logistic regression, MDM), have traditionally been used to get baseline performance. In contrast, deep learning and hybrid architectures (convolutional neural networks, autoencoders, etc.) promise automated spatio-temporal feature learning, but they tend to require large datasets to generalize well. In practice, when data is scarce or when tasks aren't large motor imagery datasets, deep models often overfit or fail to outperform strong classical baselines.

In this work, we investigate whether it is feasible to improve over the baseline in small-to-moderate-size EEG datasets using deep learning, and if not, whether careful tuning of classical machine learning pipelines can still yield meaningful gains. We use the EEG-ExPy framework as our experimental benchmark; as EEG-ExPy does not include motor imagery tasks, our experiments focus on other ERP/SSVEP/N170 paradigms. It provides consistent data and pipelines that make baseline comparisons credible.

We conduct experiments across three task types: N170, P300, and SSVEP. For each task, we compare baseline classical ML pipelines (from the library) with both

(i) deep learning architectures (autoencoders, CNNs, GRU hybrids) and

(ii) fine-tuned classical classifiers (modifying feature extraction, regularization, model hyperparameters, etc.).

Our aim is to make incremental improvements to the baseline models.

Our findings indicate that for N170 and P300 paradigms, generalization is poor for both baseline methods and tuned models across subjects/sessions; models often barely surpass chance when data is pooled. However, in the SSVEP paradigm, both baseline and tuned classical ML methods generalize considerably better, with tuned classical classifiers achieving the strongest and most reliable performance. Deep learning models did not outperform classic pipelines reliably in any setting under our data limitations.

Our findings suggest that while deep learning shows potential, its benefits are limited under small EEG datasets such as EEG-ExPy. In contrast, well-tuned classical pipelines consistently outperform or match these models with far less complexity. These results highlight the continued relevance of classical EEG decoding methods and emphasize the importance of model interpretability and data efficiency in real-world BCI applications.

II. Related Work

Recent developments in EEG-based Brain–Computer Interfaces (BCIs) have focused on enhancing the decoding of neural responses such as event-related potentials (ERPs) and steady-state visual evoked potentials (SSVEPs). Traditional classifiers, including Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Task-Related Component Analysis (TRCA), have historically achieved strong within-subject accuracy but often fail to generalize across subjects or paradigms. To overcome these challenges, deep learning and hybrid architectures have been explored for their capacity to model the complex spatio–temporal structure of EEG signals.

Research into N170 components, primarily associated with face and visual stimulus recognition, initially relied on handcrafted spectral features, which were limited by low signal-to-noise ratios and poor trial-level discrimination. More recent work applies convolutional and recurrent networks to learn discriminative temporal–spatial patterns directly from raw EEG [1], [2]. For example, Liu et al. [1] proposed a compact CNN architecture that achieved notable accuracy improvements across visual paradigms, while Zhang et al. [3] demonstrated that spatial PCA can significantly enhance SVM-based EEG decoding. These results highlight the benefits of feature learning but also reveal that generalization across paradigms and subjects remains a persistent limitation.

In P300 classification, numerous studies have sought to improve detection robustness for target responses. Classical methods such as stepwise LDA and Bayesian classifiers have gradually been supplemented by hybrid and multi-scale architectures. Borra et al. [4] introduced MS-EEGNet, a lightweight CNN that maintained high decoding accuracy with reduced computational cost, while Afrah et al. [5] proposed an unsupervised CLSTM autoencoder that improved cross-session stability. Domain adaptation and data fusion methods, such as those of Du and Li [6] and Ermaganbet et al. [7], extended this further through centralized multi-person datasets and dual-input CNNs. Recent advances in self-supervised and transfer learning, such as SpellerSSL [8], show that representation learning can enhance robustness in speller BCIs, though they remain computationally demanding and data-hungry.

For SSVEP paradigms, canonical correlation analysis (CCA) and TRCA have long served as reliable baselines, particularly for real-time frequency detection. Deep architectures, however, have recently shown strong performance in capturing temporal and spectral dependencies. Models such as SSVEPformer [9], SSVEP-TFFNet [10], and xLSTM-based spatial–attention frameworks [11] demonstrate that transformer and attention-based mechanisms can effectively decode periodic neural dynamics. Yet, their training requirements and high parameter counts make them impractical for smaller datasets or low-resource settings.

Despite the growing success of deep learning, several critical gaps remain. Many architectures are optimized for specific paradigms, limiting their cross-domain adaptability. Furthermore, the lack of standardized preprocessing and evaluation practices reduces reproducibility and complicates fair comparisons. The EEG-ExPy benchmark directly addresses these issues by providing an open-source, reproducible platform for consistent EEG-based experimentation across paradigms including N170, P300, and SSVEP. However, few studies have conducted systematic comparisons of deep and classical methods within this standardized framework.

This research builds upon that gap by empirically evaluating both deep and classical models within EEG-ExPy, emphasizing reproducibility, generalization, and data efficiency. Unlike prior work that prioritizes complex architectures, this study investigates whether carefully tuned classical pipelines, featuring covariance-based representations, tangent space mapping, and regularized classifiers can achieve competitive performance on small EEG datasets. This perspective contributes to a more balanced understanding of where model complexity truly adds value in EEG-based BCI research.

III. Methodology

A. Dataset

This study utilizes the EEG-ExPy benchmark developed by NeuroTechX, an open-source framework providing standardized EEG datasets and reproducible pipelines across multiple paradigms. EEG-ExPy consolidates several well-known public datasets into a unified structure, supporting the N170, P300, and SSVEP and more experimental paradigms, each representing distinct neural responses commonly used in Brain–Computer Interface (BCI) research.

- N170 datasets capture event-related potentials associated with face and object perception, typically involving visual stimuli of human faces and scrambled images.
- P300 datasets include oddball paradigms designed to elicit target and non-target responses, widely used in speller and attention-related BCI tasks.
- SSVEP datasets record periodic visual responses evoked by flickering stimuli at fixed frequencies, supporting frequency-based command classification.

Each paradigm provides multi-subject data, pre-segmented into trials with corresponding event labels. EEG signals are recorded at sampling rates between 250 Hz and 512 Hz, depending on the original dataset, and include 32–64 scalp electrodes following the 10–20 international system.

All datasets were accessed and preprocessed using the built-in EEG-ExPy pipelines to ensure reproducibility and consistency across paradigms. These pipelines perform standardized filtering, channel normalization, and epoch segmentation, providing a common baseline for algorithmic evaluation.

B. Experimental Setup

All experiments were conducted using the EEG-ExPy framework and its accompanying Jupyter notebooks as the base environment. Rather than building independent scripts, the project extended and modified the official EEG-ExPy examples to ensure full compatibility with its preprocessing, data-loading, and evaluation pipelines. This approach allowed consistent reproducibility across paradigms and minimized the introduction of uncontrolled preprocessing differences.

The environment was implemented in Python 3.10 using the packages bundled within EEG-ExPy (NumPy, MNE, scikit-learn, and PyTorch). All analyses were executed on a workstation running Windows 11 with Intel i7 CPU, 16 GB RAM.

The experimental notebooks were customized to each available paradigms to load and use the limited subject/session sets in EEG-ExPy.

Each paradigm, N170, P300, and SSVEPwas evaluated using the standard EEG-ExPy preprocessing routines, including band-pass filtering, epoch segmentation, and normalization. For consistency, we retained the framework's default filtering ranges (typically 1–30 Hz) and trial lengths.

The workflow was organized into three primary experimental stages:

- **Baseline Replication**: The provided EEG-ExPy classification baselines were first replicated to confirm that results matched the original benchmark outputs.
- **Model Integration and Testing**: Custom pipelines were then inserted into the EEG-ExPy framework, enabling rapid switching between classical and deep models within the same data flow. This ensured identical preprocessing and evaluation procedures across approaches.
- **Evaluation and Logging**: Each model was evaluated using the benchmark's default cross-validation scheme, and results including per-subject accuracy, area under the curve, and runtime, were logged for comparison. All training and testing procedures were deterministic (with fixed random seeds) to maintain reproducibility.

By leveraging EEG-ExPy's reproducible design and aligning all experiments within a controlled notebook environment, the study ensured that any observed performance differences stemmed from model behavior rather than preprocessing or implementation inconsistencies.

C. Model and Algorithmic Framework

The experimental design incorporated both classical signal-processing pipelines provided within EEG-ExPy framework and deep learning architectures extended from the EEG-ExPy framework. This hybrid approach allowed a systematic comparison between feature-based and end-to-end learning strategies across multiple paradigms (N170, P300, and SSVEP).

1. Classical Pipelines

For the classical machine-learning setting, experiments were primarily based on EEG-ExPy's built-in configurations utilizing Common Spatial Patterns (CSP) and covariance-based feature extraction. These methods transform raw EEG trials into low-dimensional spatial representations, improving the separability of class-specific brain responses.

The extracted features were then passed through standard classifiers, including:

- Regularized Linear Discriminant Analysis (RegLDA)
- Minimum Distance to Mean (MDM)
- Logistic Regression (LR)

Each classifier was evaluated both independently and within scikit-learn pipelines, such as CSP + RegLDA or Cov + Tangent Space + LR, ensuring reproducibility and streamlined comparison. Hyperparameters (e.g., shrinkage factor for LDA, CSP filter count) were tuned minimally to maintain the benchmark's integrity while optimizing validation accuracy.

2. Deep Learning Architectures

To explore unsupervised and end-to-end representation learning, a series of autoencoder-based architectures were implemented using PyTorch.

These models were designed to learn robust, compressed representations of EEG data, with several architectural and regularization variants:

- **Autoencoder (AE)**: A baseline EEGAutoencoder trained for reconstruction-based feature learning and dimensionality reduction.
- **Denoising Autoencoder (DAE)**: A DenoisingEEGAutoencoder trained to recover clean signals from noisy inputs, improving robustness.
- **Sparse Autoencoder (SAE)**: The SparseDenoisingEEGAutoencoder, incorporating an L1 sparsity penalty on the latent layer to promote compact and discriminative feature selection.
- **1D Convolutional Layers**: Both encoder and decoder were constructed using nn.Conv1d and nn.ConvTranspose1d, exploiting temporal structure inherent in EEG sequences.

Training and Regularization

Training employed a combination of standard and custom regularization techniques to improve generalization on limited data:

- **Data Augmentation**: Applied via an augment_batch() routine, including Gaussian noise injection, temporal shifting, amplitude scaling, and channel dropout.
- **Dropout (0.2)**: Used in encoder layers to reduce overfitting.
- **Batch Normalization**: Stabilized and accelerated convergence.
- **L1 Regularization**: Applied to latent activations to enforce sparsity.
- **Loss Function**: Mean Squared Error (MSE) reconstruction loss.
- **Optimizer**: Adam with learning rates between 1e-3 and 1e-4.

Each model was trained for up to a few dozen epochs with early stopping based on validation loss, given the limited EEG-ExPy sample sizes.

3. Evaluation Integration

Both classical and deep-learning models were wrapped into EEG-ExPy's standard evaluation interface, ensuring consistent preprocessing, cross-validation, and metric computation.

Performance was primarily assessed using classification accuracy, with confusion matrices and loss curves for interpretability.

This unified modeling framework enabled a fair and transparent comparison between interpretable classical pipelines and representation-based autoencoder systems, illustrating how much can be gained from architectural and regularization choices under constrained EEG data settings.

D. Procedure

The experimental workflow was structured to systematically compare classical and deep learning EEG decoding pipelines across three of the paradigms provided by EEG-ExPy: N170, P300, and SSVEP. Each paradigm was evaluated under two conditions: single subject, single session and all subjects, all sessions, providing insights into both within-subject performance and multi-subject generalization.

1. Classical Pipeline Evaluation

For classical feature-based pipelines, each paradigm was processed following the default EEG-ExPy notebooks, including the preconfigured covariance-based transformations and CSP mappings. Model evaluation was conducted using k-fold cross-validation, with the number of folds set to match the baseline configuration for each paradigm: 20 folds for N170, 10 folds for P300, and 20 folds for SSVEP.

Hyperparameters for tuned models (e.g., regularization strength, number of filters) were manually adjusted to improve validation performance. Performance metrics were primarily reported as mean, standard deviation, and maximum AUC across folds. This setup allowed a fair comparison against the baseline pipelines provided by EEG-ExPy.

2. Deep Learning Evaluation

Deep learning models (AE, DAE, SAE, and 1D CNN variants) were trained using the same EEG data splits. Standard preprocessing, such as channel-wise normalization, was applied to ensure model stability. Some experiments employed 20-fold cross-validation, while others trained autoencoder models without explicit class labels, focusing on unsupervised feature reconstruction.

Accuracy served as the evaluation metric for deep learning models, although direct comparison with classical pipelines was limited due to metric differences. Across all paradigms, deep models failed to surpass chance-level performance, particularly in the low-data N170 and P300 experiments.

3. Experimental Steps

The general procedure for all experiments followed a consistent sequence:

- **Data Loading**: EEG epochs were loaded for the selected subjects and sessions.
- **Preprocessing**: Minimal preprocessing was applied, mimicking the baseline pipeline.
- **Model Selection**: Classical pipelines or deep learning architectures were instantiated.
- **Training**: Models were trained using the prescribed folds, with early stopping applied to prevent overfitting for deep networks.
- **Evaluation**: Metrics were computed on held-out folds, and cross-fold averages were recorded.
- **Result Logging**: Performance statistics, including mean, standard deviation, and maximum AUC or accuracy, were stored for analysis.

This structured procedure ensured consistency across paradigms and modeling approaches, providing a controlled framework for assessing the relative merits of classical and deep learning approaches on the EEG-ExPy datasets.

IV. Results

This section presents the performance of the evaluated pipelines and models on the three EEG paradigms, N170, P300, and SSVEP, under both single-subject single-session and all-subject all-session conditions. Classical pipelines were evaluated using the Area Under the Curve (AUC) metric to maintain consistency with the EEG-ExPy benchmark. Deep learning models, in contrast, were evaluated using classification accuracy due to their different training objectives.

A. N170 Paradigm

Table 1 summarizes the AUC performance across models for the N170 visual ERP task. The XdawnCov + TS and ERPCov + TS pipelines consistently outperformed other configurations. Tuning of covariance-based tangent space (TS) models provided marginal improvements over baselines, achieving up to 0.724 mean AUC for the single-subject setting.

TABLE I
N170 Classification Performance (AUC).

Method	Mean	Std	Max
Single subject, single session			
ERPCov + MDM	0.689	0.018	0.707
ERPCov + TS	0.704	0.030	0.734
ERPCov + TS + Ridge	0.687	0.033	0.720
Vect + LR	0.645	0.023	0.668
Vect + RegLDA	0.665	0.023	0.688
XdawnCov + MDM	0.661	0.023	0.684
XdawnCov + TS	0.708	0.029	0.737
ERPCov + TS finetuned	0.721	0.029	0.750
XdawnCov + TS finetuned	0.723–0.724	0.027–0.028	0.750–0.752
All subjects, all sessions			
ERPCov + MDM	0.544	0.029	0.573
ERPCov + TS	0.587	0.018	0.605
ERPCov + TS + Ridge	0.584	0.017	0.601
Vect + LR	0.562	0.020	0.582
Vect + RegLDA	0.573	0.018	0.591
XdawnCov + MDM	0.532	0.029	0.561
XdawnCov + TS	0.587	0.017	0.604
ERPCov + TS finetuned	0.590	0.015	0.605
XdawnCov + TS finetuned	0.583–0.589	0.012–0.015	0.597–0.602

Tuned pipelines showed small but consistent gains in both evaluation settings. However, cross-subject generalization remained weak, with average AUCs dropping from 0.72 (single-subject) to around 0.58 (all-subject).

B. P300 Paradigm

Table 2 lists AUC results for the P300 paradigm. This task achieved generally higher within-subject scores compared to N170, with ERPCov + TS and XdawnCov + TS reaching approximately 0.784 AUC on average.

Across all pipelines, results suggest that P300 decoding benefited from the tangent space representations and regularized

TABLE II
P300 CLASSIFICATION PERFORMANCE (AUC).

Method	Mean	Std	Max
Single subject, single session			
ERPCov + MDM	0.777	0.041	0.818
ERPCov + TS	0.784	0.039	0.823
ERPCov + TS + Ridge	0.781	0.041	0.822
Vect + LR	0.668	0.039	0.707
Vect + RegLDA	0.757	0.046	0.803
XdawnCov + MDM	0.762	0.044	0.806
XdawnCov + TS	0.784	0.038	0.822
ERPCov + TS finetuned	0.786	0.040	0.826
XdawnCov + TS finetuned	0.783–0.784	0.039–0.041	0.823–0.825
All subjects, all sessions			
ERPCov + MDM	0.555	0.029	0.584
ERPCov + TS	0.636	0.012	0.648
ERPCov + TS + Ridge	0.636	0.012	0.648
Vect + LR	0.579	0.016	0.595
Vect + RegLDA	0.593	0.017	0.610
XdawnCov + MDM	0.542	0.028	0.570
XdawnCov + TS	0.636	0.012	0.648
ERPCov + TS finetuned	0.634	0.012	0.646
XdawnCov + TS finetuned	0.628–0.633	0.012	0.640–0.645

TABLE III
SSVEP CLASSIFICATION PERFORMANCE (AUC).

Method	Mean	Std	Max
Single subject, single session			
CSP + Cov + TS	0.949	0.030	0.979
CSP + RegLDA	0.941	0.030	0.971
Cov + MDM	0.896	0.048	0.944
Cov + TS	0.952	0.026	0.978
CSP + Cov + TS finetuned	0.948	0.029	0.977
CSP + RegLDA finetuned	0.940	0.034	0.974
Cov + TS finetuned	0.951	0.025	0.976
Cov + TS + SVM	0.910	0.030	0.940
Xdawn + TS + LR	0.935	0.033	0.968
All subjects, all sessions			
CSP + Cov + TS	0.864	0.045	0.909
CSP + RegLDA	0.850	0.042	0.892
Cov + MDM	0.758	0.057	0.815
Cov + TS	0.869	0.030	0.899
CSP + Cov + TS finetuned	0.861	0.047	0.908
CSP + RegLDA finetuned	0.858	0.040	0.898
Cov + TS finetuned	0.867	0.034	0.901
Cov + TS + SVM	0.814	0.040	0.854
Xdawn + TS + LR	0.815	0.048	0.863

classifiers. As with N170, performance declined substantially when models were applied across subjects, confirming high inter-subject variability.

C. SSVEP Paradigm

Table 3 shows results for the SSVEP paradigm. Compared to N170 and P300, SSVEP models demonstrated much stronger generalization. Covariance-based tangent space pipelines achieved mean AUCs above 0.95 for single-subject and around 0.86–0.87 for all-subject conditions.

Overall, the SSVEP models produced the most stable and transferable representations, indicating the robustness of steady-state signals compared to transient ERPs.

D. Deep Learning Models

Deep learning models, including EEGAutoencoder, Denoising Autoencoder, and Sparse Autoencoder variants were evaluated both on reconstruction performance and downstream classification accuracy.

While training accuracies occasionally reached 70–80%, the corresponding test accuracies consistently remained within 0.48–0.52, effectively at chance level. This clear disparity between training and test performance indicates severe overfitting, suggesting that the models were memorizing subject- or session-specific EEG patterns rather than learning generalizable neural representations.

Even after applying multiple forms of regularization, dropout, and data augmentation, the networks showed no improvement in generalization. The limited dataset size, high inter-subject variability, and relatively small number of training trials per subject made deep architectures unsuitable in this context, especially compared to the lightweight, covariance-based pipelines that achieved strong results with much less data.

V. Discussion

The results demonstrate a clear divide between the performance of traditional covariance-based pipelines and the tested deep learning models across all three EEG paradigms. Overall, classical methods provided reliable and interpretable performance, while deep models consistently failed to generalize beyond training data, despite showing promising reconstruction or training accuracy.

A. Traditional Pipelines and Paradigm Differences

Covariance-based tangent space models once again proved to be robust baselines for EEG decoding, particularly under data-scarce conditions. Across paradigms, tangent space representations combined with regularized classifiers (e.g., RegLDA, Ridge) delivered stable results with minimal tuning effort. Incremental improvements observed from finetuned variants suggest that even minor adjustments to preprocessing or classifier parameters can yield measurable performance gains.

The differences between paradigms highlight the varying levels of signal consistency and generalizability. SSVEP, being a steady-state response, showed the highest and most stable AUCs across both single-subject and all-subject conditions, with mean values around 0.95 and 0.86, respectively. This reinforces the notion that frequency-tagged paradigms produce more robust and transferable representations across individuals. In contrast, the transient ERP paradigms, N170 and P300, exhibited significant drops in cross-subject performance (from approximately 0.72–0.78 down to 0.58–0.63), reflecting their higher inter-subject variability and dependence on temporal precision.

B. Limitations of Deep Learning in Low-Data EEG Contexts

Despite incorporating multiple forms of regularization, data augmentation, and architectural variations, all autoencoder-based models failed to achieve meaningful classification performance. While training accuracies occasionally exceeded 70–80%, the corresponding test accuracies remained within the 0.48–0.52 range, essentially random chance. This severe generalization gap highlights a key limitation of deep learning in the small-sample EEG regime: overparameterized networks tend to memorize subject-specific noise rather than capture task-relevant neural structure.

The lack of improvement even with denoising or sparse variants suggests that traditional augmentation and regularization alone are insufficient under such constraints. EEG data's high dimensionality, low signal-to-noise ratio, and strong inter-individual variability all contribute to this failure. Without extensive pretraining or large-scale multi-subject datasets, purely supervised or autoencoding deep models are unlikely to outperform classical pipelines in this domain.

C. Generalization and Future Directions

The generalization gap observed between single-subject and all-subject conditions underscores the challenge of building subject-independent EEG decoders. Traditional covariance-based models still degrade under cross-subject testing, though SSVEP results indicate that paradigms with stable frequency components may offer a more transferable foundation. Future work should explore subject adaptation and transfer learning strategies, such as fine-tuning pretrained models on limited new-subject data, or using contrastive and self-supervised pretraining to capture invariant representations across sessions.

Another promising direction is the integration of hybrid pipelines: combining Riemannian geometry–based feature extraction with shallow neural networks, or leveraging pretrained autoencoders to augment classical classifiers. Such approaches may balance interpretability and representation power without incurring the heavy data demands of deep models.

In summary, these findings reinforce that for small EEG datasets, simpler covariance-based pipelines remain not only competitive but superior to deep architectures. While deep learning continues to dominate other modalities, EEG decoding still benefits most from well-understood, data-efficient feature representations.

VI. Conclusion

This study benchmarked both traditional Riemannian geometry–based pipelines and modern deep learning approaches across three EEG paradigms, N170, P300, and SSVEP, using the EEG-ExPy benchmark. The results clearly demonstrated that classical covariance-based methods continue to outperform deep models in small-sample EEG scenarios. Among these, tangent space representations combined with linear classifiers achieved the most consistent results, particularly for the SSVEP paradigm, where accuracies remained high and stable across subjects and sessions.

In contrast, deep learning models, despite achieving strong training accuracies, failed to generalize, indicating severe overfitting to subject-specific features. These outcomes highlight the practical limitations of current deep architectures in EEG decoding when data availability is limited. The findings reaffirm that simplicity and domain-aligned feature extraction still provide the most reliable decoding performance in such contexts.

Overall, this work underscores the importance of data efficiency, interpretability, and the continued relevance of traditional methods in neurotechnology research. Until larger-scale EEG datasets or more robust pretraining techniques become standard, Riemannian and tangent-space pipelines remain the most dependable choice for EEG-based BCI tasks.

VII. Future Work

While the presented approaches provide strong baselines, there is substantial room for advancing EEG decoding performance. Future research should focus on three key areas:

A. Cross-subject Adaptation and Transfer Learning

The largest performance drops were observed when moving from single-subject to multi-subject conditions. This gap could be mitigated through domain adaptation, fine-tuning pretrained models for new subjects, or implementing transfer learning frameworks that leverage shared feature spaces across individuals.

B. Self-supervised and Contrastive Representation Learning

Rather than relying solely on supervised objectives, future deep models could adopt self-supervised learning (SSL) or contrastive objectives to learn invariant EEG representations without large label requirements. Such pretraining strategies have shown promise in other biosignal domains and may help bridge the generalization gap.

C. Hybrid Model Architectures

Combining the strengths of both paradigms using Riemannian features as structured inputs to shallow neural networks could yield more data-efficient models. Alternatively, pretrained autoencoders may serve as feature extractors to enhance classical classifiers instead of acting as standalone decoders.

Beyond methodological innovation, future work should also emphasize reproducibility, cross-dataset validation, and real-time deployment feasibility. As EEG-BCI research advances, striking the balance between interpretability, robustness, and computational efficiency will remain central to achieving practical and scalable neural decoding systems.

Code Availability

The code implementation and experimental scripts used in this study are available at

https://github.com/aaivu/In21-S7-CS4681-AML-Research-Projects/tree/main/projects/210146N-Neurotechnology_Brain-Computer-Interface.

References

[1] Y. Liu, Z. Quince, S. Goh, S. Teragawa, and T. Low, "A lightweight deep learning model for eeg classification across visual stimuli," University of Southern Queensland preprint, 2023.

[2] J. A. O'Reilly, J. Wehrman *et al.*, "Neural correlates of face perception modeled with a convolutional recurrent neural network," *bioRxiv*, 2023.

[3] G. Zhang *et al.*, "Assessing the effectiveness of spatial pca on svm-based decoding of eeg data," *NeuroImage*, 2023.

[4] D. Borra *et al.*, "Ms-eegnet: A lightweight multi-scale convolutional neural network for p300 decoding," *Journal of Neural Engineering*, 2021.

[5] R. Afrah, Z. Amini, and R. Kafieh, "An unsupervised feature extraction method based on clstm-ae for accurate p300 classification," *Journal of Biomedical Physics & Engineering*, 2024.

[6] P. Du, P. Li *et al.*, "Single-trial p300 classification algorithm based on centralized multi-person data fusion cnn," *Frontiers in Neuroscience*, 2023.

[7] Z. Ermaganbet, A. Mussabayeva *et al.*, "Subject-independent p300 speller classification using time-frequency representation and double input cnn with feature concatenation," in *IEEE DSP Conference*, 2023.

[8] J. Hong, G. Mackellar, and S. Ghane, "Spellerssl: Self-supervised learning with p300 aggregation for speller bcis," arXiv preprint, 2025.

[9] J. Chen, Y. Zhang, P. Peng *et al.*, "A transformer-based deep neural network model for ssvep classification (ssvepformer)," arXiv preprint, 2022.

[10] Y. Dai, Z. Chen *et al.*, "A time-frequency feature fusion-based deep learning network for ssvep frequency recognition (ssvep-tffnet)," *Frontiers in Neuroscience*, 2025.

[11] W. Dong, C. Xu *et al.*, "Enhanced ssvep bionic spelling via xlstm-based deep learning and spatial attention," *Biomimetics*, 2025.

[12] J. A. O'Reilly *et al.*, "Blind source separation of event-related potentials using a recurrent neural network," *bioRxiv*, 2024.

[13] B. Aristimunha, R. Y. de Camargo, W. H. Lopez Pinaya, S. Chevallier, A. Gramfort, and C. Rommel, "Evaluating the structure of cognitive tasks with transfer learning," arXiv preprint, 2023.

[14] R. Kessler, A. Enge, and M. A. Skeide, "How eeg preprocessing shapes decoding performance," *Communications Biology*, 2025.

[15] V. Marochko *et al.*, "Integrated gradients for enhanced interpretation of p3b-erp classifiers trained with eeg-superlets in traditional and virtual environments," in *CEUR Workshop*, 2025.

[16] Y. Ravipati, N. Pouratian *et al.*, "Evaluating deep learning performance for p300 neural signal classification," in *AMIA Annual Symposium Proceedings*, 2024.

[17] S. Zang, X. Ding, M. Wu, and C. Zhou, "An eeg classification-based method for single-trial n170 latency detection and estimation," *Computational and Mathematical Methods in Medicine*, 2022.