Detecting Arousal States from Physiological Signals Using EEG-Based Features

Anirudh Natarajan
Ph.D., Biomedical Engineering
Columbia University
New York, NY, USA
amn2225@columbia.edu

Michael Zhou
M.S., Data Science
Columbia University
New York, NY, USA
mgz2112@columbia.edu

Sophie Yangyi
Ph.D., Biomedical Engineering
Columbia University
New York, NY, USA
yy3389@columbia.edu

Zhiyu Sun
M.S., Biomedical Engineering
Columbia University
New York, NY, USA
zs2710@columbia.edu

***Abstract*—Overview. This paper presents a digital signal processing pipeline for detecting and classifying human arousal states from EEG signals, systematically exploring frequency-band filtering, feature extraction, and classification techniques. Problem Statement. The classic inverse-U-shaped Yerkes-Dodson curve indicates that optimal cognitive performance occurs at an intermediate level of arousal, making accurate arousal detection crucial for applications such as EEG-based neurofeedback. Approach. Drawing on datasets from Faller et. al (2019) [4], we applied a 4th order Butterworth bandpass filter to isolate multiple EEG frequency bands (Delta, Theta, Alpha, Beta, Gamma), extracted statistical features, and then evaluated three classifiers (LDA, SVM, Logistic Regression) using Leave-One-Out Cross-Validation (LOOCV). We also investigated the impact of various downsampling factors (1, 2, 5, 10) to assess how reducing data rates influences classification accuracy. Key Results. For one subject, our pipeline with SVM achieved up to 0.93 LOOCV accuracy using all frequency bands, generally outperforming LDA and Logistic Regression. While multi-band features often surpassed single-band approaches, certain bands (notably Gamma) remained competitive on their own. Crucially, Gamma-band features and the LDA classifier demonstrated remarkable resilience under substantial downsampling, maintaining above-chance accuracy even at a 10x reduction in data. Conclusion. These results highlight that careful frequency-band selection, combined with appropriate classifiers and preprocessing steps such as downsampling, can yield robust stress/arousal detection using only EEG signals. This DSP-driven pipeline can readily incorporate additional physiological signals, more sophisticated filters, and advanced feature extraction methods for improved performance and broader applicability.**

***Index Terms*—EEG-based signal processing, arousal state classification, frequency-band filtering, downsampling, feature extraction, LDA, SVM, Logistic Regression, LOOCV, neurofeedback**

I. Introduction

A. Motivation and Problem Statement

A proper understanding of how human arousal states influence cognitive performance is crucial, especially in high-pressure environments where optimal arousal can determine success or failure. Consider the aviation scenario in which a pilot's heightened arousal can induce catastrophic pilot-induced oscillations (PIO), or similarly, a surgeon performing a life-saving operation who must maintain composure under extreme stress. Even mundane tasks like walking on a pristine carpet with a mug of coffee can become stressful when compared to walking outside on the sidewalk. According to the Yerkes-Dodson Law, both overstimulation and understimulation negatively affect performance, underscoring the need for achieving an optimal, moderate level of arousal. Fortunately, multiple physiological signals can indicate stress (high arousal) versus a more relaxed state (low arousal). Among these signals are brain waves, measured noninvasively via electroencephalography (EEG). Accurate detection using EEG signals can guide interventions like neurofeedback to achieve and maintain optimal arousal levels during demanding tasks, as shown in [4].

B. Related Work

Prior research has leveraged EEG and other physiological signals to measure and modulate arousal, demonstrating that real-time feedback and training can improve performance in challenging tasks. Early studies established that EEG features, particularly power in specific frequency bands, correlate with stress and arousal levels [1][2]. EEG-based classification methods have been employed in both controlled laboratory settings and real-world scenarios [3][4]. Extending beyond classification, researchers have focused on enabling users to self-regulate arousal through neurofeedback. In Faller et al. (2019), for example, EEG-based real-time neurofeedback improved participants' performance in a demanding sensory-motor task. This project aligns with existing literature showing the reliability of using frequency-band features for classifying arousal states.

C. Our Contribution

Building on previous work, we develop a pipeline that integrates EEG frequency-band filtering, feature extraction, and multiple classification methods to robustly detect arousal states in offline conditions. Specifically, we extract features from Delta, Theta, Alpha, Beta, and Gamma bands to classify trials as "easy" versus "hard," using these conditions as proxies for different arousal levels. While the original study supporting this dataset performed real-time classification for neurofeedback, our approach focuses on an offline, proof-of-concept analysis. We evaluate three classifiers—LDA, SVM, and Logistic Regression—and, with the SVM, achieve up to 0.92 LOOCV accuracy for one subject when combining

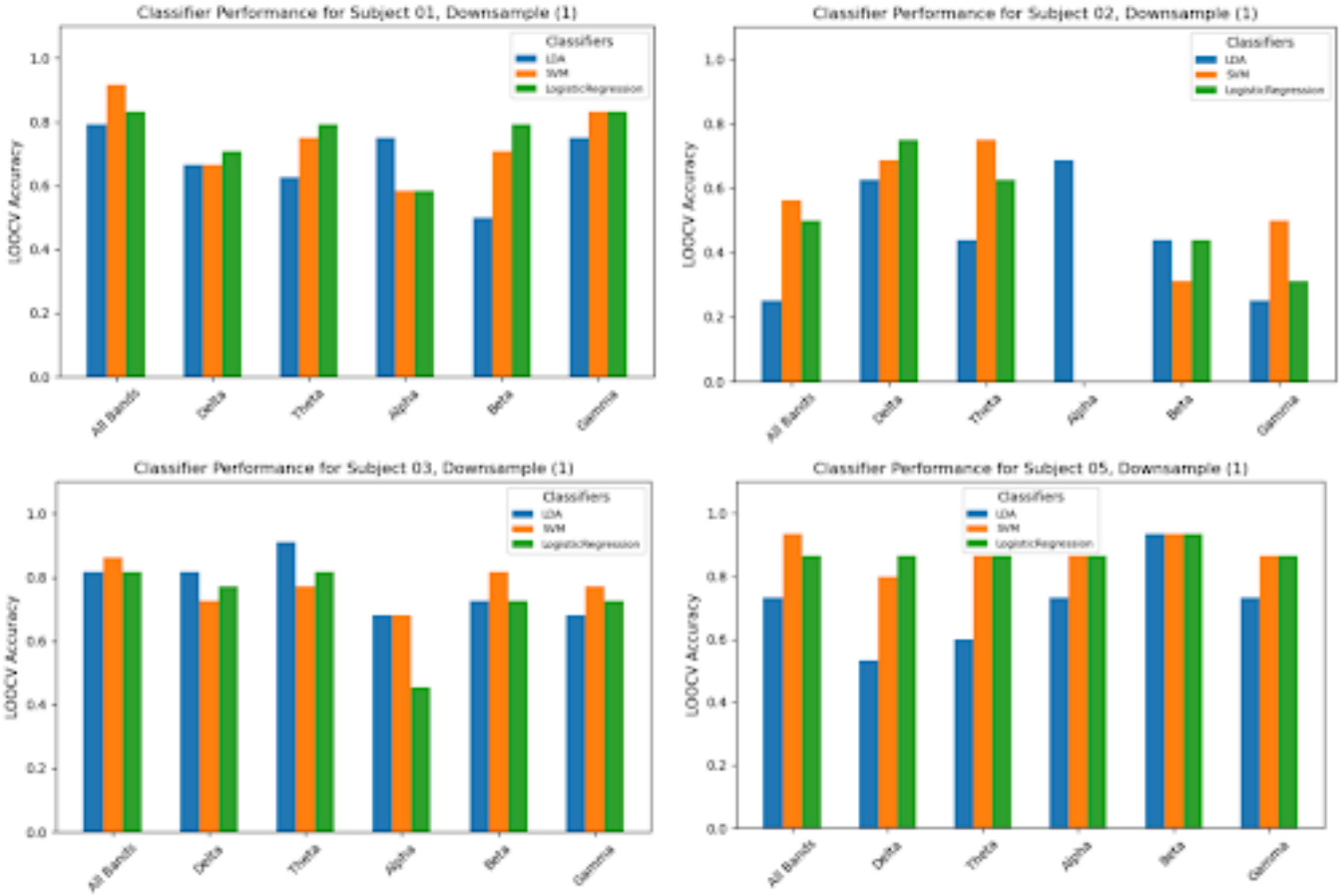

Fig. 1. Leave-One-Out Cross-Validation (LOOCV) Accuracy for individual subjects (S01, S02, S03, and S05) across different frequency bands. High-frequency Gamma and Beta bands showed superior classification accuracy for S01 and S05, while lower-frequency bands (Theta and Delta) generally performed worse. Notable inter-subject variability is observed, with S03 achieving the highest Theta band accuracy (0.91) with LDA and S02 showing poorer overall performance. Combining all bands generally improved classification accuracy by leveraging comprehensive neural activity representation.

all frequency bands. Additionally, we investigate the impact of downsampling on classifier performance. Consistent with existing literature, we find that high-frequency bands, especially Gamma, remain crucial contributors to above-chance classification accuracy even under heavy downsampling. Our pipeline can easily incorporate other physiological signals, such as heart rate variability or pupil size, as additional features or for validation of arousal state.

II. Technical Approach

A. Data and Experimental Setup

Our analysis begins with a publicly available EEG dataset from Faller et al. (2019), which offers a suitable testbed for examining the relationship between arousal states and task difficulty. Participants performed a Boundary Avoidance Task in a visual flight simulation, with EEG signals recorded at 256 Hz from a 64-channel electrode array. While the dataset does not directly label arousal levels, we inferred a proxy by labeling the first 30 seconds of each trial as "easy" (low stress) and the last 30 seconds as "hard" (high stress), based on the increasing difficulty and eventual failure point of each trial. This process yielded 12 easy and 12 hard trials per subject, each of consistent duration.

B. Preprocessing Steps

To efficiently handle the EEG signals and highlight informative features, we employed frequency-band filtering and downsampling. First, we applied a 4th order Butterworth bandpass filter to isolate five canonical frequency bands: Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–12 Hz), Beta (12–25 Hz), and Gamma (25–100 Hz). Next, we explored various downsampling factors (1, 2, 5, 10) to reduce computational load and assess the effects on classification performance. This approach allowed us to determine which frequency bands remained most informative under heavily reduced data rates.

C. Feature Extraction

After filtering the EEG data into the selected frequency bands, we computed statistical features to capture discriminative patterns linked to arousal. For each band and each trial, we computed the mean and standard deviation across all channels, yielding a compact feature vector that summarized each trial's spectral characteristics. We compared single-band feature sets

to multi-band concatenations, finding that multi-band features often improved accuracy, although high-frequency bands like Gamma were sometimes individually competitive. After filtering the EEG data into the selected frequency bands, we computed statistical features to capture discriminative patterns linked to arousal. For each band and each trial, we computed the mean and standard deviation across all channels, yielding a compact feature vector that summarized each trial's spectral characteristics. We compared single-band feature sets to multi-band concatenations, finding that multi-band features often improved accuracy, although high-frequency bands like Gamma were sometimes individually competitive.

D. Classification Methods

With feature vectors ready, we applied three traditional machine learning classifiers—Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and Logistic Regression—to distinguish between easy (low arousal) and hard (high arousal) trials. We selected these algorithms because they are well-established in EEG-based classification, offering interpretability, speed, and a useful baseline. In our experiments, SVM generally outperformed LDA and Logistic Regression, but the latter two methods still served as valuable benchmarks and helped validate the feature selection process.

E. Cross-Validation Strategy

To obtain robust performance estimates with limited trial counts, we employed Leave-One-Out Cross-Validation (LOOCV). By training on all but one trial and then testing on the remaining trial, LOOCV maximizes data usage. This approach, well-suited for relatively small datasets, provided a reliable measure of how effectively each classifier differentiated easy (low-stress) from hard (high-stress) EEG epochs, using classification accuracy as the key performance metric.

III. Experiments

A. Experimental Protocol

In our experiments, we conducted within-subject classification of easy (low-stress) versus hard (high-stress) trials across multiple participants, both individually and in aggregate, to assess the robustness of our pipeline. We chose a within-subject approach due to the likely personalized application of such methods, making generalizability less critical. For a detailed single-subject analysis, we selected four out of twenty subjects (S01, S02, S03, S05). We then expanded our evaluation by averaging performance across all available subjects, providing a broader view of the pipeline's overall effectiveness.

B. Results Across Frequency Bands

Examining individual frequency bands revealed variable predictive power, yet combining multiple bands often yielded superior accuracy. As shown in Fig. 1, there was significant inter-subject variability. For S01 and S05, classification using high-frequency Gamma and Beta bands reached up to 0.83 and 0.93 accuracy, respectively, while lower-frequency bands (Theta and Delta) performed somewhat lower, consistent with the literature. However, S03 attained up to 0.91 accuracy with LDA on the Theta band, and S02 exhibited relatively poor performance overall. Combining all bands typically improved accuracy beyond the average of individual bands, as it captured a more complete representation of neural activity. In Fig. 2, the aggregated results across all subjects also show more robust performance for the all-bands condition, though the anticipated superior performance of high-frequency bands alone was less pronounced at the group level.

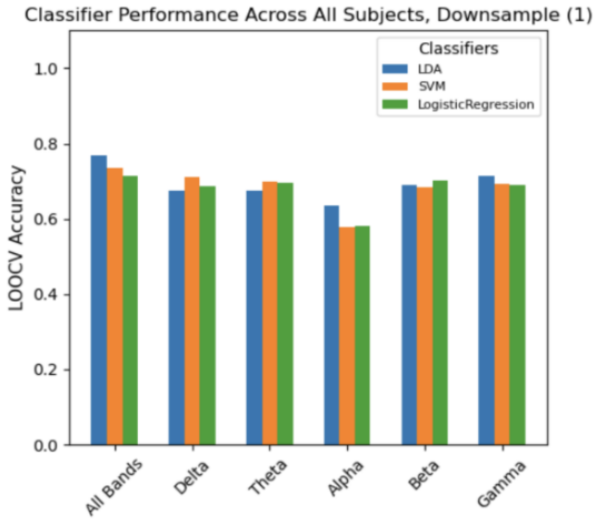

Fig. 2. Group-level LOOCV Accuracy averaged across all subjects, comparing individual frequency bands and the all-bands condition. The all-bands condition demonstrated the most robust classification performance, capturing information across multiple neural activity bands. However, the expected dominance of high-frequency bands (Gamma and Beta) was less evident at the group level.

C. Impact of Preprocessing Parameters

Altering preprocessing steps, particularly downsampling, revealed that while accuracy generally declined at higher downsampling factors, certain frequency bands—especially Gamma—remained discriminative. Fig. 3 illustrates the effects of downsampling by factors of 1, 2, 5, and 10 across all subjects. Heavy downsampling drastically reduced accuracy for single-band conditions and led to a modest decrease for the all-bands combination. Notably, Gamma band features demonstrated exceptional resilience, consistently performing above chance and remaining comparable to all-band performance even under a 10x downsampling condition.

D. Comparison of Classifiers

Across subjects and conditions, the SVM often emerged as the top-performing classifier, although LDA and Logistic Regression served as valuable baselines. While SVM did not always produce the absolute highest accuracy, its performance remained stable and less prone to unexpected fluctuations across different subjects and conditions. Interestingly, LDA

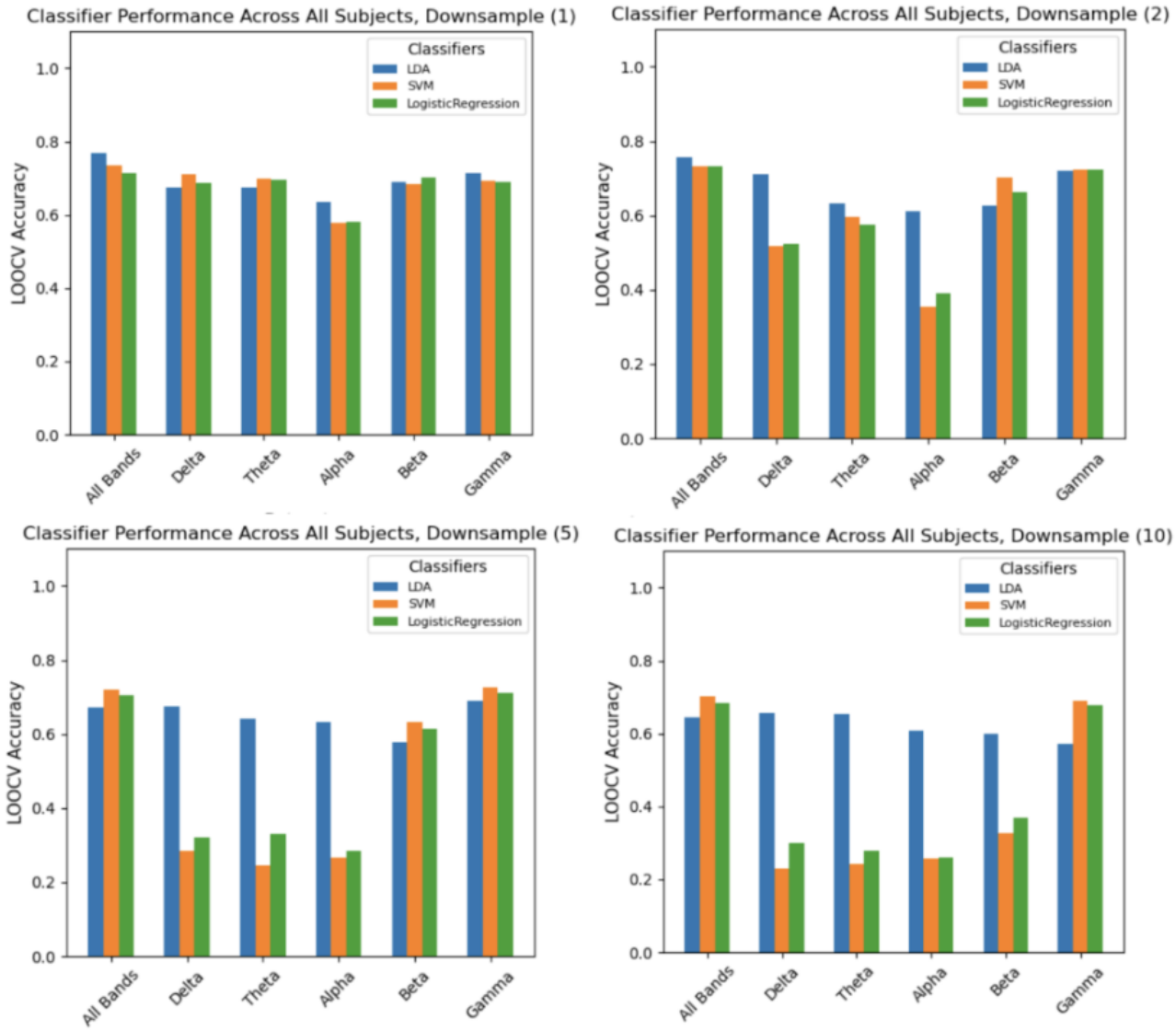

Fig. 3. LOOCV Accuracy across all subjects for various downsampling factors (1, 2, 5, and 10), comparing individual frequency bands and the all-bands condition. Higher downsampling factors led to a general decline in accuracy, particularly for single-band conditions. However, the Gamma band demonstrated remarkable robustness, maintaining discriminative power and achieving performance levels comparable to the all-bands condition, even under a 10x downsampling factor. The all-bands condition showed reduced sensitivity to downsampling, highlighting its ability to integrate features across bands to mitigate information loss.

maintained relatively high accuracy despite downsampling, exhibiting less degradation compared to the other classifiers—a point we will explore further in the subsequent discussion section.

E. Testing and Verification

To ensure robustness in EEG signal preprocessing, we evaluated the classification performance across five commonly used filter types: Butterworth, Chebyshev I, Chebyshev II, Elliptic, and Bessel. Fig. 4 presents the LOOCV accuracy averaged across all subjects for these filter types at a downsampling rate of 1. The results were compared across frequency bands (Delta, Theta, Alpha, Beta, Gamma) and the combined "All Bands" condition for three classifiers (LDA, SVM, Logistic Regression). The Butterworth filter demonstrated consistent and high performance across all frequency bands and classifiers, making it the most reliable choice for this study. Both Chebyshev types I and II filters showed slight variability, particularly in mid-frequency bands (Theta and Alpha), likely due to their sharper roll-off characteristics which may affect signal continuity. While maintaining high accuracy overall, the Elliptic filter exhibited sensitivity to specific bands, particularly in lower frequencies, which could lead to inconsistencies in performance. Known for its maximally flat phase response, the Bessel filter performed competitively in some bands but showed reduced robustness compared to the

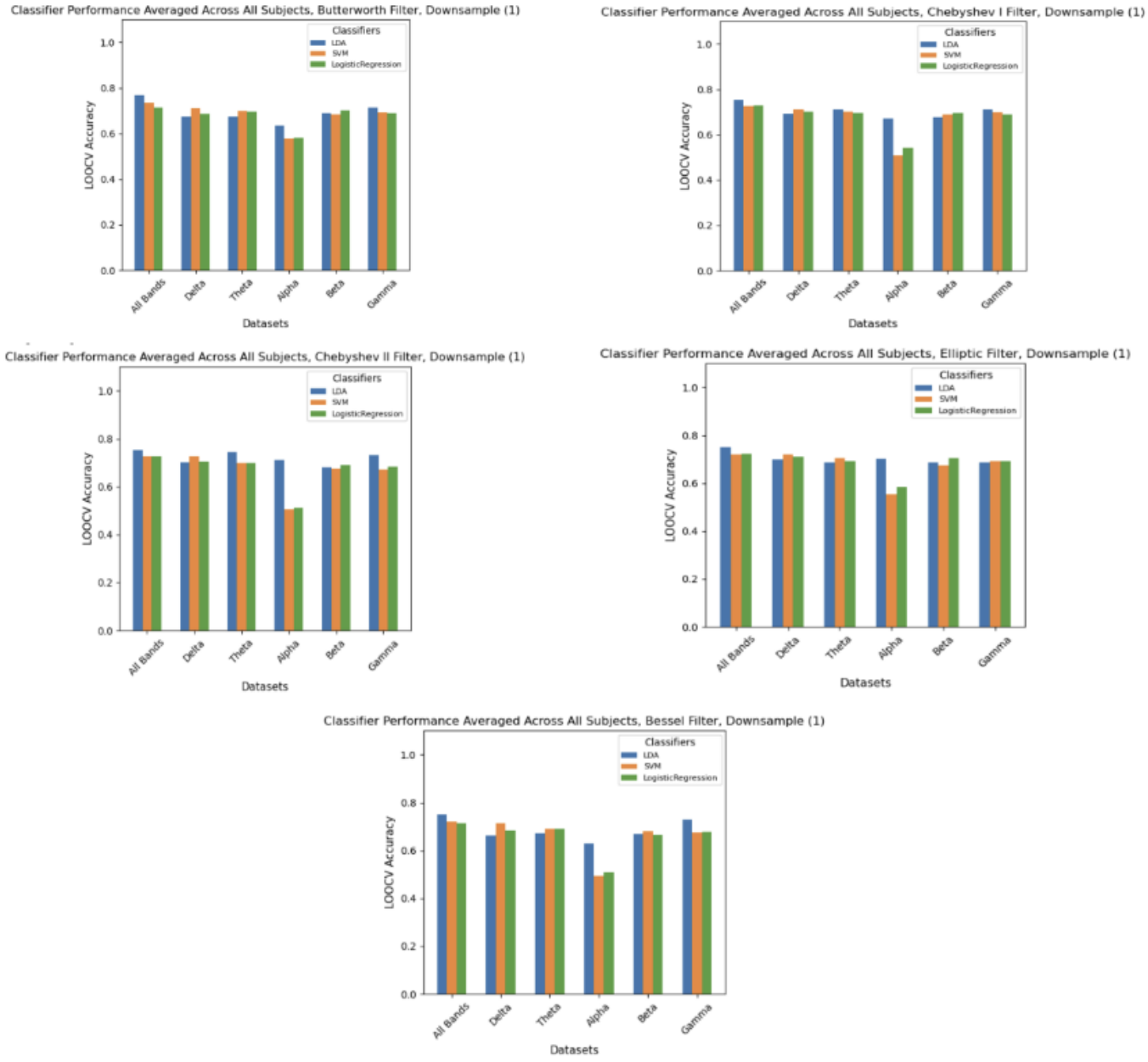

Fig. 4. Comparison of LOOCV accuracy averaged across all subjects using different filter types (Butterworth, Chebyshev I, Chebyshev II, Elliptic, and Bessel) at a downsampling rate of 1. The Butterworth filter exhibited consistent and robust performance across all frequency bands and classifiers. Chebyshev filters (I and II) showed variability, particularly in mid-frequency bands such as Theta and Alpha. The Elliptic filter maintained high accuracy but displayed sensitivity to downsampling. The Bessel filter, known for its maximally flat phase response, showed competitive performance in certain bands but was less robust overall compared to the Butterworth filter. These results underscore the Butterworth filter's suitability for EEG signal preprocessing in this study.

Butterworth filter.

The experimental results reaffirm the suitability of the Butterworth filter for EEG preprocessing in this study. Its balance between amplitude response and phase continuity ensures minimal distortion while preserving critical neural signal characteristics. As downsampling rate increases, the variability between different frequency bands also increases. This analysis highlights the importance of filter selection in EEG-based classification tasks.

IV. Discussion

A. Interpretation of Results

Our findings suggest that certain EEG frequency bands, particularly higher-frequency bands like Gamma and Beta, provide strong discriminative features for classifying arousal states. Moreover, integrating multiple bands generally enhances overall accuracy. While representative subjects, such as S01 and S05, achieved high accuracy (up to 0.83–0.93) using Gamma and Beta bands, we observed that when averaging across all subjects, the advantage of high-frequency bands was less pronounced. Nonetheless, the robustness of the Gamma band stood out under heavy downsampling conditions, demonstrating resilience even at a 10x reduction in data. This aligns with existing literature that links higher-frequency EEG activity to stress and arousal.

Although classification accuracy per se was not the primary focus of this project, the differences in classifier performance are instructive. SVMs, known for handling high-dimensional, nonlinear datasets effectively, delivered stable and generally strong results. Notably, LDA showed remarkable robustness under severe downsampling. Its emphasis on linear separability and reliance on global statistical properties—along with built-in dimensionality reduction—likely renders it less sensitive to reduced data density. In sum, high-frequency bands and LDA-based classification emerged as particularly resilient under stringent preprocessing constraints.

In the filter comparison experiments (see Section III-E), the Butterworth filter demonstrated superior performance across all classifiers and frequency bands, validating its choice for EEG signal preprocessing in this study.

B. Domain of Applicability

This pipeline appears best suited for scenarios with relatively high-quality, high-sampling-rate EEG recordings and stable trial counts. Exclusive reliance on EEG signals, however, may limit the generalizability of our approach. Incorporating additional physiological measures such as heart rate variability or pupil size could strengthen the validity of arousal classification, moving beyond easy/hard trial labels toward more direct indicators of stress. Such multimodal integration would likely yield a more comprehensive understanding of arousal and improve classification performance across diverse populations and conditions.

C. Future Directions

Looking ahead, there are several avenues for improving and extending this work. First, exploring alternative filter designs—such as Chebyshev, Elliptic, or Bessel filters—could refine frequency-band isolation and potentially enhance feature quality. Additionally, advanced signal processing methods, including Independent Component Analysis (ICA), could help remove artifacts and improve EEG signal purity. Beyond preprocessing, feature extraction techniques like power spectral density (PSD), nonlinear dynamics measures, or deep learning-based representations may capture subtle arousal-related patterns more effectively than simple statistical descriptors.

Finally, as was done in Faller et. al (2019), implementing this pipeline in a closed-loop, real-time system—such as a neurofeedback environment—would allow for on-the-fly arousal regulation, offering practical benefits in demanding operational tasks and high-stress professional settings. Such developments hold the promise of moving from offline analysis toward adaptive, responsive interventions that optimize human performance and well-being.

Acknowledgment

We would like to thank Professor John Wright for putting together such a wonderful course in Digital Signal Processing, as well as the Teaching Assistants for their support along the way.

References

[1] J. F. Alonso, S. Romero, M. R. Ballester, R. M. Antonijoan, and M. A. Mañanas, "Stress assessment based on EEG univariate features and functional connectivity measures," Physiological Measurement, vol. 36, no. 7, pp. 1351–1365, May 2015, doi: https://doi.org/10.1088/0967-3334/36/7/1351.

[2] A. Asif, M. Majid, and S. M. Anwar, "Human stress classification using EEG signals in response to music tracks," Computers in Biology and Medicine, vol. 107, pp. 182–196, Apr. 2019, doi: https://doi.org/10.1016/j.compbiomed.2019.02.015.

[3] M. Bagheri and S. D. Power, "EEG-based detection of mental workload level and stress: the effect of variation in each state on classification of the other," Journal of Neural Engineering, vol. 17, no. 5, p. 056015, Oct. 2020, doi: https://doi.org/10.1088/1741-2552/abbc27.

[4] J. Faller, J. Cummings, S. Saproo, and P. Sajda, "Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task," Proceedings of the National Academy of Sciences, vol. 116, no. 13, pp. 6482–6490, Mar. 2019, doi: https://doi.org/10.1073/pnas.1817207116.