

Benchmarking Feature-Level and Decision-Level Fusion Strategies for Hybrid EEG-fNIRS Cognitive State Detection

Ainara Carpio-Chicote^{1,2,3}, Tudor-Andrei Deaconu⁵, Enrique J. Gómez^{1,2,3,4}, George Kontaxakis⁵

¹Biomedical Engineering and Telemedicine Centre, ETSI Telecomunicación

²Center for Biomedical Technology, Universidad Politécnica de Madrid

³imas12 Hospital 12 de Octubre Madrid, Spain

⁴CIBER-BBN, Madrid, Spain

⁵Information Processing and Telecommunications Center, ETSI Telecomunicación, Universidad Politécnica de Madrid

Email: ainara.carpio.chicote@fgupm.upm.es

ABSTRACT: The integration of Electroencephalography (EEG) and functional Near-Infrared Spectroscopy (fNIRS) has emerged as a robust framework for passive Brain-Computer Interface (pBCI) applications, particularly in the field of cognitive workload monitoring. This study benchmarks feature-level (early) and decision-level (late) fusion strategies using a subject-independent Leave-One-Subject-Out (LOSO) validation framework. Leveraging an open-access dataset of 26 participants performing an n-back task (0-back vs. 2/3-back), we developed a MATLAB-Python pipeline to contrast unimodal and multimodal performance. Our results demonstrate that while EEG provides high temporal responsiveness, its inter-subject variability limits generalization (65.9% LOSO accuracy). In contrast, fNIRS offers stable hemodynamic markers (76.9% accuracy). Multimodal integration significantly outperformed unimodal baselines, with decision-level fusion achieving the highest subject-independent performance (85.3% accuracy, 0.977 AUC) using an SVM-Poly meta-classifier. These findings demonstrate that hybrid EEG-fNIRS integration provides a more reliable framework than unimodal approaches for detecting distinct cognitive states. While further validation is needed for granular workload levels, this study establishes a robust baseline for the development of future pBCI applications in operational environments. **KEYWORDS:** Hybrid EEG-fNIRS, Multimodal Fusion, Cognitive Workload.

INTRODUCTION

Monitoring cognitive states in high-risk operational scenarios is a fundamental challenge for the development of reliable Brain-Computer Interfaces (BCIs) [1] [2]. In these environments, the saturation of working memory (WM) resources can lead to critical performance degradation [1]. While traditional BCIs focus on active control, passive BCIs (pBCIs) aim to monitor mental states—such as cognitive workload—without requiring conscious effort from the user [3]. However, the deployment of pBCIs

in real-world settings is often hindered by the limitations of unimodal recordings. Electroencephalography (EEG), while providing excellent temporal resolution, is highly susceptible to artifacts, whereas functional near-infrared spectroscopy (fNIRS) offers hemodynamic stability but suffers from an inherent metabolic delay [4], [5], [6].

Hybrid EEG-fNIRS systems have been proposed to overcome these modality-specific weaknesses by leveraging neurovascular coupling (NVC) [7]. Although the potential of hybrid systems is well-documented, the optimal architectural strategy for integrating these different data streams, specifically regarding feature-level versus decision-level fusion, remains a key area of investigation in neural engineering [8]. This paper establishes a performance benchmark for multimodal fusion strategies by evaluating their ability to detect working memory engagement using the n-back test [9]. By contrasting minimal (0-back) versus high (2/3-back) cognitive demand [10], the aim is to identify the most robust integration framework, providing a baseline for the transition of pBCI technologies from controlled laboratory settings to demanding operational applications.

MATERIALS AND METHODS

Dataset and Experimental Paradigm:

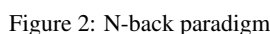
We utilized the open-access hybrid EEG-fNIRS dataset provided by Shin et al. (2018) [11]. The data consists of simultaneous recordings from 26 healthy participants performing a mental workload task based on the n-back paradigm.

EEG activity was recorded at 1,000 Hz using 30 active electrodes (BrainAmp) following the 10-5 system. Simultaneously, fNIRS data were acquired at 10.4 Hz (NIRScout) through 36 channels formed by 16 sources and 16 detectors. Sensors were strategically distributed across frontal, motor, parietal, and occipital regions, with a high-density focus on the prefrontal cortex (16 fNIRS channels) to monitor executive demand (see Fig. 1).

The n-back paradigm is widely used to assess working



- 0-back (Control): Minimal WM demand, used as the baseline.
- 2-back and 3-back (High Load): Increased demand for maintenance, updating, and inhibition of information.



Preprocessing Pipeline:

EEG Preprocessing:

The raw EEG signal (1,000 Hz) was band-pass filtered between 0.5 and 40 Hz using a zero-phase Butterworth filter. Then, artifact rejection was performed via Independent Component Analysis (ICA) using the FastICA algorithm [12]. To ensure objectivity, artifactual components were automatically identified and removed using the Multiple Artifact Rejection Algorithm (MARA) [13], supplemented by statistical thresholds (variance/kurtosis > 3 SD). Finally, the cleaned signal was re-referenced to a Common Average Reference (CAR) [14]. The summarized pipeline can be seen in Fig. 3.



fNIRS Preprocessing:

The fNIRS raw light intensities had already been transformed into relative concentration changes of oxygenated (ΔHbO) and deoxygenated (ΔHbR) hemoglobin via the Modified Beer-Lambert Law (MBLL) and downsampled to 10 Hz.

To eliminate cardiac and respiratory artifacts, a 4th-order zero-phase IIR low-pass filter with a 0.2 Hz cutoff was applied. Each block-level epoch was baseline-corrected using a 3-second window (from -5 s to -2 s prior to block onset) to remove slow physiological drifts and isolate task-related hemodynamic responses. The summarized fNIRS preprocessing pipeline can be seen in Fig. 4.



Segmentation and Multi-level Alignment:

Due to the differences in temporal resolutions, a block-aggregation strategy was adopted. For EEG, spectral features (PSD) and statistical metrics were first computed for each trial-locked epoch (-0.1 to 1.0 s). These features were then averaged across all trials within a 40 s block. This approach ensures that high-frequency spectral information (e.g., beta and gamma bands) is preserved before aggregation, providing a synchronized feature set that matches the 40 s fNIRS block duration.

Feature Extraction: Following block-level aggregation, a multi-domain feature set was extracted to capture the complex dynamics of cognitive load. For each task block, a unified feature vector was created.

- EEG Features: Three domains were calculated for each channel:
 - Spectral: Power Spectral Density (PSD) in Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz) and Gamma (30–40 Hz) bands using Welch’s method.
 - Temporal: P300 component amplitude and latency within a 300–600 ms post-stimulus window.
 - Statistical: Mean, variance, skewness, and kurtosis.
- fNIRS Features: For each channel and chromophore (ΔHbO and ΔHbR), four metrics were extracted: mean concentration, standard deviation, Area Under the Curve (AUC), and the hemodynamic slope estimated via first-order polynomial fitting.

Neurovascular Coupling (NVC) Analysis:

To ensure that the classification accuracy comes from legitimate physiological changes rather than statistical artifacts, a Neurovascular Coupling (NVC) analysis was performed as a biological validation step [15].

- Correlation Analysis: Spearman’s correlation (r_s) was calculated between EEG Alpha power (a marker of cortical inhibition) and HbO concentration in the prefrontal cortex. The aim was to identify the classic inverse relationship: a decrease in Alpha power (event-related desynchronization) correlated with an increase in the hemodynamic response during high-load (2/3-back) conditions.
- Granger Causality: A Granger Causality test was applied to verify that changes in neuronal electrical activity temporally precede fluctuations in hemoglobin metabolism with a physiological delay of 2–5 seconds, validating the neurovascular activation chain.

To account for the metabolic delay of the fNIRS signal, the correlation analyses considered a physiological lag of 2–5 seconds. The use of 40 s blocks further stabilizes the comparison, focusing on sustained hemodynamic trends rather than instantaneous changes.

Classification Framework and Fusion Strategies:

The classification pipeline was implemented in Python using scikit-learn and XGBoost. To ensure a rigorous comparison between unimodal and multimodal approaches, a consistent evaluation protocol was performed in all experiments.

Unimodal Baselines:

Two independent pipelines were developed for EEG and fNIRS [16].

To avoid overfitting due to high-dimensional data, a supervised feature selection using SelectKBest with an ANOVA F-test was implemented. This allowed identifying the most discriminative features (those with the highest variance between cognitive load levels). To reduce dimensionality and mitigate overfitting, EEG features were

selected using an ANOVA F-test, retaining features with a significance threshold of $p < 0.05$.

To ensure robustness, a grid search for hyperparameter tuning was performed, evaluating a diverse set of classifiers. The specific configurations for the best performing models are summarized in Table 1.

Table 1: Classification Models and Hyperparameter Configurations

Model	Core Parameters	Details/Kernel
LDA	Solver: SVD	No shrinkage.
RF	Estimators: 100	Gini; Max depth: None.
XGBoost	LR: 0.1	Max depth: 6; Sub. 0.8.
SVM-RBF	$C \in [0.1, 10]$, γ	RBF kernel.
SVM-Poly	Deg: 3; $C = 1.0$	Poly (Meta-classifier).

Feature-Level Fusion (Early Fusion):

EEG and fNIRS feature vectors were concatenated into a high-dimensional joint representation [17]. To prevent overfitting and exploit cross-modal interactions, we applied dimensionality reduction and hyperparameter tuning within the training folds.

Decision-Level Fusion (Late Fusion):

This modular approach combined the probability outputs of the best-performing unimodal classifiers. These probabilities were used as inputs for a meta-classifier (e.g., SVM-Poly), which generated the final prediction [18]. This strategy preserves the unique characteristics of each signal while leveraging their complementary predictive power.

Validation Protocol:

To ensure a comprehensive evaluation, we employed a dual-validation scheme.

First, a stratified block-level hold-out split (80/20) was used to establish a performance baseline for the hybrid features. To stabilize the high-dimensional feature space, a global scaling transformation (Z-score) and supervised feature selection (SelectKBest, $p < 0.05$) were applied based on the training pool distribution.

Second, to assess subject-independent generalization, a LOSO cross-validation was performed. In this scheme, the model was iteratively trained on $N - 1$ subjects and tested on the remaining participant. This protocol evaluates the BCI’s ability to handle inter-subject variability, which is the primary challenge in real-world deployment. The overall classification and fusion pipeline is illustrated in Fig. 5.

RESULTS

Unimodal Performance and Cross-Subject Robustness:

The first stage of our evaluation established the baseline performance of EEG and fNIRS in a subject-independent scenario. To ensure the results represent the system’s ability to generalize to new users, we prioritized Leave-One-Subject-Out (LOSO) cross-validation over simple hold-out splits.

EEG-based Classification:

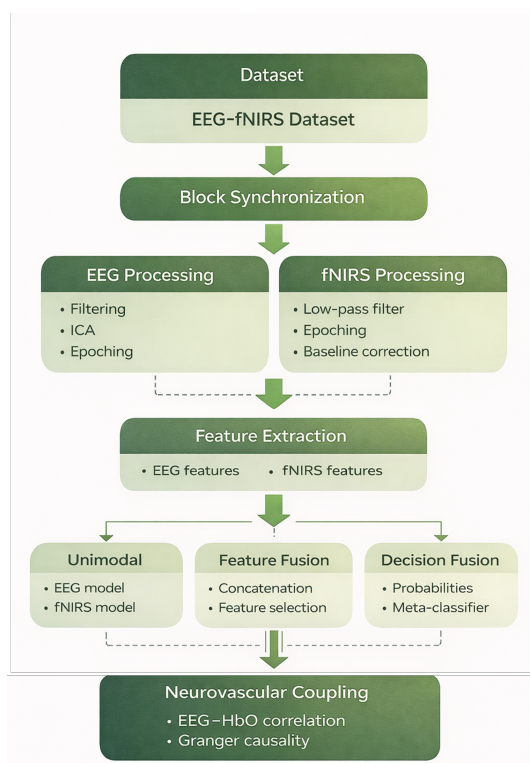


Figure 5: Classification framework

The RF classifier achieved a peak accuracy of 87.2% on the hold-out set. However, performance significantly decreased under LOSO validation to a mean accuracy of 65.95% ($\pm 20.6\%$) and an AUC of 0.709. This drop, combined with the high standard deviation, reflects the high inter-subject variability of electrophysiological signals. While EEG captures fast cognitive changes, its patterns vary significantly between individuals.

fNIRS-based Classification:

Hemodynamic features, evaluated via SVM-RBF, showed a lower peak accuracy on the hold-out set (82.9%) but demonstrated better stability across subjects. The LOSO mean accuracy remained at 76.92% ($\pm 12.8\%$) with an AUC of 0.805. These results suggest that prefrontal hemodynamic responses—specifically oxygenation trends—are more consistent markers of cognitive workload across different participants than EEG spectral densities.

Table 2: Performance Benchmark: Unimodal vs. Multimodal Fusion Strategies. Results reported as Mean (%) $\pm$ SD for Acc. and Mean $\pm$ SD for AUC.

Strategy	Hold-out Acc (%) $\pm$ SD	LOSO Acc (%) $\pm$ SD	LOSO AUC $\pm$ SD
<i>Unimodal</i>			
EEG	87.2 $\pm$ 5.4	66.0 $\pm$ 20.6	0.71 $\pm$ 0.25
fNIRS	82.9 $\pm$ 4.2	76.9 $\pm$ 12.8	0.81 $\pm$ 0.18
<i>Multimodal</i>			
Feature Fusion	90.1 $\pm$ 3.1	80.1 $\pm$ 11.7	0.86 $\pm$ 0.16
Decision Fusion	89.7 $\pm$ 2.8	85.3 $\pm$ 19.1	0.98 $\pm$ 0.05

Multimodal Feature-Level Fusion (Early Fusion):

This strategy integrates EEG and fNIRS information at

Unimodal LOSO Performance: Accuracy Distribution (N=26)

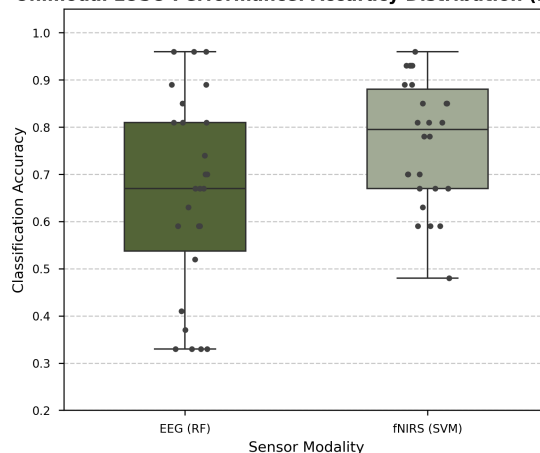


Figure 6: Comparison of unimodal classification accuracy and variability under LOSO cross-validation.

the input stage by concatenating their feature vectors into a single, high-dimensional joint representation. This approach allows the models to learn cross-modal dependencies directly from the synchronized neurophysiological data.

As shown in Table 3, the XGBoost classifier achieved a peak accuracy of 90.1% on the hold-out set. However, the subject-independent LOSO validation, which represents a more realistic scenario for new users, obtained an accuracy of 80.1%.

Table 3: Feature-Level Fusion Performance (Early Fusion). Metrics reported as Mean (%) $\pm$ SD for Acc. and Mean $\pm$ SD for AUC.

Model	Hold-out Acc (%) $\pm$ SD	LOSO Acc (%) $\pm$ SD	LOSO AUC $\pm$ SD
XGBoost	90.1 $\pm$ 3.1	80.1 $\pm$ 11.7	0.86 $\pm$ 0.16
Random Forest	84.4 $\pm$ 4.5	75.3 $\pm$ 14.1	0.81 $\pm$ 0.19
SVM-RBF	85.8 $\pm$ 3.8	74.2 $\pm$ 15.5	0.80 $\pm$ 0.21

The feature importance analysis (Fig. 7) confirms that both EEG and fNIRS contribute significantly to the classification. Frontal Alpha power and prefrontal HbO emerged as the most relevant features. This is consistent with the role of the prefrontal cortex in working memory tasks.

Fig. 7 illustrates the relative contribution (Gain) of the top features in the hybrid model. The decision-making process is clearly driven by a balanced integration of both modalities: while Frontal Alpha power (EEG) serves as a primary neural trigger, prefrontal HbO metrics (fNIRS) provide metabolic stability to the classification. This confirms the biological relevance of our features, as the most discriminative variables are located over the prefrontal cortex, the main hub for working memory processing.

Decision-Level Fusion (Late Fusion):

In this strategy, we combined the posterior probabilities from the best unimodal classifiers (EEG-RF and fNIRS-SVM). As shown in Table 4, the SVM with a Polynomial kernel (degree 3) emerged as the most robust meta-

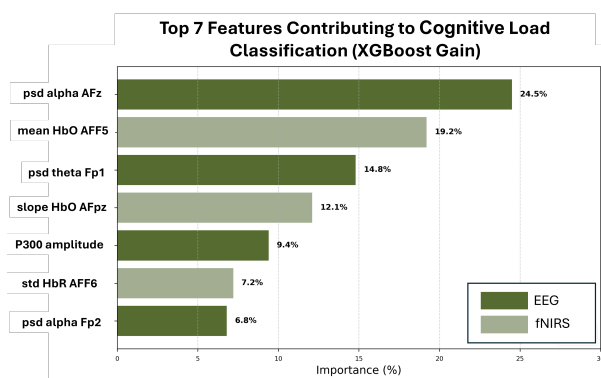


Figure 7: Feature importance ranking (XGBoost Gain) for the early fusion model.

classifier. Under LOSO validation, it achieved a mean accuracy of 85.3% ($\pm 19.1\%$) and an AUC of 0.977 (± 0.05). This high AUC value demonstrates a strong ability to separate workload states across the entire cohort as illustrated by the ROC curve in Fig. 8.

Table 4: Decision-Level Fusion Performance. Comparison of different meta-classifiers. Metrics reported as Mean (%) $\pm$ SD for Acc. and Mean $\pm$ SD for AUC.

Meta-Classifier	Hold-out Acc (%) $\pm$ SD	LOSO Acc (%) $\pm$ SD	LOSO AUC $\pm$ SD
SVM-Poly (d=3)	89.7 $\pm$ 2.8	85.3 $\pm$ 19.1	0.98 $\pm$ 0.05
XGBoost	88.5 $\pm$ 3.5	81.2 $\pm$ 14.5	0.91 $\pm$ 0.12
Random Forest	86.4 $\pm$ 4.1	79.8 $\pm$ 16.2	0.88 $\pm$ 0.15

A key finding is the balanced contribution of both modalities. As detailed in Table 5 and Fig. 9, the SVM-Poly meta-classifier assigns nearly equal weights to each signal (48.1% for EEG and 51.9% for fNIRS). This improvement happens because the model combines the strengths of both signals: the fast response of the EEG and the stability of the fNIRS. By integrating both electrical and hemodynamic data, the model can compensate for the weaknesses of each modality, leading to a more reliable prediction than unimodal methods.

Table 5: Relative contribution of EEG and fNIRS probabilities to the final decision-level fusion across different meta-classifiers.

Meta-Classifier	EEG Influence (%)	fNIRS Influence (%)
SVM-Poly (d=3)	48.1	51.9
XGBoost	51.8	48.2
Random Forest	55.2	44.8
LDA	45.5	54.5

DISCUSSION

The unimodal results highlight a clear difference between the two modalities. EEG shows a high potential to detect changes but fails to generalize well to new users (approximately 66% accuracy). This is likely due to the natural differences in brain shape and electrical patterns between individuals. In contrast, fNIRS acts as a more sta-

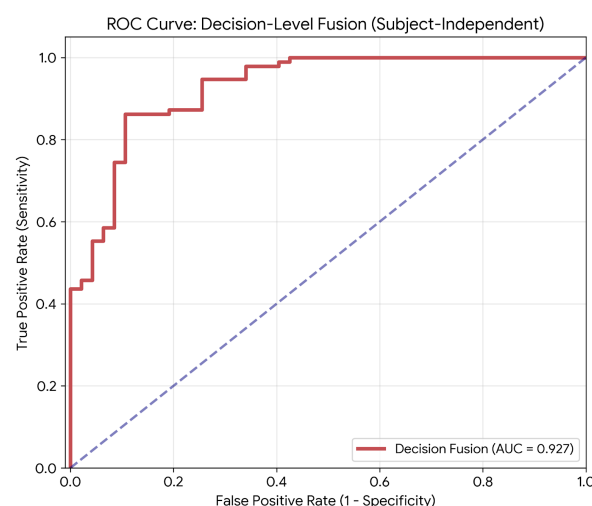


Figure 8: ROC curve analysis for the top-performing Decision-Level Fusion model.

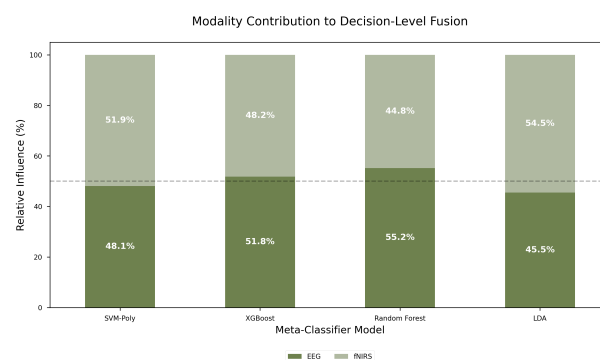


Figure 9: Relative influence of EEG and fNIRS probabilities across different meta-classifiers.

ble marker of workload across the group (77% accuracy), even though it is a slower signal.

The results of the Early Fusion strategy (90.1% peak accuracy) confirm that EEG and fNIRS work better together by capturing complementary aspects of working memory. The XGBoost model effectively identifies moments where both signals align, such as when a drop in Alpha waves (EEG) occurs at the same time as an increase in oxygenation (fNIRS). However, the 10% drop in LOSO accuracy shows that combining too many variables can make the model too specific to the training group. Regarding Late Fusion, the high AUC (0.977) should be viewed carefully. This result is partly due to our test comparing two very different states (very low effort vs. high effort). While this shows the system is very effective at detecting clear workload changes, its main strength is how the stability of fNIRS helps correct the variability of the EEG. This makes the hybrid approach more reliable for potential pBCI applications. Although the use of ICA and MARA is mainly an offline approach, it allows us to establish a clear performance benchmark for fusion. For future real-time applications, these steps could be replaced by faster online artifact rejection methods without changing the core fusion logic.

This dual-source confirmation is essential for transitioning passive BCI (pBCI) systems from controlled laboratory settings to real-world, safety-critical applications.

CONCLUSIONS

In conclusion, our results show that hybrid EEG-fNIRS fusion helps to mitigate the high inter-subject variability of EEG by using fNIRS as a stable reference. Our findings establish that:

1. **Multimodal Reliability:** Hybrid architectures consistently outperform unimodal baselines by using fNIRS as a stabilizing anchor for variable EEG signals.
2. **Generalizability:** Achieving 85% accuracy in subject-independent (LOSO) scenarios confirms that the system captures reliable biomarkers of cognitive load that work across different users.
3. **Future Work:** While the system is highly effective at detecting clear workload transitions, future research should evaluate its performance in more granular states to further define its operational limits.

BIBLIOGRAPHY

- [1] Aricò P, Borghini G, Di Flumeri G, Sciaraffa N, Colosimo A, Babiloni F. Passive BCI in operational environments: Insights, recent advances, and future trends. *IEEE Transactions on Biomedical Engineering*. 2017;64(7):1431–1436.
- [2] Fazli S et al. Enhanced performance by a hybrid NIRS–EEG brain computer interface. *NeuroImage*. 2012;59(1):519–529.
- [3] Aricò P, Borghini G, Di Flumeri G, Sciaraffa N, Babiloni F. Passive BCI beyond the lab: Current trends and future directions. *Physiological Measurement*. 2018;39(8):08TR02.
- [4] Chiarelli AM, Zappasodi F, Di Pompeo F, Merla A. Simultaneous functional near-infrared spectroscopy and electroencephalography for monitoring of human brain activity and oxygenation: A review. *Neurophotonics*. 2017;4(4):041411.
- [5] Shibasaki H. Human brain mapping: Hemodynamic response and electrophysiology. *Clinical Neurophysiology*. 2008;119(4):731–743.
- [6] Pinto-Orellana MA, Khan H, Ombao H, Mirtaheri P. Emerging neuroimaging approach of hybrid EEG-fNIRS recordings: Data collection and analysis challenges. *Data Science in Science*. 2024;3(1):2426785.
- [7] Yeung MK, Chu VW. Viewing neurovascular coupling through the lens of combined EEG–fNIRS: A systematic review of current methods. *Psychophysiology*. 2022;59(6):e14054.
- [8] Lee HT, Shim M, Liu X, et al. A review of hybrid EEG-based multimodal human–computer interfaces using deep learning: Applications, advances, and challenges. *Biomedical Engineering Letters*. 2025;15:587–618.
- [9] Kane MJ, Conway ARA, Miura TK, Colflesh GJH. Working memory, attention control, and the N-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2007;33(3):615–622.
- [10] Aghajani H, Garbey M, Omurtag A. Measuring mental workload with EEG+fNIRS. *Frontiers in Human Neuroscience*. 2017;11:359.
- [11] Shin J et al. Open access dataset for EEG and fNIRS single-trial classification. *Scientific Data*. 2018;5(1):1–16.
- [12] Hyvärinen A, Oja E. Independent component analysis: Algorithms and applications. *Neural Networks*. 2000;13(4-5):411–430.
- [13] Marriott Haresign I et al. Automatic classification of ICA components from infant EEG using MARA. *Developmental Cognitive Neuroscience*. 2021;52:101024.
- [14] Ludwig KA, Miriani RM, Langhals NB, Joseph MD, Anderson DJ, Kipke DR. Using a common average reference to improve cortical neuron recordings from microelectrode arrays. *Journal of Neurophysiology*. 2009;101(3):1679–1689.
- [15] Croce P, Zappasodi F, Merla A, Chiarelli AM. Exploiting neurovascular coupling: A Bayesian sequential Monte Carlo approach applied to simulated EEG–fNIRS data. *Journal of Neural Engineering*. 2017;14(4):046029.
- [16] Qiu L et al. Multi-modal integration of EEG-fNIRS for characterization of brain activity evoked by preferred music. *Frontiers in Neurorobotics*. 2022;16:823435.
- [17] Khan MU, Hasan MAH. Hybrid EEG-fNIRS BCI fusion using multi-resolution singular value decomposition (MSVD). *Frontiers in Human Neuroscience*. 2020;14:599802.
- [18] Rabbani MHR, Islam SMR. Deep learning networks based decision fusion model of EEG and fNIRS for classification of cognitive tasks. *Cognitive Neurodynamics*. 2024;18:1489–1506.