Application of Large-Scale Foundation Models and Multimodal Fusion for Stress Detection in Glider Pilots Under Real-World Flight Conditions

Antonina Wolszczak
Poznan University of Technology
The Faculty of Computing and Telecommunications
Poznan, Poland
antonina.wolszczak@student.put.poznan.pl

Maria De Marsico
Sapienza University of Rome
Department of Computer Science
Rome, Italy
demarsico@di.uniroma1.it

Extended Abstract

Monitoring the cognitive and emotional workload of pilots during flight operations remains a critical challenge in aviation safety. While mobile brain-computer interfaces (BCIs) offer continuous neurophysiological monitoring capabilities, predictive models trained in pristine, simulator-based environments (e.g., NASA MATB-II) poorly generalize "in-the-wild" due to massive environmental artifacts and pronounced inter-subject physiological variance [1]-[3]. This study rigorously evaluates various machine learning paradigms – spanning classical algorithms, deep convolutional neural networks (CNNs), and novel large-scale foundation models – for stress detection using multimodal data collected from six pilots in an actual SZD-9 BIS "Bocian" glider cockpit. To overcome the unfeasibility of in-flight questionnaires, ground truth labels were generated via automated facial expression analysis using the OpenFace framework [4] and Facial Action Coding System (FACS) [5]. A novel physical marker protocol utilizing inertial detection of pilot head nods was implemented to synchronize the high-frequency EEG stream (recorded via an 8-channel Unicorn Hybrid Black amplifier) with visual behavioral labels. The cleaned, overlapping 5-second EEG epochs were subsequently evaluated using a subject-independent, Leave-One-Group-Out cross-validation scheme to prevent data leakage.

The results, detailed in Table I, expose critical bottlenecks in deploying BCI systems in real-world aviation. Classical machine learning models demonstrated a strict performance ceiling, with the XGBoost algorithm [6] achieving a maximum ROC AUC of 0.654. Deep convolutional networks (EEGNet, TSception) pre-trained on massive simulator datasets failed entirely upon fine-tuning on the target glider dataset, with their predictive performance collapsing to chance levels. This provides vital empirical evidence of the Negative Transfer and Domain Shift phenomena caused by disparate hardware topologies. Conversely, the LaBraM foundation model [7] successfully broke the classical performance ceiling, achieving an average ROC AUC of 0.730. However, evaluation revealed its extreme sensitivity to stochastic weight initialization and data segmentation, with cross-validation fold scores fluctuating drastically between 0.492 (Fold 1) and 0.857 (Fold 5). This enormous variance perfectly encapsulates the Data Starvation phenomenon, proving that massive foundation models require significantly larger datasets to stabilize their representations. Ultimately, generic, subject-independent BCI models are fundamentally insufficient for real-world aviation, necessitating a paradigm shift toward subject-specific calibration and temporal context modeling.

TABLE I.
Cross-Validation Performance (ROC AUC)

Method	Architecture	ROC AUC
Machine Learning	XGBoost	0.654
Deep Transfer Learning	TSception	~0.500 (Chance)
Large Foundation Model	LaBraM	0.730 (± 0.132)

References

[1] C. D. Wickens, "Situation awareness and workload in aviation," Current Directions in Psychological Science, vol. 11, no. 4, pp. 128-133, 2002.

[2] K. B. B., [Author 2], [Author 3], [Author 4], and [Author 5], "Multi-attribute task battery II (MATB-II) for cognitive workload assessment," NASA Technical Reports Server, 2020.

[3] F. Lotte et al., "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," Journal of Neural Engineering, vol. 15, no. 3, p. 031005, 2018.

[4] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 59-66.

[5] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto, 1978.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[7] W. Jiang et al., "LaBraM: Large brain model for learning generic representations with neural decoding," arXiv preprint arXiv:2505.23042, 2024