# Classifying Cognitive Load in a Quasi-Realistic Scenario Based on Multimodal Data Fusion

## Cross-Subject Prediction Using Self-Reports Compared to Task Load

Sabrina Gado
Sabrina.Gado@uni-wuerzburg.de
University of Würzburg

Katharina Lingelbach
Katharina.Lingelbach@uni-oldenburg.de
University of Oldenburg
Fraunhofer Institute for Industrial Engineering IAO

Mathias Vukelić
Mathias.Vukelic@iao.fraunhofer.de
Fraunhofer Institute for Industrial Engineering IAO

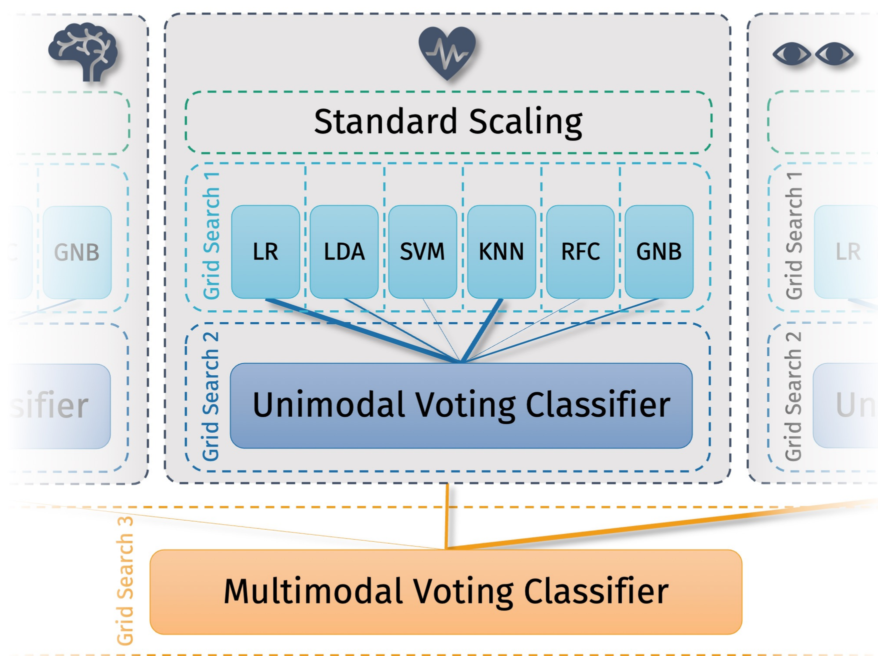
## 1 BACKGROUND

- When situational demands exceed available cognitive resources, people experience cognitive overload which often leads to erroneous behavior [1, 2, 3].
- In naturalistic scenarios, various situational distractions occur that may impede maintenance of goal-directed behavior [4].
- To prevent incidents in safety critical contexts, **closed-loop human-computer systems** should adapt flexibly to users' current cognitive resources.
- Therefore, **robust, non-intrusive measures of cognitive load** as well as suitable classification procedures are required.

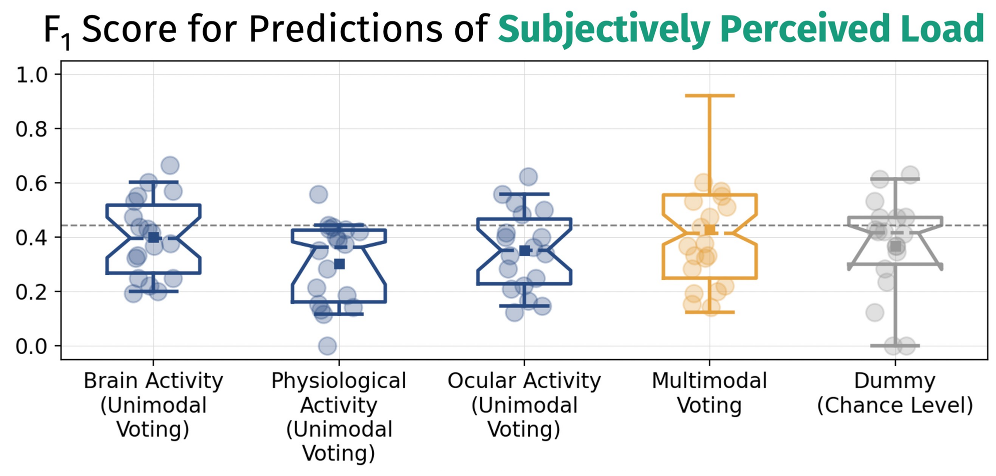
## 2 METHODS

- We conducted a **multimodal study** with 18 participants (9 female, mean age = 25.9 ± 3.8 years, range = 21 - 35).
- Participants performed an adapted version of the **warship commander task** (WCT) [5] with concurrent **emotional speech distractions** taken from the Berlin Database of Emotional Speech (Emo-DB) [6].

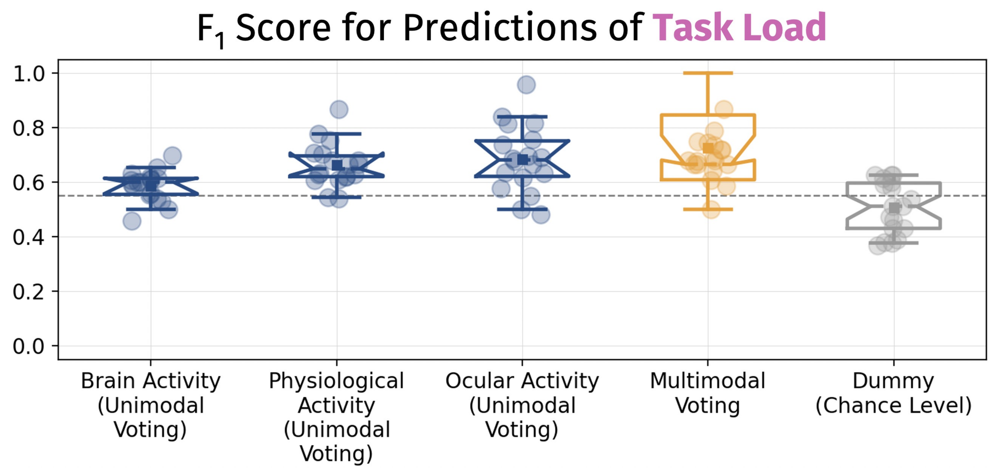
- Participants' **current cognitive load** (high vs. low) was operationalized as
  1) **Subjectively perceived load** (based on self-reports acquired with the Nasa TLX effort subscale [7])
  2) **Task load** (induced by different levels of difficulty in the experimental conditions).
- We recorded **brain activity** (fNIRS), **physiological activity** (heart rate, respiration, and body temperature), and **ocular activity** (pupil dilation and fixations).
- Aggregated features were then fed into a **multilevel data fusion and classification architecture** comprising unimodal and multimodal combinations of classifiers.
- To evaluate the models' performances, we computed the average $F_1$ score with each subject serving as test set once (**cross-subject leave-one-out classification**).




$F_1$ Score for Predictions of **Subjectively Perceived Load**



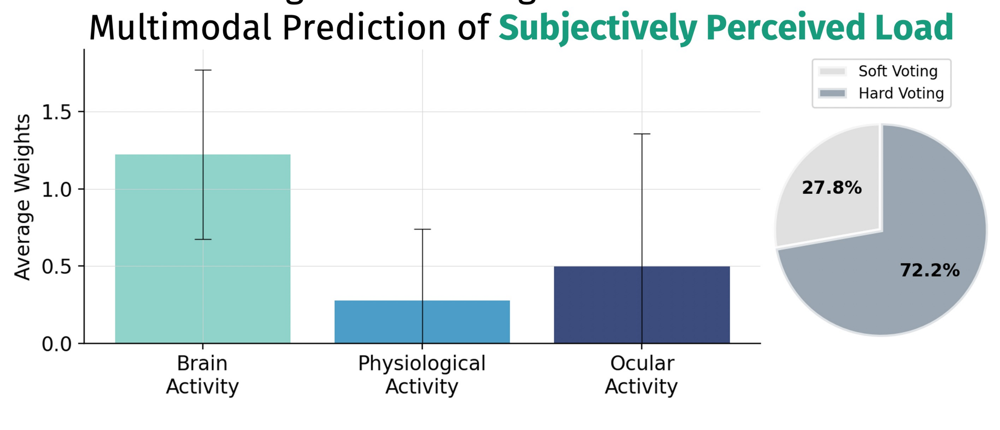

$F_1$ Score for Predictions of **Task Load**



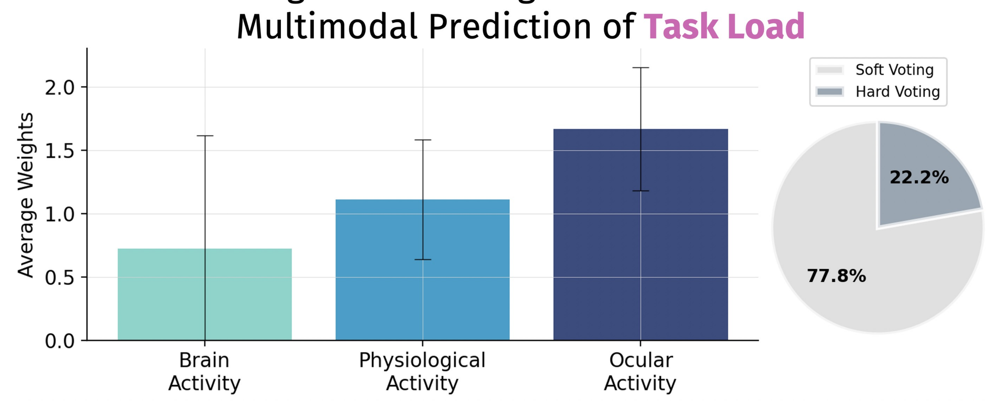

Confusion Matrix for the Predictions of

| | **Subjectively Perceived Load** | | **Task Load** | |
|---|---|---|---|---|
| | Correctly Identified High Load (True Positive) 15.36% | Falsely Identified High Load (False Positive) 24.78% | Correctly Identified High Load (True Positive) 41.11% | Falsely Identified High Load (False Positive) 28.89% |
| | Correctly Identified Low Load (True Negative) 37.39% | Falsely Identified Low Load (False Negative) 22.46% | Correctly Identified Low Load (True Negative) 21.39% | Falsely Identified Low Load (False Negative) 8.61% |

Weights and Voting Procedure of the Multimodal Prediction of **Subjectively Perceived Load**




Weights and Voting Procedure of the Multimodal Prediction of **Task Load**




## 3 RESULTS

- We observed substantial **between-subject variation** in the classifiers' performances and in the weighting of the different modalities.
- The choice of the ground truth affected the classifiers' performances substantially:
  1) **Subjectively perceived load**: We could not reliably predict the subjectively perceived cognitive load for any modality neither by a unimodal combination of classifiers nor in a multimodal approach (recall: 40.6%, precision: 38.3%).
  2) **Task load**: The classification of the experimentally induced task load was significantly above chance level for all modalities with **high average performances**.

     The multimodal voting classifier could also predict task load with an average recall of 82.7% and precision of 58.7%.

     **Ocular activity was weighted highest** for the multimodal prediction. **Soft voting was used more often** than hard voting to combine the different modalities.

## 4 DISCUSSION & CONCLUSION

- Our proposed multimodal classification approach contributes to the development of **ecologically valid monitoring systems of cognitive load across individuals**.
- We provide **insights into characteristics of different data fusion and classification strategies** that allow researchers and practitioners to select appropriate methods.
- Deviations between the two ground truth approaches might be explained by the retrospective nature of self-reports. Because they depend on the individual's perception, reasoning, and **unverifiable introspection** they are vulnerable to various perceptual and **response biases** as well as **automatic evaluation processes** [8, 9]
- Our results further highlight the need for suitable methods
  a) to identify **"odd" subjects** who are potentially difficult to predict due to their heterogeneity compared to the training set,
  b) to facilitate **transfer learning** for these individuals and **generalizability** of the models.

We would like to thank Maria Wirzberger, Ron Becker, Alina Schmitz-Hübsch, Michael Bui, and Sophie Felicitas Böhm for their contribution to the project.

Follow me on twitter @sabrinagado

### REFERENCES

[1] Bakker, A. B., & Demerouti, E. (2007). The Job Demands-Resources model: State of the art. *Journal of Managerial Psychology*, *22*(3), 309–328.
[2] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.
[3] Bowling, N. A., Alarcon, G. M., Bragg, C. B., & Hartman, M. J. (2015). A meta-analytic examination of the potential correlates and consequences of workload. *Work & Stress*, *29*(2), 95–113.
[4] Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science*, *19*(3), 143–148.
[5] Becker, R., Stasch, S.-M., Schmitz-Hübsch, A., & Fuchs, S. (2021). Quantitative scoring system to assess performance in experimental environments. *Proceedings of the 14th International Conference on Advances in Computer-Human Interactions (ACHI 2021)*, 91–96.
[6] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, 1517-1520.
[7] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland.
[8] Ranchet, M., Morgan, J. C., Akinwuntan, A. E., & Devos, H. (2017). Cognitive workload across the spectrum of cognitive impairments: A systematic review of physiological measures. *Neuroscience & Biobehavioral Reviews*, 80, 516–537.
[9] Matthews, G., De Winter, J., & Hancock, P. A. (2020). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science*, *21*(4), 369–396.

Follow our project repository on OSF