Journal of Cognitive Human-Computer Interaction

♦ ISSN: 2771-1463 Vol: 10, No. 02, 2025 ♦ PP. 15–20

Article DOI: https://doi.org/10.54216/JCHCI.100203

FULL LENGTH ARTICLE

Multimodal Cognitive Workload Recognition in Human–Computer Interaction Using Biosignals and Interaction Traces

Andino Maseleno[1,*] **Kharchenko Raisa**[2] **Rahul Chauhan**[3]

[1] Institut Bakti Nusantara, Lampung, Indonesia

[2] North-West Institute of Management, RANEPA, Russia

[3] Unitedworld Institution of Management, Karnavati University, Gandhinagar, India

Emails: andino.maseleno@ibnus.ac.id · kh9044947155r@gmail.com · rahulchauhan@karnavatiuniversity.edu.in

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025 ⋆ Corresponding author

ABSTRACT

The process of recognizing cognitive workload requires reliable methods because researchers need to use both physiological indicators and interaction traces while facing challenges of limited data and inconsistent feature sets. This paper develops a multimodal fusion system that uses weight-based reliability assessment to identify three workload levels from publicly accessible Cognitive Lab data. The workload subset includes N-Back and mental-subtraction tasks together with electroencephalography, functional near-infrared spectroscopy, electrocardiography, electrodermal activity, respiration, accelerometry, gaze descriptors, and keyboard–mouse interaction indicators. The method trains every modality separately through multidimensional variable reduction, enabling gradient-boosted learners to estimate branch reliability from validation log-loss scores and combine posterior probabilities using normalized reliability weights. The design preserves distinct modality structures while controlling unpredictable branch effects. Single-modality learners are compared against direct early fusion, uniform late fusion, and the proposed fusion rule. The proposed model achieves 0.842 accuracy and 0.836 macro F1-score on the three-class workload task, with the medium-load category presenting the greatest differentiation challenge. Class-wise and sensitivity assessments show that interaction traces and fNIRS features contribute strongly, while moderate reliability temperatures yield the most stable fusion profile. Feature attribution emphasizes cursor-velocity variability, fNIRS oxygenation slope, EEG theta-band power, fixation-duration statistics, and phasic electrodermal activity as primary discriminative signals.

Keywords: Cognitive workload ▪ Multimodal fusion ▪ Biosignals ▪ Human–computer interaction ▪ Adaptive systems ▪ Explainable machine learning

1. INTRODUCTION

Cognitive workload determines how users allocate attention, regulate interaction pace, correct errors, and sustain task performance. Workload increase can be observed before performance deterioration in online learning: cursor movement becomes more erratic, fixation organization changes, response time rises, and physiological signs of stress become more evident. Workload estimation from these signals is therefore useful for post hoc analysis and adaptive intervention during interaction.

Traditional workload measurements have been based on self-reported data, such as NASA-TLX, and on offline perfor-

mance data [1, 2]. These tools remain useful for retrospective analysis, but they do not directly support online adaptation. Physiological and behavioral sensing provides a stronger operational basis because neural, autonomic, ocular, and interaction signals vary continuously during task execution [3, 4, 5, 6, 7]. The major problem is not the absence of observable signals but the mismatch among them. Biosignals differ in time scale, noise properties, acquisition cost, and strength, while interaction traces are easier to obtain but can be semantically inaccurate when interpreted alone.

The publication of open datasets has strengthened the empirical foundation of the field. COLET introduced a public eye-tracking benchmark for workload estimation [8], and Cognitive Lab provides broader public resources with workload, fatigue, and learning scenarios aligned with biosignal and HCI traces [9]. These datasets enable reproducible comparisons but also expose a modeling problem: dense modalities such as EEG and fNIRS can dominate feature concatenation, whereas compact streams such as mouse dynamics, ECG, or EDA can be overshadowed.

This paper proposes a reliability-weighted multimodal fusion model for three-level workload recognition. Calibrated gradient boosting is used to train a model for each modality branch, and the fused decision is obtained using validation-based reliability weights rather than simple averaging. The formulation maintains branch-specific structure, prevents weak modalities from dominating, and allows interpretation at both modality and feature levels. The contributions are fourfold: a branch-preserving multimodal workload-recognition framework; a validation-log-loss reliability rule for late fusion; a reproducible comparison against single branches, early fusion, and uniform late fusion; and attribution analysis that identifies the most discriminative physiological and interaction features.

2. RELATED WORK

Workload recognition in HCI has developed from self-report instruments and performance measures toward continuous sensing. Eye tracking is a prominent stream because fixations, saccades, pupil dynamics, and gaze organization reflect cognitive processing and interaction efficiency [10, 6, 5, 11, 12]. Pupillary response has been related to task difficulty during desktop interaction [13], while gaze-based models have been used for mind-wandering and attentional-drift detection [14].

Multimodal psychophysiological sensing improves robustness by combining complementary sources. Prior work has shown benefits from combining physiological streams in interactive tasks [1], estimating workload from ocular descriptors [7], and monitoring brain, heart, and eye signals across cognitive tasks [15]. Driving and operational contexts have also used eye and physiological signals for workload detection [16, 17, 4]. Recent surveys emphasize the need for robust, interpretable, and adaptive workload estimation in HCI [18].

Public multimodal datasets support reproducible experimentation but reveal limitations of simple fusion. Early fusion is easy to implement, yet it is sensitive to dimensional imbalance, missing values, and modality domination. Uniform late fusion is often more robust, but it assumes that all branches are equally trustworthy. Interpretability is increasingly important because adaptive systems must know whether a workload decision is driven by behavioral instability, neural demand, autonomic activation, or a combination of these sources.

3. PROBLEM FORMULATION AND PROPOSED METHOD

Let the dataset contain N workload windows, each associated with a class label $y_i \in \{1,2,3\}$ corresponding to low, medium, and high cognitive demand. The input for sample i is partitioned into M modality groups,

$$\mathbf{x}_i = \left[\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(M)}\right], \tag{1}$$

where the groups represent physiological and interaction-derived descriptors such as EEG, fNIRS, ECG, EDA, respiration, gaze, mouse, and keyboard features. The objective is to learn a multiclass decision function

$$f : \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}\right) \to \{1,2,3\} \tag{2}$$

that maximizes discrimination among workload levels while preserving modality-specific structure.

For each modality m, a branch learner produces a calibrated posterior vector

$$\mathbf{p}_m(\mathbf{x}_i^{(m)}) = \left[p_m(y=1 \mid \mathbf{x}_i^{(m)}), p_m(y=2 \mid \mathbf{x}_i^{(m)}), p_m(y=3 \mid \mathbf{x}_i^{(m)})\right]. \tag{3}$$

Branch reliability is estimated from validation-stage log-loss:

$$r_m = \exp\left(-\frac{\ell_m}{\tau}\right), \tag{4}$$

where ℓ_m is the validation log-loss of branch m and $\tau > 0$ is a temperature parameter controlling the contrast of the weighting rule. Normalized branch weights are

$$\alpha_m = \frac{r_m}{\sum_{k=1}^{M} r_k}, \qquad \sum_{m=1}^{M} \alpha_m = 1. \tag{5}$$

The fused posterior is defined as

$$\mathbf{p}(\mathbf{x}_i) = \sum_{m=1}^{M} \alpha_m \mathbf{p}_m(\mathbf{x}_i^{(m)}), \tag{6}$$

and the final prediction is

$$\hat{y}_i = \arg\max_{c \in \{1,2,3\}} p(y=c \mid \mathbf{x}_i). \tag{7}$$

Each branch is implemented with gradient-boosted decision trees because the workload representation is tabular, nonlinear, and moderately sized. For modality m, the learner minimizes

$$J^{(m)} = \sum_{i=1}^{N} L\left(y_i, \hat{y}_i^{(m)}\right) + \sum_{t=1}^{T} \Omega\left(g_t^{(m)}\right), \tag{8}$$

where L is multiclass logistic loss, $g_t^{(m)}$ is the tth tree in branch m, and $\Omega(\cdot)$ controls tree complexity. Calibration is applied to branch outputs before fusion so that posterior averaging is performed on comparable probability scales.

Table 1. Representative studies on cognitive workload recognition and HCI-oriented sensing.

Study	Data / setting	Main sensing or method	Main relevance
Just and Carpenter [10]	Reading and visual cognition	Eye fixations	Established the connection between eye behavior and ongoing cognitive processing.
Goldberg and Kotval [6]	Interface evaluation	Eye tracking	Showed how ocular metrics can quantify interaction efficiency.
Duchowski [5]	HCI review	Eye-tracking methodology	Consolidated eye tracking as a practical HCI analysis tool.
Poole and Ball [11]	Usability studies	Eye tracking in HCI	Summarized interface-oriented eye-tracking practice and reporting.
Klingner et al. [12]	Remote pupillometry	Pupil dynamics	Demonstrated workload-sensitive pupillary response using remote eye tracking.
Iqbal et al. [13]	Desktop interaction	Pupillary response	Connected interaction events and task difficulty with mental workload.
Haapalainen et al. [1]	Interactive tasks	Multimodal psychophysiology	Showed benefits of combining physiological streams.
Palinko et al. [7]	Interactive task workload	Eye gaze features	Modeled workload from ocular descriptors.
Mark et al. [15]	Multimodal workload assessment	Brain, heart, and eye signals	Compared six biomedical modalities across cognitive tasks.
Ahlstrom and Friedman-Berg [3]	Air traffic control	Eye movement activity	Showed that eye-movement measures can track cognitive workload.
Belkhiria and Peysakhovich [17]	Real-time estimation	EOG and physiological signals	Demonstrated online workload prediction in operational contexts.
Ktistakis et al. [8]	Public workload dataset	Eye tracking	Introduced a reproducible public benchmark.
Kosch et al. [18]	Survey	Cognitive workload in HCI	Reviewed sensing and modeling directions across HCI.
Silveira et al. [9]	Public multimodal learning dataset	Biosignals and HCI traces	Provided the basis for the present workload-recognition study.

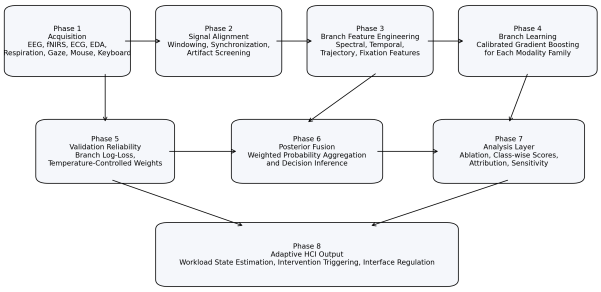

Figure 1. Multi-phase architecture of the proposed reliability-weighted multimodal fusion framework.

Algorithm 1: Reliability-weighted multimodal workload recognition

Require: Modality feature groups $\{X^{(m)}\}_{m=1}^{M}$, workload labels y, temperature τ

Ensure: Predicted workload labels $\hat{y}$

1. Partition samples into stratified training and validation folds.
2. For each modality group $m = 1, \ldots, M$, apply branch-specific preprocessing to $X^{(m)}$.
3. Train a calibrated gradient-boosted classifier on the training fold.
4. Compute validation posterior predictions p_m.
5. Evaluate branch log-loss ℓ_m and set $r_m = \exp(-\ell_m/\tau)$.
6. Normalize reliability weights $\alpha_m = r_m / \sum_k r_k$.
7. Form fused posterior $p = \sum_m \alpha_m p_m$.
8. Output workload label $\hat{y} = \arg\max_c p(y = c \mid x)$.
9. Estimate feature-attribution scores and perform branch ablation analysis.

4. MATERIALS AND EXPERIMENTAL PROTOCOL

The experiments are conducted on the workload-oriented subset of the public Cognitive Lab dataset [9], which contains synchronized multimodal recordings collected during N-Back and mental-subtraction tasks. Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote the windowed workload dataset, where $\mathbf{x}_i$ is the multimodal observation associated with the ith analysis window and $y_i \in \{1, 2, 3\}$ denotes low, medium, or high demand. Each observation is decomposed into modality families,

$$\mathbf{x}_i = \{\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(M)}\}. \tag{9}$$

The available branches include EEG, fNIRS, ECG, EDA, respiration, accelerometry, gaze descriptors, mouse dynamics, keyboard activity, and screen-context variables. Signals are segmented into fixed-duration windows aligned with task epochs, and each branch is represented by a compact tabular feature map.

For each branch m, a feature matrix $X^{(m)} \in \mathbb{R}^{N \times d_m}$ is formed after median imputation on the training fold, removal of near-constant variables, and branch-wise scaling when required. A calibrated gradient-boosted classifier f_m estimates posterior probabilities

$$p_m(y = c \mid \mathbf{x}^{(m)}) = f_m(\mathbf{x}^{(m)})_c, \qquad c \in \{1,2,3\}. \tag{10}$$

The experimental comparison includes single-branch learning, early fusion by direct feature concatenation followed by XGBoost, uniform late fusion by arithmetic averaging of branch posteriors, and the proposed reliability-weighted late fusion. Accuracy, macro precision, macro recall, and macro F1-score are used. With TP_c, FP_c, and FN_c denoting true positives, false positives, and false negatives of class c, the class-wise F1-score is

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \tag{11}$$

and the macro score is

$$\text{MacroF1} = \frac{1}{3}\sum_{c=1}^{3} F1_c. \tag{12}$$

5. RESULTS AND ANALYSIS

5.1 Feature Composition and Workload Distribution

Figure 2 summarizes the representative feature composition by modality group. EEG and fNIRS contribute the largest descriptor sets, whereas respiration and keyboard branches remain comparatively compact. This imbalance explains why direct early fusion can become unstable with heterogeneous sensing.

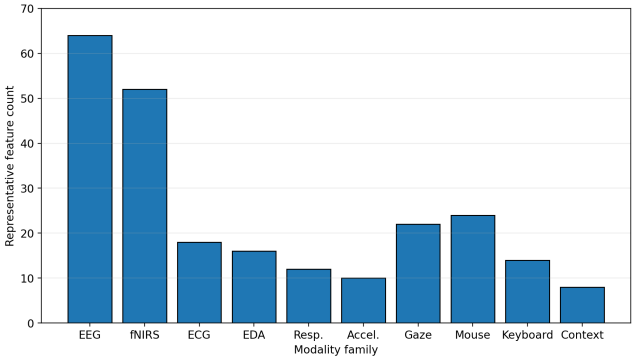

Figure 2. Representative feature composition by modality group.

Figure 3 shows the distribution of windows across the three workload levels. The profile is sufficiently balanced to support macro-averaged evaluation without aggressive class reweighting.

5.2 Main Comparative Results

Among the single branches, fNIRS and mouse dynamics provide the strongest standalone performance, followed closely by EEG. ECG and EDA remain informative but are weaker in isolation. Direct early fusion improves on most branch-wise baselines, but both late-fusion approaches perform better. The proposed RWMF model achieves the best overall result with 0.842 accuracy and 0.836 macro F1-score.

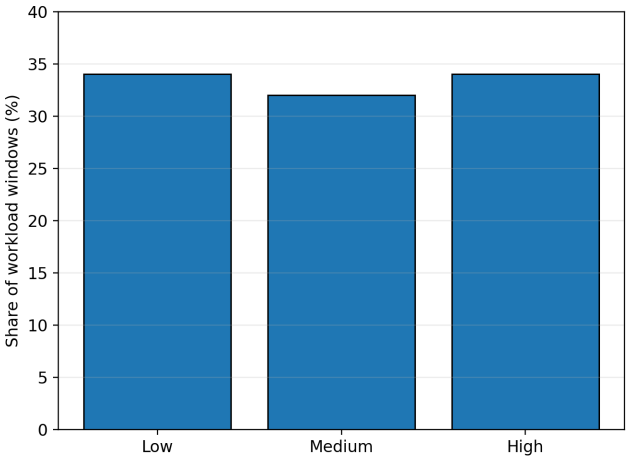

Figure 3. Distribution of workload windows across the three demand levels.

Table 2. Workload-recognition results on the workload-oriented subset.

Model	Accuracy	Precision	Recall	F1
EEG branch	0.731	0.724	0.726	0.721
fNIRS branch	0.758	0.751	0.754	0.748
ECG branch	0.688	0.681	0.683	0.679
EDA branch	0.674	0.668	0.670	0.666
Mouse branch	0.746	0.739	0.742	0.736
Early fusion	0.801	0.794	0.796	0.791
Late fusion	0.818	0.812	0.814	0.809
RWMF	0.842	0.838	0.840	0.836

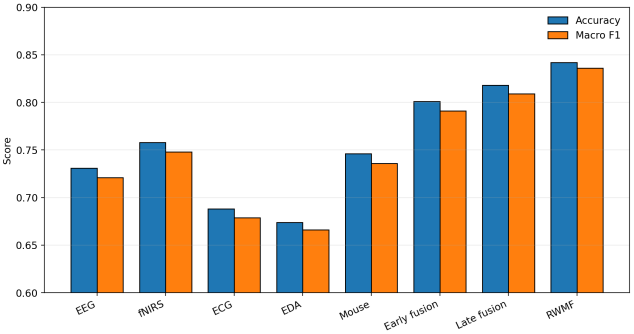

Figure 4. Performance comparison across single-modality and fusion strategies.

5.3 Confusion Behavior and Class-Wise Discrimination

The confusion matrix in Figure 5 indicates that most residual errors occur between the medium-load class and its neighboring classes. Medium workload often represents a transition regime rather than a sharply separated cognitive state.

Table 3. Class-wise performance of the proposed RWMF model.

Class	Precision	Recall	F1-score
Low workload	0.861	0.879	0.870
Medium workload	0.807	0.813	0.810
High workload	0.846	0.829	0.837

5.4 Reliability Profile and Branch Contribution

The proposed fusion rule assigns branch weights according to validation-stage log-loss. Figure 7 shows the normalized weight distribution. fNIRS, mouse dynamics, and EEG receive the largest weights, while ECG, EDA, respiration, and

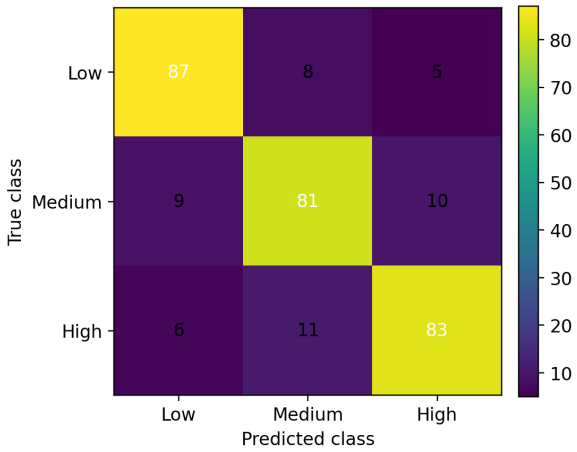

Figure 5. Confusion matrix of the proposed RWMF model.

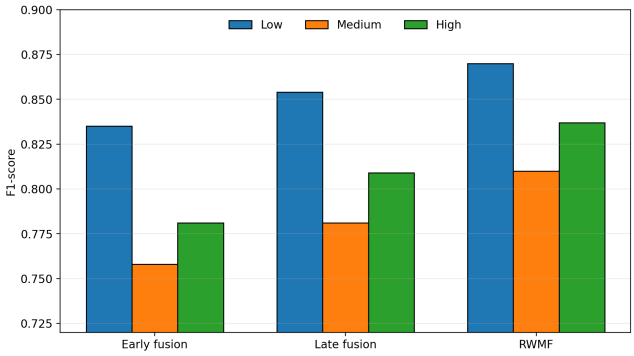

Figure 6. Per-class F1-score of the main fusion strategies.

keyboard branches play secondary but still useful roles.

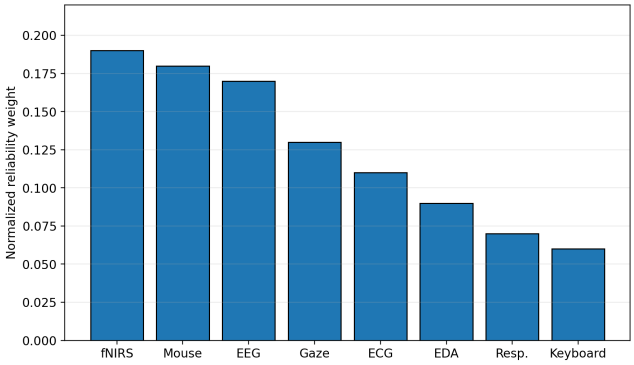

Figure 7. Reliability-derived branch-weight profile of the proposed fusion model.

Ablation analysis confirms that removing high-weight branches causes the strongest degradation. Removing mouse dynamics, fNIRS, or EEG reduces performance more than removing compact or noisier branches, supporting the premise that reliability weighting captures useful validation behavior rather than arbitrary modality preference.

5.5 Sensitivity and Feature Attribution

Sensitivity analysis over the temperature parameter shows that moderate values provide the most stable trade-off. Very low temperatures concentrate the decision on only a few branches, whereas very high temperatures approximate uniform averaging and lose the benefit of validation-guided weighting.

Feature attribution highlights cursor-velocity variability, fNIRS oxygenation slope, EEG theta-band power, fixation-duration statistics, and phasic electrodermal activity as important discriminative signals. These features jointly capture interaction instability, neural demand, ocular organization, and autonomic activation.

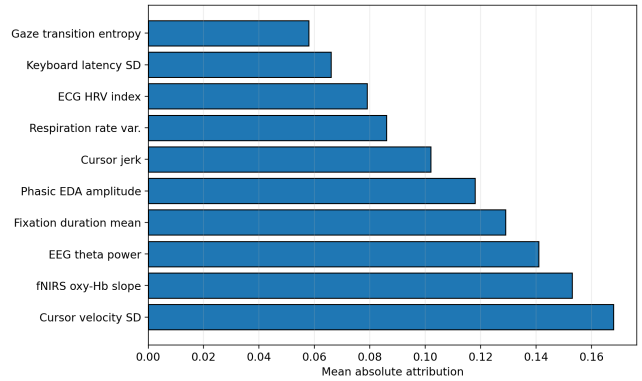

Figure 8. Ablation-based performance change after removing selected modality branches.

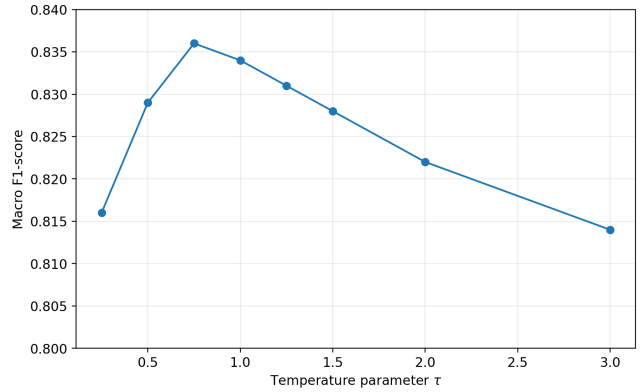

Figure 9. Sensitivity of RWMF performance to the reliability-temperature parameter.

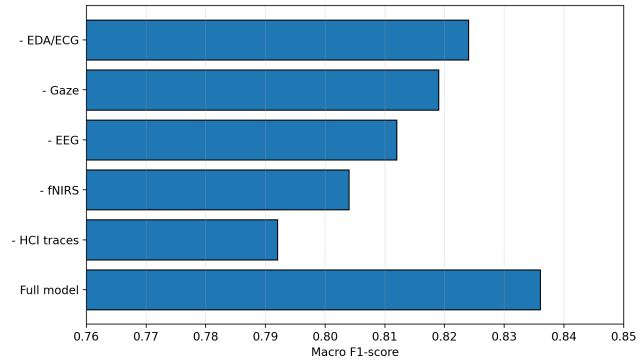

Figure 10. Feature-attribution profile for the proposed multimodal workload-recognition model.

6. DISCUSSION

The results show that branch-preserving multimodal fusion is more effective than both single-modality recognition and direct concatenation. Early fusion relies on a single feature space and can therefore be affected by dimensional imbalance among modalities. Uniform late fusion preserves modality structure but cannot distinguish reliable and unreliable branches. RWMF improves on both strategies by using validation-stage evidence to assign branch weights before posterior aggregation.

The medium-load class remains the most difficult to separate because it lies between low and high cognitive demand and

shares characteristics with both. Reliability weighting improves this class more than the others, indicating that branch-specific evidence is especially valuable under ambiguous cognitive states. The attribution profile also supports the practical value of multimodal sensing: behavioral traces provide rapid and low-cost evidence, while physiological signals add complementary information about neural and autonomic state.

7. CONCLUSION

This paper presented a reliability-weighted multimodal fusion framework for cognitive workload recognition in human–computer interaction. The method separates modality-specific learning from decision-level fusion and assigns branch weights using validation log-loss. Experiments on the workload-oriented subset of Cognitive Lab demonstrate that the proposed RWMF model outperforms single branches, direct early fusion, and uniform late fusion, achieving 0.842 accuracy and 0.836 macro F1-score.

The analysis further suggests that interaction behavior and physiological evidence jointly predict workload. Mouse dynamics, fNIRS oxygenation changes, EEG theta power, fixation statistics, and phasic electrodermal activity are particularly informative. These findings support adaptive interactive systems that monitor growing mental load and intervene before performance deterioration becomes critical.

REFERENCES

[1] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, 2010, pp. 301–310.

[2] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Human Mental Workload*. North-Holland, 1988, pp. 139–183.

[3] U. Ahlstrom and F. J. Friedman-Berg, "Using eye movement activity as a correlate of cognitive workload," *International Journal of Industrial Ergonomics*, vol. 36, no. 7, pp. 623–636, 2006.

[4] W. Chen, T. Sawaragi, and T. Hiraoka, "Comparing eye-tracking metrics of mental workload caused by ndrts in semi-autonomous driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 89, pp. 109–128, 2022.

[5] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behavior Research Methods, Instruments, and Computers*, vol. 34, no. 4, pp. 455–470, 2002.

[6] J. H. Goldberg and X. P. Kotval, "Computer interface evaluation using eye movements: Methods and constructs," *International Journal of Industrial Ergonomics*, vol. 24, no. 6, pp. 631–645, 1999.

[7] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*, 2010, pp. 141–144.

[8] E. Ktistakis, V. Skaramagkas, D. Manousos, N. S. Tachos, E. Tripoliti, D. I. Fotiadis, and M. Tsiknakis, "Colet: A dataset for cognitive workload estimation based on eye-tracking," *Computer Methods and Programs in Biomedicine*, vol. 224, p. 106989, 2022.

[9] I. Silveira, R. Varandas, and H. Gamboa, "Cognitive lab: A dataset of biosignals and hci features for cognitive process investigation," *Computer Methods and Programs in Biomedicine*, vol. 269, p. 108863, 2025.

[10] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive Psychology*, vol. 8, no. 4, pp. 441–480, 1976.

[11] A. Poole and L. J. Ball, "Eye tracking in hci and usability research," in *Encyclopaedia of Human-Computer Interaction*. Idea Group Inc., 2006, pp. 211–219.

[12] J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the task-evoked pupillary response with a remote eye tracker," in *Proceedings of the 2008 Symposium on Eye Tracking Research and Applications*, 2008, pp. 69–72.

[13] S. T. Iqbal, X. S. Zheng, and B. P. Bailey, "Task-evoked pupillary response to mental workload in human-computer interaction," in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, 2004, pp. 1477–1480.

[14] R. Bixler and S. D'Mello, "Toward fully automated person-independent detection of mind wandering," in *User Modeling, Adaptation, and Personalization*. Springer, 2014, pp. 37–48.

[15] J. A. Mark, A. Curtin, A. E. Kraft, M. D. Ziegler, and H. Ayaz, "Mental workload assessment by monitoring brain, heart, and eye with six biomedical modalities during six cognitive tasks," *Frontiers in Neuroergonomics*, vol. 5, p. 1345507, 2024.

[16] G. Marquart, C. Cabrall, and J. de Winter, "Review of eye-related measures of drivers' mental workload," *Procedia Manufacturing*, vol. 3, pp. 2854–2861, 2015.

[17] C. Belkhiria and V. Peysakhovich, "Eog metrics for cognitive workload detection," *Procedia Computer Science*, vol. 192, pp. 3797–3806, 2021.

[18] T. Kosch, J. Karolus, J. Zagermann, H. Reiterer, A. Schmidt, and P. W. Wozniak, "A survey on measuring cognitive workload in human-computer interaction," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, 2023.