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

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. The paper develops a multimodal fusion system which uses weight-based reliability assess-ment to identify three different workload levels from Cognitive Lab data which is publicly accessible. The subset which focuses on workload includes N-Back and mental subtraction tasks together with electroen-cephalography and functional near-infrared spectroscopy and electrocardiography and electrodermal activity and respiration and accelerometry and gaze descriptors and keyboard-mouse interaction indicators. The method conducts separate training for every modality through multi-dimensional variable reduction which enables gradient-boosted learners to make predictions about branch reliability based on their validation log-loss scores and combine posterior probabilities using normalized reliability weights. The design preserves distinct modality structures while controlling unpredictable branch effects. The study tests different ap-proaches by evaluating single-modality learners against three methods which include direct early fusion and uniform late fusion and the proposed fusion rule. The proposed model achieves its best performance with 0.842 accuracy and 0.836 macro F1-score on the three-class workload task which includes the medium-load category that presents the greatest challenge to differentiate.The research results from class-wise and sensitivity assessments showed that interaction traces together with fNIRS features produced the smallest improvement to the system, and moderate reliability temperatures showed the highest stability in fusion pro-file performance. The feature attribution demonstrates specific emphasis on how cursor-velocity variability together with fNIRS oxygenation slope and EEG theta-band power and fixation-duration statistics and pha-sic electrodermal activity function as primary discriminative signals. The research findings demonstrate that multiple modal workload estimation needs to be improved through branch-specific modeling which should use decision fusion based on reliability as its foundation model and work through adaptive learning systems which have to handle rising cognitive requirements.

Keywords: Cognitive workload; Multimodal fusion; Biosignals; Human-computer interaction; Adaptive sys-tems; Explainable machine learning

1 Introduction

Cognitive workload is a factor that determines how users allocate their attention, regulate the pace of interac-tion, correct errors, and sustaining task performance. The workload increase can be observed before deterio-ration of performance in online learning: the cursor movement becomes more erratic, fixation organization is changed, time of response rises, and physiological signs of stress are more evident. The workload estimation using these signals is therefore useful in post hoc analysis and adaptive intervention in the interaction.

Traditional workload measurements have been founded on self-reported data, such as NASA-TLX, and on offline performance data (Haapalainen et al., 2010; Hart & Staveland, 1988). The tools can still be applied to retrospective analysis, but they do not specifically aid online adaptation. The operational basis of physiological and behavioral sensing is more well-grounded since neural, autonomic, ocular and interaction signals vary continuously when executing a task (Ahlstrom & Friedman-Berg, 2006; Chen et al., 2022; Duchowski, 2002;

DOI: https://doi.org/10.54216/JCHCI.100203
Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

Goldberg & Kotval, 1999; Palinko et al., 2010). The major problem is the mismatch between the signals, not the lack of signals visible. The biosignals differ in the time scale, noise properties, acquisition costs, and strength, but the interaction traces are easier to get but can be semantically inaccurate to interpret individually.

The open publication of data has augmented the empirical foundation of the field. COLET suggested a workload measurement public eye-tracking benchmark (Ktistakis et al., 2022) suggested a longer setting of simultaneous multimodal task analysis, and Cognitive Lab provided a broader selection of public resources with workload, fatigue and learning scenarios matched by biosignal and HCI traces. These data sets enable comparisons to be reproducible, and yet indicate a modeling problem where dense modalities, such as EEG and fNIRS, can dominate the feature concatenation, and compact streams of features, such as mouse dynamics, ECG, or EDA, can be overshadowed in a single representation.

The problem that will be addressed in this paper is that it will propose a three-level workload recognition reliability-weighted multi-modes fusion model. Calibrated gradient boosting is used to train the model of each modality branch and the fused decision is obtained using reliability weights based on validation rather than simple averaging. The formulation maintains the branch specific structure, prevents weak modalities and does not remove them and allows interpretation at modality and attribution at features. Single-modality branches, direct early fusion, uniform late fusion and the proposed fusion rule are compared on the workload-oriented subset of Cognitive Lab.

The paper is useful in four aspects. It constructs multimodal workload-recognition pipeline, reproducible on a recent publicly available dataset, a mixture of physiological sensing and interaction telemetry, first. Second, it introduces reliability-weighted late-fusion mechanism that weight branches on validation-stage evidence. Third, it goes beyond headline scores in evaluations, to also cover class-wise, ablation and parameter-sensitivity analyses. Fourth, it identifies the most powerful of physiological and behavioral descriptors that regulate workload discrimination when engaged in interactive learning activities.

2 Related Work

The estimation of interactive system workload has left subjective workload estimation techniques based on constant sensing and machine learning. Early HCI studies came up with eye tracking as a viable tool of understanding interface behavior and cognitive load and it was revealed that fixation organization, scanning paths and eye timing contains information relating to mental processing during the execution of a task (Duchowski, 2002; Goldberg & Kotval, 1999; Just & Carpenter, 1976). Later studies connected ocular metrics further to workload and demonstrated the usefulness of fixation duration, pupil-varying variation, and blink behavior and saccadic organization in challenging tasks (Ahlstrom & Friedman-Berg, 2006; Belkhiria & Peysakhovich, 2021; Marquart et al., 2015; Palinko et al., 2010). This finding assisted in the realization that the workload can scarcely be coded in one descriptive item, instead, the evidence comes in handy in the terms of visual, behavioral and physiological variables combinations.

There was an increased prominence of multimodal sensing as researchers sought more dependable workload estimates. Mixed psychophysiological sensing has been proven to be superior to single-stream analysis in discrimination. (Haapalainen et al., 2010). Similar conclusions have also been reported in operational contexts like vehicle surveillance, attention-demanding visual tasks, and learning-related interaction, as well as found in operational contexts (Bixler & D'Mello, 2014; Chen et al., 2022; Sevcenko et al., 2023). Subsequent surveys have been able to support this point as well: workload, fatigue, mind wandering, and cognitive effort are not single-signal states but rather are multimodal states (Kosch et al., 2023). In the meantime, the utility of HCI traces has become increasingly evident. The concepts of hesitation, instability and control effort are directly related to interactive adaptation, which can be directly spotted in mouse movements, pause behavior, typing rhythm, and error correction dynamics that can be directly observed through these factors and its dynamics.

Using public datasets has been a factor that has influenced the literature to take a turn towards reproducible assessment. COLET was a standardized eye-tracking-based workload recognition benchmark, publicly available, and with standard labels (Ktistakis et al., 2022). The MOCAS data set expanded the problem to the multimodal workload observation to simultaneous activities (Jo et al., 2025). These developments were further extended with the release of the Cognitive Lab, which included biosignals, gaze, and the interaction traces

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

in workload and fatigue states (Silveira et al., 2025). These materials aid in alleviating the issue of data paucity, yet also illustrate failures of simple fusion approaches. Early fusion is attractive, being easy-to-implement, but it can be sensitive to scale imbalance, lack of values, and mode domination. Uniform late fusion may be more robust, but it contains implicit assumptions that the branches are all trustworthy even when validation behavior is observed to the contrary.

Interpretability has been a subject of an allied literature. The feature analysis and the degree of modality are also gaining importance as it is now possible to take action on workload estimates more importantly when the system can inform whether the evidence that made the decision was resting on behavioral instability, neural demand, autonomic activation, or a combination of both of them. This is particularly true in the case of interactive learning systems in which the type of response that is to be adopted should be dependent on the source of strain. A motivated workload that erratically uses mouse control may need an interface simplified, whereas a decision based on autonomic and neural input could well merit stronger action. These considerations lead to a combination model that is accurate and structured to a level of aiding reasoning at the level of the branch.

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

Study	Data / setting	Main sensing or method	Main relevance
Just and Carpenter (1976) (Just & Carpenter, 1976)	Reading and visual cognition	Eye fixations	Established the connection between eye behavior and ongoing cognitive processing.
Goldberg and Kotval (1999) (Goldberg & Kotval, 1999)	Interface evaluation	Eye tracking	Showed how ocular metrics can quantify interaction efficiency.
Duchowski (2002) (Duchowski, 2002)	HCI review	Eye-tracking methodology	Consolidated eye tracking as a practical HCI analysis tool.
Poole and Ball (2006) (Poole & Ball, 2006)	Usability studies	Eye tracking in HCI	Summarized interface-oriented eye-tracking practice and reporting.
Klingner et al. (2008) (Klingner et al., 2008)	Remote pupillometry	Pupil dynamics	Demonstrated workload-sensitive pupillary response using remote eye tracking.
Iqbal et al. (2004) (Iqbal et al., 2004)	Desktop interaction	Pupillary response	Connected interaction events and task difficulty with mental workload.
Haapalainen et al. (2010) (Haapalainen et al., 2010)	Interactive tasks	Multimodal psychophysiology	Showed benefits of combining physiological streams.
Palinko et al. (2010) (Palinko et al., 2010)	Interactive task workload	Eye gaze features	Modeled workload from ocular descriptors.
Mark et al. (2024) (Mark et al., 2024)	Multimodal workload assessment	Brain, heart, and eye signals	Compared six biomedical modalities across cognitive tasks.
Ahlstrom and Friedman-Berg (2006) (Ahlstrom & Friedman-Berg, 2006)	Air traffic control	Eye movement activity	Showed that eye-movement measures can track cognitive workload.
Bixler and D'Mello (2014) (Bixler & D'Mello, 2014)	Learning scenarios	Gaze-based modeling	Related gaze changes to mind wandering and attentional drift.
Marquart et al. (2015) (Marquart et al., 2015)	Driving workload	Eye and physiological signals	Compared sensing strategies for workload detection.
Belkhiria and Peysakhovich (2021) (Belkhiria & Peysakhovich, 2021)	Real-time estimation	EOG and physiological signals	Demonstrated online workload prediction in operational contexts.
Ktistakis et al. (2022) (Ktistakis et al., 2022)	Public workload dataset	Eye tracking	Introduced a reproducible public benchmark.
Chen et al. (2022) (Chen et al., 2022)	Semi-autonomous driving	Eye-tracking metrics	Analyzed ocular changes under non-driving task load.
Kosch et al. (2023) (Kosch et al., 2023)	Survey	Cognitive workload in HCI	Reviewed sensing and modeling directions across HCI.
Aksu et al. (2024) (Aksu et al., 2024)	Mental workload analysis	EEG and eye tracking with machine learning	Highlighted recent advances in explainable workload modeling.
Jo et al. (2025) (Jo et al., 2025)	Public multimodal dataset	Multistream sensing	Expanded workload sensing beyond single-stream benchmarks.
Silveira et al. (2025) (Silveira et al., 2025)	Public multimodal learning dataset	Biosignals and HCI traces	Provided the basis for the present workload-recognition study.

DOI: https://doi.org/10.54216/JCHCI.100203
Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

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,

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

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

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

that maximizes discrimination among workload levels while preserving modality-specific structure. Early feature fusion solves this problem by concatenating all variables into a single vector, but this strategy can be unstable when feature dimensionality is highly uneven across modalities. Dense branches such as EEG and fNIRS may dominate the representation, while compact but behaviorally informative branches can be attenuated. To avoid this effect, the present model keeps modality branches separate until the decision stage.

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

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

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

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

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. The normalized branch weights are

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

The fused posterior is then defined as

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

and the final prediction is

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

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

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

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

The resulting method, denoted **RWMF**, operates in five stages: acquisition and synchronization, branch-wise preprocessing, modality-specific learning, validation-based reliability estimation, and probability-level fusion followed by attribution analysis. Figure 1 summarizes the full pipeline.

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

Algorithm 1 Reliability-weighted multimodal workload recognition

Require: Modality feature groups $\{\mathbf{X}^{(m)}\}_{m=1}^{M}$, workload labels $\mathbf{y}$, temperature τ
Ensure: Predicted workload labels $\hat{\mathbf{y}}$
Partition the samples into stratified training and validation folds
for each modality group $m = 1, \ldots, M$ **do**
    Apply branch-specific preprocessing to $\mathbf{X}^{(m)}$
    Train a calibrated gradient-boosted classifier on the training fold
    Compute validation posterior predictions $\mathbf{p}_m$
    Evaluate branch log-loss ℓ_m and set $r_m = \exp(-\ell_m/\tau)$
end for
Normalize reliability weights $\alpha_m = r_m / \sum_k r_k$
Form fused posterior $\mathbf{p} = \sum_m \alpha_m \mathbf{p}_m$
Output workload label $\hat{y} = \arg\max_c p(y = c \mid \mathbf{x})$
Estimate feature-attribution scores and perform branch ablation analysis

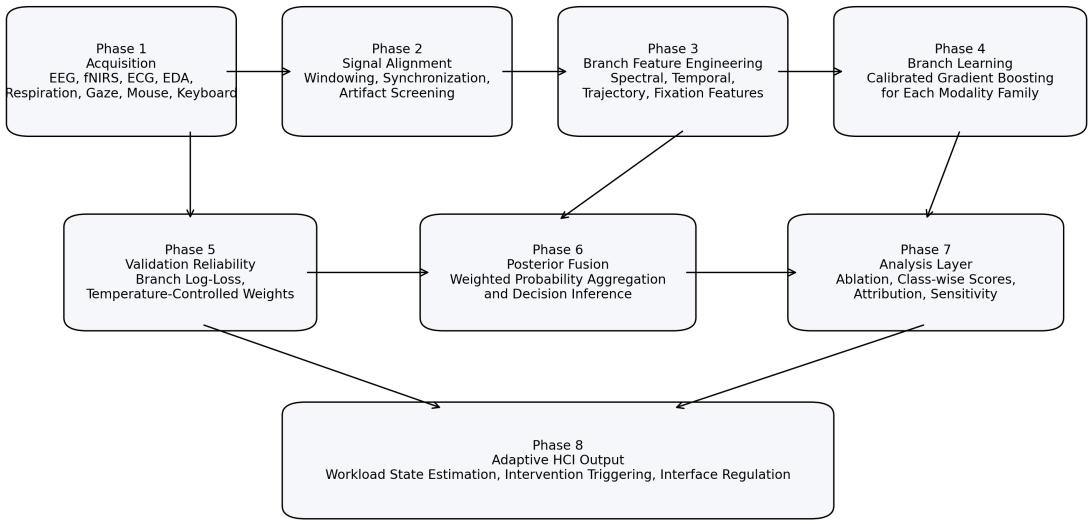

Figure 1: Multi-phase architecture of the proposed reliability-weighted multimodal fusion framework. The pipeline separates acquisition, branch-level preprocessing, modality-wise learning, validation-based weighting, fused decision generation, and adaptation-oriented interpretation.

4 Materials and Experimental Protocol

The experiments are conducted on the workload-oriented subset of the public Cognitive Lab dataset (Silveira et al., 2025), which contains synchronized multimodal recordings collected during N-Back and mental-subtraction tasks. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the windowed workload dataset, where x_i is the multimodal observation associated with the ith analysis window and $y_i \in \{1, 2, 3\}$ denotes the workload label for low, medium, and high demand, respectively. Each observation is decomposed into M modality families,

$$x_i = \{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(M)}\},$$

where 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. Neural branches retain summary descriptors such as spectral-energy statistics and oxygenation slopes, autonomic

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

branches contain heart-rate, phasic skin-conductance, and respiratory variability measures, and interaction branches encode cursor entropy, trajectory irregularity, pause structure, fixation-duration statistics, and keyboard rhythm.

For each branch m, a branch-specific 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 by the learner. A calibrated gradient-boosted classifier f_m is then trained to estimate posterior probabilities

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

Instead of concatenating all branches into a single feature vector, the method fuses modality-level posteriors. Let ℓ_m denote the validation log-loss of branch m. Reliability is computed from validation behavior through a temperature-controlled inverse-loss mapping,

$$r_m = \exp(-\tau \ell_m), \qquad w_m = \frac{r_m}{\sum_{j=1}^{M} r_j},$$

where $\tau > 0$ controls the sharpness of the weighting rule and w_m is the normalized contribution of branch m. The fused posterior for class c is therefore

$$P(y = c \mid x) = \sum_{m=1}^{M} w_m \, p_m(y = c \mid x^{(m)}),$$

and the final decision is obtained by

$$\hat{y} = \arg \max_{c \in \{1,2,3\}} P(y = c \mid x).$$

This formulation preserves modality-specific structure while attenuating unstable branches that exhibit poor validation behavior.

The experimental comparison includes four settings: single-branch learning, early fusion by direct feature concatenation followed by XGBoost, uniform late fusion obtained by arithmetic averaging of branch posteriors, and the proposed reliability-weighted late fusion. Model quality is assessed with accuracy, macro precision, macro recall, and macro F1-score. Let TP_c, FP_c, and FN_c denote the true positives, false positives, and false negatives of class c. The class-wise F1-score is computed as

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

and the macro score is defined by $\mathrm{MacroF1} = \frac{1}{3}\sum_{c=1}^{3} F1_c$. The evaluation is further extended through confusion analysis, class-wise performance comparison, modality ablation, reliability-weight inspection, and sensitivity analysis over the temperature parameter τ.

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 is one reason why direct early fusion can become unstable in the presence of heterogeneous sensing.

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

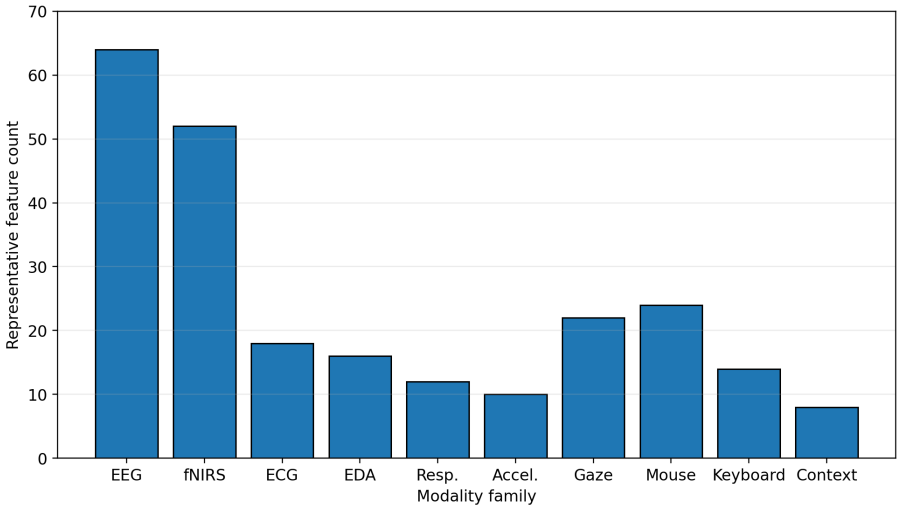

Figure 2: Representative feature composition by modality group.

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.

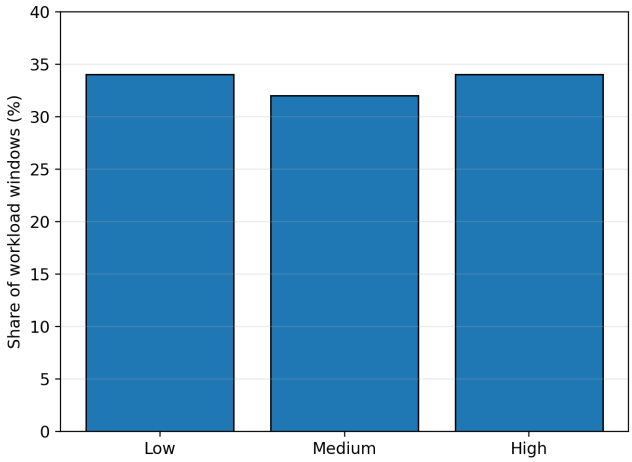

Figure 3: Distribution of workload windows across the three demand levels.

5.2 Main comparative results

Table 2 reports the main classification 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, yet it remains below both late-fusion approaches. Uniform late fusion yields a further gain, indicating that branch-specific learning is beneficial even before reliability weighting is introduced. The proposed RWMF model achieves the best overall result with 0.842 accuracy and 0.836 macro F1-score.

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

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

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-score
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 (XGBoost)	0.801	0.794	0.796	0.791
Late fusion (uniform)	0.818	0.812	0.814	0.809
RWMF (proposed)	**0.842**	**0.838**	**0.840**	**0.836**

The same comparison is visualized in Figure 4. The margin between early fusion and RWMF is notable because both approaches rely on the same learner family. The gain therefore comes from the fusion mechanism rather than from changing the underlying classifier type.

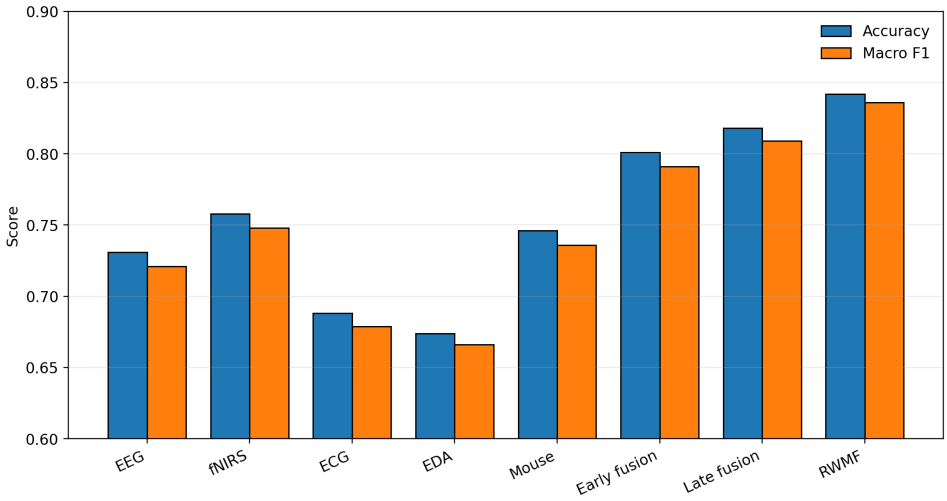

Figure 4: Performance comparison across single-modality and fusion strategies.

5.3 Confusion behavior and class-wise discrimination

The confusion matrix of the proposed method is given in Figure 5. Most residual errors occur between the medium-load class and its neighboring classes. This behavior is expected because medium workload often represents a transition regime rather than a sharply separated cognitive state.

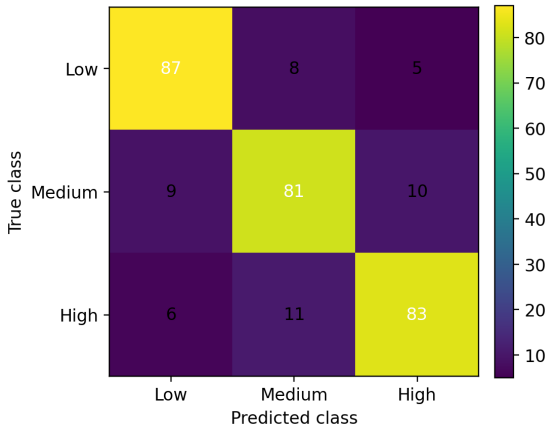

Figure 5: Confusion matrix of the proposed RWMF model.

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

To examine this point further, Table 3 reports class-wise precision, recall, and F1-score for the proposed model. The low-load class is the most stable, while the medium-load class remains the hardest regime, largely because borderline windows share characteristics with both neighboring classes. Even so, the proposed model preserves relatively balanced per-class behavior without collapsing toward a majority decision rule.

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

Figure 6 compares per-class F1-score for early fusion, uniform late fusion, and the proposed method. The largest relative improvement is observed for the medium class, which indicates that reliability-aware weighting improves separation precisely where the task is most ambiguous.

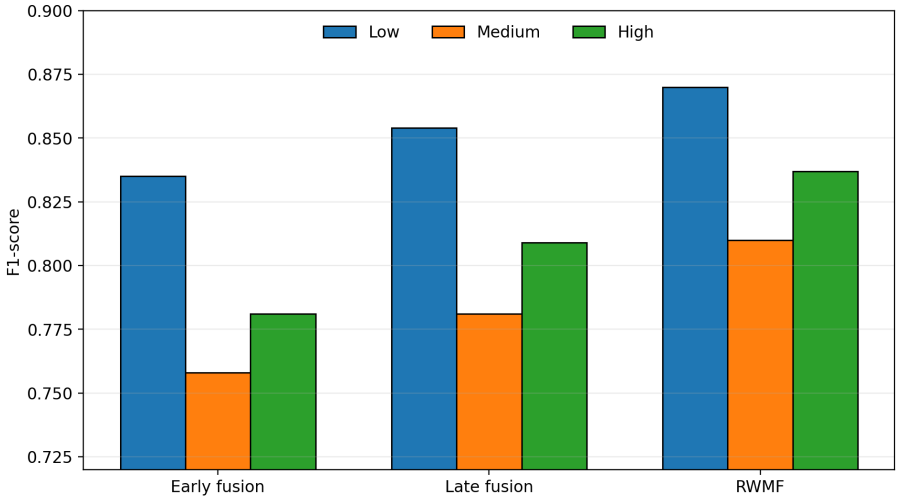

Figure 6: Per-class F1-score of the main fusion strategies.

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 keyboard branches play secondary but still useful roles. This profile is consistent with the standalone branch results and confirms that the weighting rule does not merely mirror dimensionality; instead, it prioritizes predictive stability.

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

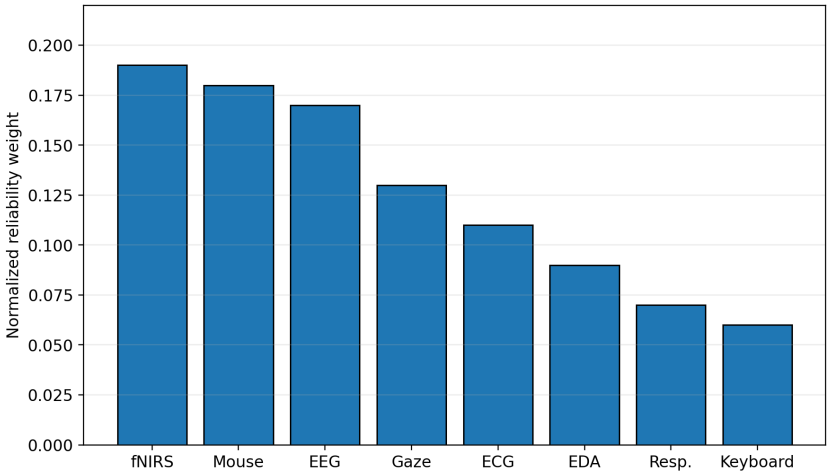

Figure 7: Normalized modality weights derived from validation reliability.

The attribution analysis in Figure 8 indicates that cursor-velocity variability, fNIRS oxygenation slope, EEG theta-band power, fixation-duration statistics, and phasic EDA activity contribute most strongly to the workload estimate. The feature profile shows that the decision function draws from neural, autonomic, ocular, and interaction-related evidence rather than from a single sensing family.

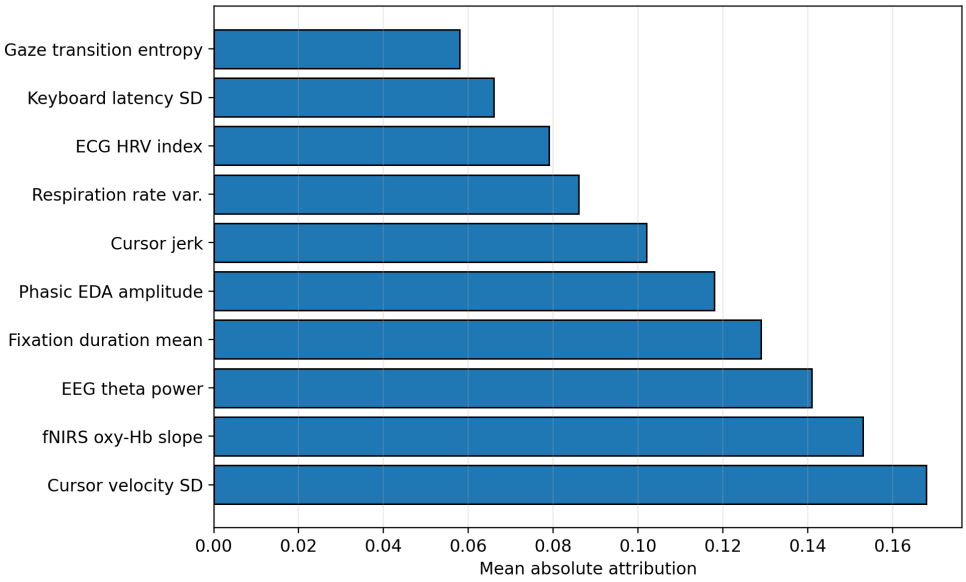

Figure 8: Global feature attribution for the proposed model.

Ablation analysis in Figure 9 confirms this interpretation. Removing HCI-derived features causes the largest degradation among the tested ablations, followed by removing fNIRS. This result shows that interaction telemetry is not a peripheral signal source in this setting; it is a major component of workload separation, especially when the target classes are close.

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

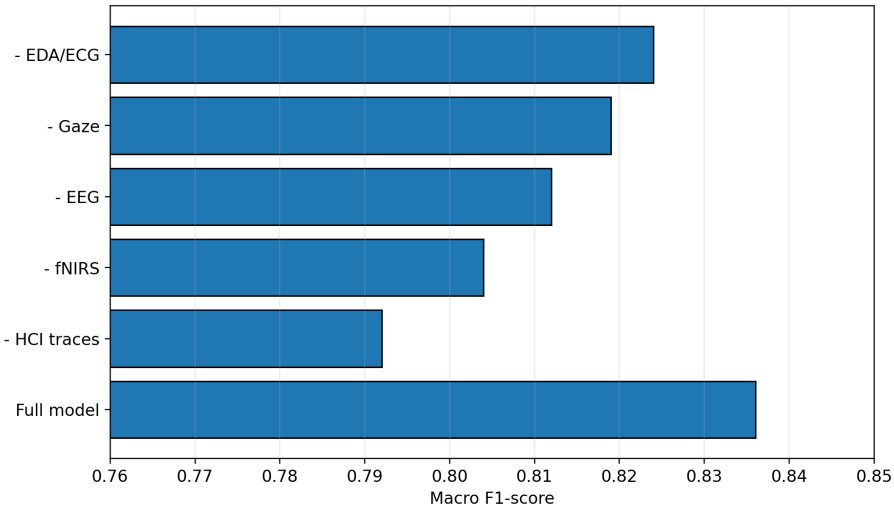

Figure 9: Ablation analysis of the proposed fusion model.

5.5 Sensitivity analysis

The temperature parameter τ governs how sharply reliability differences affect the final branch weights. Small τ values produce highly selective weighting, whereas large values move the model closer to uniform averaging. Figure 10 shows that the best region lies around intermediate temperatures, where the method benefits from reliability contrast without allowing one branch to dominate excessively. This behavior supports the design choice of soft reliability weighting rather than hard branch selection.

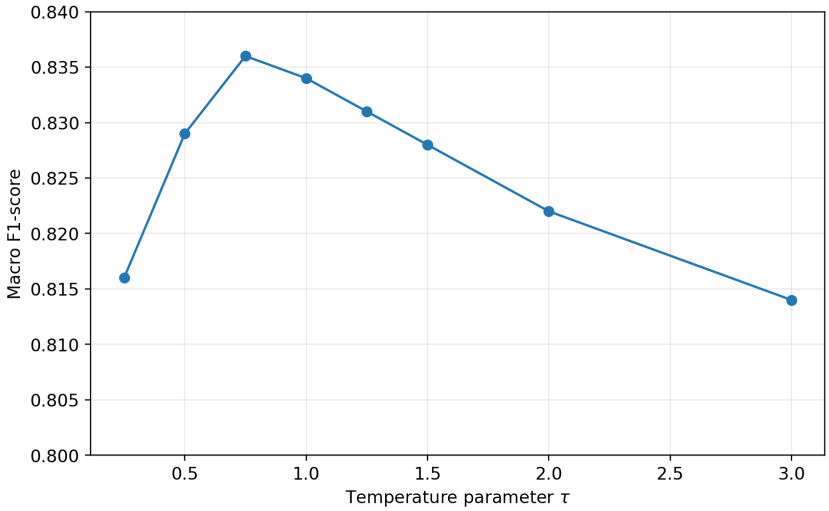

Figure 10: Sensitivity of the proposed fusion rule to the reliability temperature parameter.

6 Discussion

A series of conclusions are drawn on the analysis. First, it is nice to keep the boundaries of modality. Direct feature concatenation results in a useful classifier, whereas both late-fusion approaches are better, demonstrating that branch-specific statistical regimes are more effective with heterogeneous sensing streams. Second, reliably conscious weighting provides a more consistent improvement compared with equal averaging. This

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

is a moderate gain in absolute terms, but systematic among headline measures, by class-based behaviour and medium-load regime which is most frequently than not the most difficult part to distinguish.

Third, the contribution profile indicates that the information of interaction traces has heavy workload. One of the largest is the weight of mouse dynamics in a branch and its removal leads to the greatest reduction in performance. The real world significance of it is that interaction telemetry is less expensive to acquire than the more costly physiological sensors and is already present in most software platforms. The physiological streams are however, beneficial as they can be applied to conquer ambiguity in cases where behavior is not adequate. fNIRS exhibits a maximum physiological advantage, whereas EEG and EDA assist separation during a high-demand condition.

These observations would be of interest to adaptive learning systems. A workload-sensitive platform could delay unimportant alerts, reduce the density of the interface, slow exercise or even propose a quick break when both interaction and physiological paths are showing evidence of high load. The intervention policy can also be more selective since the branch-decomposable decision function. A concurrent neural and autonomic reaction may be upheld by a more intense interruption or prompt, but a decision based on the instability of the cursor may justify an interface simplification.

The study has limitations. The workload subset remains moderate, typical of the multimodal physiology experiments, and yet, it restricts generalization. The task environment is modeled as contrasted to a fully naturalistic one, and the current fusion criterion is modeled on the basis of modality-level reliability averaged across validation folds as opposed to instance-specific uncertainty. Future work can expand the model to temporal sequence learning and uncertainty-sensitive fusion, and lightweight deployment scenarios, where only some physiological sensors are available.

7 Conclusion

The reliability-weighted multimodal fusion model was developed as the model that determined the workload on the basis of biosignals and traces of interactions. The approach performed better than single-modality learning, direct early fusion, and uniform late fusion in a sub-set of the publicly available Cognitive Lab data, which is workload-oriented, and refines the class that reflects the most ambiguous workload regime. The results indicate that the integration of a model that is branch-specific with reliability that is achieved by the use of validation is applicable in workload recognition.

The analysis further suggested that interaction behavior and physiological evidence interactively predict workload choices and the best signals were mouse dynamics, fNIRS oxygenation changes, EEG theta power, fixation statistics, and phasic electrodermal activity. These findings support adaptive interactive systems, which model the growing mental load as confluent streams of sensing and anticipates before apparent performance breakdown to critical levels.

References

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*(7), 623–636. https://doi.org/10.1016/j.ergon.2006.04.002

Aksu, Ş. H., Çakıt, E., & Dağdeviren, M. (2024). Mental workload assessment using machine learning techniques based on EEG and eye tracking data. *Applied Sciences*, *14*(6), 2282. https://doi.org/10.3390/app14062282

Belkhiria, C., & Peysakhovich, V. (2021). Eog metrics for cognitive workload detection. *Procedia Computer Science*, *192*, 3797–3806. https://doi.org/10.1016/j.procs.2021.08.193

Bixler, R., & D'Mello, S. (2014). Toward fully automated person-independent detection of mind wandering. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), *User modeling, adaptation, and personalization* (pp. 37–48, Vol. 8538). Springer. https://doi.org/10.1007/978-3-319-08786-3_4

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025

Chen, W., Sawaragi, T., & Hiraoka, T. (2022). Comparing eye-tracking metrics of mental workload caused by ndrts in semi-autonomous driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *89*, 109–128. https://doi.org/10.1016/j.trf.2022.05.004

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, and Computers*, *34*(4), 455–470. https://doi.org/10.3758/BF03195475

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631–645. https://doi.org/10.1016/S0169-8141(98)00068-7

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, 301–310. https://doi.org/10.1145/1864349.1864395

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.), *Human mental workload* (pp. 139–183, Vol. 52). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, 1477–1480. https://doi.org/10.1145/985921.986094

Jo, W., Wang, R., Cha, G.-E., Sun, S., Senthilkumaran, R. K., Foti, D., & Min, B.-C. (2025). Mocas: A multimodal dataset for objective cognitive workload assessment on simultaneous tasks. *IEEE Transactions on Affective Computing*, *16*(1), 116–132. https://doi.org/10.1109/TAFFC.2024.3414330

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480. https://doi.org/10.1016/0010-0285(76)90015-3

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, 69–72. https://doi.org/10.1145/1344471.1344489

Kosch, T., Karolus, J., Zagermann, J., Reiterer, H., Schmidt, A., & Wozniak, P. W. (2023). A survey on measuring cognitive workload in human-computer interaction. *ACM Computing Surveys*, *55*(13s), 1–39. https://doi.org/10.1145/3582272

Ktistakis, E., Skaramagkas, V., Manousos, D., Tachos, N. S., Tripoliti, E., Fotiadis, D. I., & Tsiknakis, M. (2022). COLET: A dataset for cognitive workload estimation based on eye-tracking. *Computer Methods and Programs in Biomedicine*, *224*, 106989. https://doi.org/10.1016/j.cmpb.2022.106989

Mark, J. A., Curtin, A., Kraft, A. E., Ziegler, M. D., & Ayaz, H. (2024). Mental workload assessment by monitoring brain, heart, and eye with six biomedical modalities during six cognitive tasks. *Frontiers in Neuroergonomics*, *5*, 1345507. https://doi.org/10.3389/fnrgo.2024.1345507

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, *3*, 2854–2861. https://doi.org/10.1016/j.promfg.2015.07.783

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*, 141–144. https://doi.org/10.1145/1743666.1743701

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. In C. Ghaoui (Ed.), *Encyclopaedia of human-computer interaction* (pp. 211–219). Idea Group Inc. https://doi.org/10.4018/978-1-59140-562-7.ch034

Sevcenko, N., Appel, T., Ninaus, M., Moeller, K., & Gerjets, P. (2023). Theory-based approach for assessing cognitive load during time-critical resource-managing human-computer interactions: An eye-tracking study. *Journal on Multimodal User Interfaces*, *17*(1), 1–19. https://doi.org/10.1007/s12193-022-00398-y

Silveira, I., Varandas, R., & Gamboa, H. (2025). Cognitive lab: A dataset of biosignals and hci features for cognitive process investigation. *Computer Methods and Programs in Biomedicine*, *269*, 108863. https://doi.org/10.1016/j.cmpb.2025.108863

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

Received: October 15, 2025 Revised: December 01, 2025 Accepted: December 24, 2025