A Neuro-Symbolic System for Interpretable Multimodal Physiological Signals Integration in Human Fatigue Detection

Mohammadreza Jamalifard, Yaxiong Lei, Parasto Azizinezhad,
Javier Fumanal Idocin, and Javier Andreu-Perez
School of Computer Science and Electronic Engineering
University of Essex, United Kingdom
Email: {m.jamalifard, yaxiong.lei, p.azizinezhad, j.fumanal-idocin, j.andreu-perez}@essex.ac.uk

***Abstract*—We propose a neuro-symbolic architecture that learns four interpretable physiological concepts, *oculomotor dynamics, gaze stability, prefrontal hemodynamics, and multimodal*, from eye-tracking and neural hemodynamics, functional near-infrared spectroscopy, (fNIRS) windows using attention-based encoders, and combines them with differentiable approximate reasoning rules using learned weights and soft thresholds, to address both rigid hand-crafted rules and the lack of subject-level alignment diagnostics. We apply this system to fatigue classification from multimodal physiological signals, a domain that requires models that are accurate and interpretable, with internal reasoning that can be inspected for safety-critical use. In leave-one-subject-out evaluation on 18 participants (560 samples), the method achieves 72.1% ± 12.3% accuracy, comparable to tuned baselines while exposing concept activations and rule firing strengths. Ablations indicate gains from participant-specific calibration (+5.2 pp), a modest drop without the fNIRS concept (-1.2 pp), and slightly better performance with Łukasiewicz operators than product (+0.9 pp). We also introduce *concept fidelity*, an offline per-subject audit metric from held-out labels, which correlates strongly with per-subject accuracy ($r = 0.843$, $p < 10^{-4}$).**

***Index Terms*—neuro-symbolic AI, approximate reasoning, fatigue detection, eye-tracking, oculomotor features, fNIRS, concept bottleneck, interpretable machine learning**

I. Introduction

We introduce a neurosymbolic system designed to address the lack of mechanisms supporting *subject-level auditing and troubleshooting* in models applied to noisy or distribution-shifted multimodal physiological signals [1]. Despite advances in multimodal sensing (e.g., eye tracking, EEG, fNIRS) [2], current deep learning approaches are limited by their susceptibility to inter-subject variability and struggle with real-world generalization [3]–[5]. We apply this system to the problem of fatigue, which substantially impairs human performance in safety-critical domains including transportation, healthcare, and industrial operations [6].

First, interpretable fatigue models typically rely on hand-crafted rules inspired by literature [5]. However, inter-individual and context-dependent variations often alter the magnitude, and even the direction of physiological responses, rendering fixed rule sets may be unreliable across different subjects and tasks [7]. This motivates data-adaptive concept definitions and thresholds.

We address these limitations through a neuro-symbolic framework with three key contributions:

1) We propose an attention-based concept extraction architecture that learns four interpretable physiological concepts including *oculomotor dynamics, gaze stability, prefrontal hemodynamics, and multimodal*, from eye-tracking and fNIRS features using data-adaptive thresholds explained in the methodology.
2) We introduce a differentiable approximate reasoning layer as a structured interpretability module, with transparent fixed rule templates and end-to-end learned rule weights and soft thresholds that adapt to the data distribution.
3) We conduct extensive ablation studies on normalization, concept importance, logic operator variants, and learned vs. fixed thresholds, and report concept fidelity as a reliability metric ($r = 0.843$ correlation with accuracy).

II. Related Work

A. Physiological Fatigue Detection

Multimodal approaches combining oculomotor and neuroimaging signals have shown promise for fatigue assessment. Eye-tracking features including oculomotor dynamics, blink patterns, and gaze stability have been associated with alertness states [8]. Reviews highlight a range of oculomotor metrics including saccade velocity, fixation patterns, and gaze dispersion as indicators of fatigue-related performance degradation [9]. However, the relative contribution of specific eye-tracking features varies across studies and populations, motivating data-driven feature weighting rather than a priori assumptions about indicator importance. fNIRS captures prefrontal hemodynamic changes associated with sustained attention and fatigue-related resource changes [10]. The prefrontal cortex shows increased oxygenated hemoglobin concentration under higher cognitive workload [11]. However, cross-subject generalization remains challenging due to substantial individual differences [12]. Relatively fewer works jointly model eye tracking and fNIRS for fatigue assessment under cross-subject evaluation settings.

B. Neuro-Symbolic AI and Approximate Reasoning

Neuro-symbolic systems integrate neural pattern recognition with symbolic reasoning, providing both learning flexibility and interpretability [13]. A recent systematic review found research concentrated in learning/inference (63%) and logic/reasoning (35%), with gaps in explainability [14]. Logic Tensor Networks enable differentiable reasoning through fuzzy semantics [15]. Concept Bottleneck Models (CBMs) constrain predictions through human-interpretable intermediate representations [16], with recent extensions including post-hoc CBMs [17]. Fuzzy logic provides a natural framework for physiological computing where concepts exist on continuous spectra [18], [19]. Recent surveys highlight growing interest in neuro-symbolic approaches for safety-critical applications [20].

C. Uncertainty and Reliability in Classification

For high-stakes decisions, interpretable models are inherently preferable to post-hoc explanations [21]. Model calibration is essential for trustworthy deployment [22]. MC Dropout [23] and deep ensembles [24] provide prediction-level uncertainty, but do not identify systematic subject-level factors affecting reliability. Individual differences in physiological responses pose fundamental challenges for cross-subject generalization [25]. Our concept fidelity metric addresses this gap by quantifying how well learned representations capture discriminative information for each individual.

III. Experimental Setup

A. Data Collection

We collected multimodal physiological data from 18 healthy adults (10 female, 8 male; age 27.7 ± 6.5) during an ethics-approved fatigue protocol. Participants provided written consent and reported normal/corrected-to-normal vision with no epilepsy, neurological/psychiatric disorders, or skin allergies. The 40–60 minute session occurred in a controlled lab using a chin-rest for stabilization. Standard eye-tracker calibration and drift correction were performed.

The experimental session comprised three phases: (1) baseline assessment: A battery of oculomotor tasks including pro-saccade, anti-saccade, and smooth pursuit tasks; we use eye tracking here as a sensing modality for window-level oculomotor features rather than for fine-grained cognitive inference. (2) fatigue induction: A high-load sequence consisting of sustained visual search and mental arithmetic (approximately 30 minutes) designed to deplete cognitive resources; and (3) post-task assessment: A repetition of the baseline oculomotor battery to quantify performance degradation and pattern changes.

We employed leave-one-subject-out cross-validation (LOSO-CV), training on 17 subjects and evaluating on the held-out subject in each of 18 folds. Main results (Table IV) use three random seeds (42, 123, 456) to align computational cost across methods including nested CV tuning. Ablation studies (Tables V–VIII) were conducted under the calibration configuration with the same three seeds. We verified key findings are consistent with five-seed runs ($72.1\% \pm 11.8\%$ for NeSy). All baseline methods received identical preprocessed features with the same participant-aware normalization to ensure fair comparison.

TABLE I
Acquisition and preprocessing summary

Component	Core details
Eye-tracking	EyeLink 1000 Plus (2000 Hz); timestamps reconstructed to remove LSL jitter/duplicates.
fNIRS	8-channel CW; bandpass 0.01–0.2 Hz; flat channels ($\sigma < 10^{-10}$) removed.
Pupil preprocessing	Blink detection at 2.5σ below median, 50 ms dilation, linear interpolation; bandpass 0.01–4.0 Hz.
Windows	Modalities aligned to 10 Hz; 10 s windows with 50% overlap; eye features computed at native rate then aggregated.
Labels	Alert (baseline) vs. Fatigued (post-induction); 560 samples (280/280).

TABLE II
Summary of extracted multimodal features (90 dims total).

Modality	Feature Groups & Metrics	Dim.
Eye: Pupil	**Stat:** μ, σ, range, skew, kurtosis. **Dyn:** $\mu, \sigma, \max$ of $\lvert\dot{p}\rvert$ and $\lvert\ddot{p}\rvert$. **Spec:** Bandpower (LF, HF, Ratio), SampEn, Coeff. Var	16
Eye: Oculomotor	**Dispersion:** σ_x, σ_y, corr(x, y), spatial H. **Kinematics:** Velocity v ($\mu, \sigma, \max, P_{90}$), accel. stats. **Events:** Saccade rate, fixation prop. **Trend:** Linear fit slope (x, y), angular Δ stats, SampEn(v).	18
Eye: Eyelid	**Blink:** Rate, duration/IBI stats ($\mu, \sigma, \min$ / $\max$), Percentage of Eye Closure (PERCLOS) (total & weighted).	8
fNIRS	**Global:** μ, σ, skew, range of channel-mean & deriv. **Spec:** VLF, LF, HF powers & ratios. **Regional:** μ, σ, SampEn per ROI (8 groups). **Sym:** L/R diff stats (μ, σ, ∇), corr(L,R), A/P contrast. **Cmplx:** Global SampEn, Hurst exp., outlier prop.	48

B. Baseline Methods

We compare against an ablated model (no logic layer) and standard ML baselines. Baselines are evaluated with (i) scikit-learn defaults and (ii) modest nested-CV tuning (inner-fold grid search) to avoid optimistic bias under LOSO. Normalization is performed within each fold to prevent leakage.

IV. Methods

A. Problem Formulation

The model predicts binary fatigue $y \in \{0, 1\}$ from 10 s windows (50% overlap) of eye-tracking and fNIRS. We extract 90 features (42 eye, 48 fNIRS) capturing oculomotor dynamics and prefrontal hemodynamics, then apply participant-aware normalization before an attention-based concept extractor and a differentiable approximate reasoning rule layer (Fig. 1).

B. Participant-Aware Normalization

To address the high inter-subject variability characteristic of physiological computing [25], we normalize features relative to each participant's pre-task alert state. For participant p, we compute baseline statistics from samples labeled as alert ($y = 0$):

$$\mu_p = \frac{1}{|S_p^0|} \sum_{i \in S_p^0} \mathbf{x}_i, \quad \sigma_p = \text{std}(\{\mathbf{x}_i\}_{i \in S_p^0}) \tag{1}$$

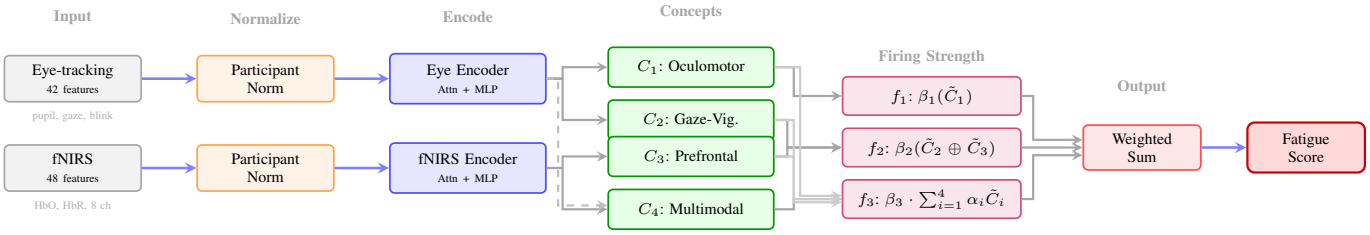

Fig. 1. Neuro-symbolic architecture: participant-normalized features → concept bottleneck (C1–C4) → differentiable Logic Rules → fatigue score.

where S_p^0 denotes the set of alert samples for participant p. Normalized features represent relative change from baseline: $\tilde{\mathbf{x}} = (\mathbf{x} - \mu_p)/(\sigma_p + \epsilon)$, where ϵ is a numerical stability constant.

Evaluation note: Pre-task alert segments are included in the held-out subject's evaluation set; they are only used to estimate μ_p and σ_p (normalization), not to tune model weights. While global normalization achieved higher accuracy in ablation studies (80.3% vs 72.1%), it assumes cohort-estimated population statistics and is less robust to subject-specific shifts and domain shift. We report participant-aware results as the primary configuration to reflect realistic deployment conditions.

C. Concept Extraction with Learned Attention

We extract four concepts capturing distinct physiological aspects of fatigue, following the concept bottleneck paradigm [16]:

1) **Oculomotor Dynamics** (C_1): Encodes gaze kinematics (acceleration, angular velocity, dispersion). It correlates most with gaze acceleration ($r = 0.43$) and strongly with fatigue labels ($r = 0.86$), consistent with degraded oculomotor control under fatigue.
2) **Gaze Stability** (C_2): Captures compensatory oculomotor patterns (gaze slope, angular dynamics). It is negatively correlated with fatigue ($r = -0.65$), suggesting vigilance-maintenance mechanisms that diminish as fatigue increases.
3) **Prefrontal Hemodynamics** (C_3): Represents prefrontal activity from fNIRS channels and is negatively correlated with fatigue ($r = -0.68$), consistent with reduced prefrontal activation in fatigued states [11].
4) **Multimodal** (C_4): Models cross-modal eye–fNIRS interactions with a positive fatigue correlation ($r = 0.70$).

We interpret concepts via post-hoc correlations between concept activations and input features over the full dataset. Notably, C_1 emphasizes gaze kinematics rather than pupil features despite both being available, indicating stronger discriminative value of dynamics in our protocol. This illustrates an advantage of learned concepts over hand-crafted definitions.

Each concept extractor applies learned attention [26] to focus on discriminative features:

$$\mathbf{a} = \sigma(\mathbf{W}_a \tilde{\mathbf{x}}), \quad \mathbf{h} = \text{GELU}(\text{LN}(\mathbf{W}_h(\tilde{\mathbf{x}} \odot \mathbf{a}))) \tag{2}$$

$$C_i = \sigma(\mathbf{v}_i^\top \mathbf{h} + b_i) \tag{3}$$

where σ denotes the sigmoid function, $\odot$ element-wise multiplication, LN is LayerNorm, and concept activations $C_i \in [0, 1]$ are interpreted as membership degrees. The hidden dimension is 64 with dropout rate 0.3.

D. Differentiable Approximate Reasoning with Learned Weights

The neuro-symbolic layer learns concept thresholds (τ_i) and rule weights (β_j) end-to-end. It also applies an attention-like weighting over concepts ($\boldsymbol{\alpha}$) *within the logic layer*, improving robustness to high inter-subject variability in physiological signals. By learning soft thresholds and rule strengths from data, the model avoids the rigidity of fixed hand-crafted rules [13].

We apply soft thresholding with learnable parameters:

$$\tau_i = \sigma(\hat{\tau}_i), \qquad \tilde{C}_i = \sigma\left((C_i - \tau_i) \cdot T\right) \tag{4}$$

where $\hat{\tau}_i$ is the learned threshold parameter (initialized to 0.0 so that τ_i starts at 0.5) and $T = 2.0$ is the temperature controlling decision sharpness. Each concept $C_i \in [0, 1]$ is transformed into a soft-thresholded activation $\tilde{C}_i$ centered around its learned threshold. The model handles inverted feature-label relationships through end-to-end learning.

Three rules combine the soft-thresholded concepts $\tilde{C}_i$ (which are already in $[0, 1]$) using logic operations, as shown in Table III. In particular, $\boldsymbol{\alpha}$ denotes *logic-layer concept weights* used only in the global evidence rule R_3; we parameterize $\boldsymbol{\alpha}$ by unconstrained logits $\mathbf{w}_\alpha \in \mathbb{R}^4$ and map them to a simplex with softmax:

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{w}_\alpha), \tag{5}$$

so that $\alpha_i \in (0, 1)$ and $\sum_{i=1}^{4} \alpha_i = 1$. Larger α_i increases the contribution of $\tilde{C}_i$ to the aggregation in R_3.

TABLE III
DIFFERENTIABLE APPROXIMATE REASONING RULE BASE WITH LEARNED WEIGHTS. EACH RULE MAPS CONCEPT ACTIVATIONS TO FATIGUE EVIDENCE.

Rule	Antecedent → Consequent	Firing Strength
R_1	IF $\tilde{C}_1$ is HIGH THEN Fatigue is LIKELY	$f_1 = \beta_1 \cdot \tilde{C}_1$
R_2	IF ($\tilde{C}_2$ is HIGH) OR ($\tilde{C}_3$ is HIGH) THEN Fatigue is LIKELY	$f_2 = \beta_2 \cdot (\tilde{C}_2 \oplus \tilde{C}_3)$
R_3	IF (Weighted Concept Sum) is HIGH THEN Fatigue is LIKELY	$f_3 = \beta_3 \cdot \left(\sum_{i=1}^{4} \alpha_i \tilde{C}_i\right)$

$\oplus$: t-conorm; default probabilistic sum $a \oplus b = a + b - ab$.
$\tilde{C}_i = \sigma((C_i - \tau_i) \cdot T)$: soft-thresholded activation (Eq. 4).
α_i: learned concept weights ($\sum_i \alpha_i = 1$); β_j: learned rule weights.
$\hat{y} = \sigma(\mathbf{w}^\top [f_1, f_2, f_3] + b)$: consequent aggregation.

Post-hoc analysis revealed differing concept polarities: C_1 correlates positively with fatigue ($r = 0.86$), while C_2 and C_3 correlate negatively ($r = -0.65$ and $r = -0.68$, respectively). Rule R_2 uses the probabilistic-sum t-conorm, $a \oplus b = a + b - ab$, which is smooth and bounded in $[0, 1]$. Rule weights $\boldsymbol{\beta}$ are learned during training. The final prediction combines rule outputs through a learned linear layer and sigmoid, adapting to concept polarities without requiring manual specification of semantics.

E. Training Objective

The total loss combines classification and interpretability regularization:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{div}} + \lambda_2 \mathcal{L}_{\text{sparse}} \tag{6}$$

where $\mathcal{L}_{\text{CE}}$ is binary cross-entropy loss. $\mathcal{L}_{\text{div}}$ penalizes correlations between different concepts to encourage disentanglement. $\mathcal{L}_{\text{sparse}}$ promotes interpretable logic by encouraging low-entropy concept weights $\boldsymbol{\alpha}$, concentrating importance on fewer concepts. We set $\lambda_1 = 0.05$ and $\lambda_2 = 0.001$. Training uses AdamW with learning rate 5×10^{-4}, weight decay 10^{-3}, batch size 32, cosine annealing, gradient clipping at 1.0, and early stopping (patience 20) for up to 150 epochs.

F. Concept Fidelity

We define concept fidelity Φ as a post-hoc, label-dependent audit metric. For subject s, fidelity is computed as the average absolute point-biserial correlation between concept activations and ground-truth labels:

$$\Phi^{(s)} = \frac{1}{k} \sum_{i=1}^{k} \left| r_{pb}(C_i^{(s)}, y^{(s)}) \right| \tag{7}$$

Rule fidelity is defined analogously using rule activations R_j. High fidelity indicates strong concept-label alignment for a given subject, while low fidelity signals potential subject–model mismatch.

V. Results

A. Classification Performance

Table IV proposes LOSO-CV accuracy across all methods. Our neuro-symbolic approach achieves 72.1% ± 12.3%, competitive with both tuned and default baselines while providing interpretable decision pathways. A Wilcoxon signed-rank test comparing NeSy to the best baseline (SVM-RBF tuned) yielded $W = 56$, $p = 0.332$, with matched-pairs rank-biserial correlation $r = 0.27$ (small-to-medium effect); the non-significant result reflects comparable accuracy, with NeSy providing the added benefit of interpretability.

B. Ablation Studies

We conducted ablation studies to isolate the contribution of each architectural component.

TABLE IV
Classification Accuracy (LOSO Cross-Validation)

Method	Accuracy (%)†	95% CI
NeSy (Ours)	**72.1 ± 12.3**	[66.4, 77.8]
No Logic (Ablation)	72.0 ± 12.1	[66.4, 77.6]
Tuned Baselines (nested CV):		
SVM-RBF	69.4 ± 13.5	[63.2, 75.7]
Random Forest	66.6 ± 12.3	[60.9, 72.3]
Extra Trees	68.3 ± 14.3	[61.7, 74.9]
Logistic Regression	67.8 ± 15.1	[60.8, 74.7]
Default Baselines:		
SVM-RBF	67.6 ± 16.4	[60.0, 75.2]
Random Forest	64.7 ± 13.6	[58.4, 71.0]

†Mean ± SD across 18 LOSO folds (seed-averaged per fold); 95% CI across folds

1) Normalization Strategies: Table V compares normalization approaches under realistic deployment conditions. Global normalization (using training-cohort statistics) achieves 80.3%, but it assumes that cohort-level feature distributions transfer to unseen users. Our participant-aware calibration approach achieves 72.1%. Removing calibration entirely drops accuracy to 66.9%, demonstrating a +5.2 pp absolute improvement.

TABLE V
Ablation: Normalization Strategies

Strategy	Accuracy (%)	Std (%)
Global*	80.3	10.1
Participant-Aware (Calibration)†	72.1	12.3
No-Calibration (Train-Only)	66.9	8.3

*Population statistics estimated from the training cohort; may not transfer to unseen users
†Primary configuration; calibration, no test labels used

2) Individual Concept Contributions: Table VI shows the accuracy drop when each concept is ablated (zeroed out) under the calibration configuration. The prefrontal-hemodynamic concept (C_3) contributes most significantly, with a 1.2 pp accuracy drop when removed. This finding aligns with the rule discrimination analysis showing R_2 (which incorporates C_3) as the dominant discriminator.

TABLE VI
Ablation: Individual Concept Contributions

Configuration	Accuracy (%)	Δ (pp)
Full Model (Calibration)	72.1	–
Without C_1 (Oculomotor Dynamics)	73.4	+1.3
Without C_2 (Gaze Stability)	72.4	+0.3
Without C_3 (Prefrontal Hemodynamics)	70.9	-1.2
Without C_4 (Multimodal)	71.2	-0.9

Post-hoc correlation analysis revealed that C_1 learned to weight gaze acceleration and angular dynamics rather than pupil-specific features, despite both being available in the input. This suggests that oculomotor control degradation provided a stronger discriminative signal than pupil dynamics in our protocol. The modest accuracy improvement when removing C_1 (+1.3 pp) and C_2 (+0.3 pp) indicate partial redundancy among eye-tracking concepts, while removing C_3

(Prefrontal-hemodynamic) causes the largest drop (-1.2 pp), consistent with R_2 being the dominant discriminator.

3) Logic Operators: Table VII compares different logic operator families under the calibration configuration. The "Product" configuration uses product t-norm ($a \cdot b$) for conjunction and probabilistic sum ($a + b - ab$) for disjunction in R_2. "Łukasiewicz" uses bounded operators: $\max(0, a+b-1)$ for conjunction and $\min(1, a + b)$ for disjunction. "Gödel" uses $\min$ / $\max$ operations. Łukasiewicz operators achieve the highest accuracy (72.9%).

TABLE VII
ABLATION: LOGIC OPERATORS

Operator	Accuracy (%)	Std (%)	Δ
Product	72.0	12.0	–
Łukasiewicz	**72.9**	12.2	+0.9
Gödel (min/max)	72.2	12.3	+0.2

4) Learned versus Fixed Thresholds: Table VIII compares learned thresholds τ_i against fixed values ($\tau_i = 0.5$ for all concepts) under the calibration configuration. Learned thresholds provide a modest improvement (+1.5 pp), allowing the model to adapt decision boundaries to each concept's distribution. The similar performance suggests the model is relatively robust to threshold initialization.

TABLE VIII
ABLATION: LEARNED VS. FIXED THRESHOLDS

Configuration	Accuracy (%)	Std (%)
Learned Thresholds	**72.1**	12.3
Fixed Thresholds ($\tau = 0.5$)	70.6	12.2

C. Concept Fidelity and Rule Discrimination

Table IX presents per-subject fidelity and rule discrimination metrics. Mean concept fidelity was $\Phi = 0.41 \pm 0.13$, with strong correlation with classification accuracy ($r = 0.843$, $p < 10^{-4}$).

TABLE IX
PER-SUBJECT FIDELITY AND RULE DISCRIMINATION

ID	Acc	C-Fid	R-Fid	R_1	R_2	R_3
P007	96.7	0.75	0.83	-0.03	+0.09	-0.01
P006	86.4	0.55	0.56	-0.05	+0.10	-0.00
P010	85.0	0.46	0.48	-0.05	+0.10	-0.00
P018	85.0	0.44	0.61	-0.04	+0.06	-0.01
P011	83.5	0.45	0.44	-0.04	+0.08	-0.00
P016	81.4	0.48	0.47	+0.01	+0.05	+0.00
P009	54.4	0.25	0.27	+0.00	+0.02	+0.00
P005	55.3	0.14	0.10	-0.01	+0.01	-0.00
P002	58.0	0.39	0.25	+0.01	+0.02	+0.01
P001	58.3	0.16	0.28	-0.03	+0.01	-0.00
Mean	72.1	0.41	0.40	-0.02	+0.05	-0.00

C-Fid: concept fidelity; R-Fid: rule fidelity; R_1–R_3: rule discrimination (Fatigued−Alert). Ten representative subjects shown.

Rule discrimination analysis reveals key patterns:

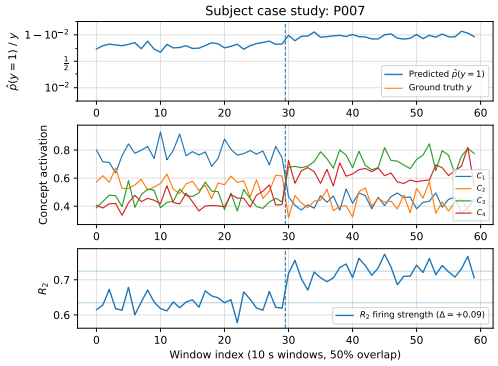

Fig. 2. Case study (P007): fatigue prediction and labels (top), concept activations C_1–C_4 (middle), and rule R_2 firing (bottom); dashed line marks the Alert→Fatigued transition.

- R_2 **(Vigilance & Prefrontal)**: Mean discrimination $+0.052$, Cohen's $d = 1.15$ (large effect). This rule combines gaze stability and prefrontal haemodynamics via probabilistic sum is the primary discriminator.
- R_1 **(Oculomotor Dynamics)**: Mean discrimination -0.021, Cohen's $d = -0.24$. The small negative discrimination at the rule level reflects the interaction between the learned concept (C_1, which correlates positively with fatigue at $r = 0.86$) and the learned rule weight β_1, which the model adjusted to balance contributions across rules.
- R_3 **(Combined)**: Near-zero discrimination (-0.001), serving as a regularizing baseline rather than active discriminator.

Concept fidelity is computed *post-hoc* from ground-truth labels (Eq. 7), and thus serves as an *offline* alignment diagnostic (useful in validation/pilot settings with occasional labels) to flag subjects needing re-calibration or sensor/assessment checks. Developing a label-free reliability proxy that tracks fidelity remains future work.

D. Per-Subject Analysis

The best-performing subject (P007, 96.7% accuracy) achieved near-perfect classification with high concept fidelity (0.75) and strong R_2 discrimination (+0.09). Several subjects (P009, P005, P001, P002) performed below 60% with low concept fidelity and weak rule discrimination, suggesting poor concept–label alignment for these individuals.

We also observe that a subset of participants exhibit weak or even reversed rule discrimination, suggesting that the effective direction of some fatigue-related patterns can vary across individuals; our rule-level diagnostics make such subject-specific deviations explicit.

VI. DISCUSSION

A. Interpretability versus Performance Trade-off

The full model (72.1%) matches the no-logic ablation (72.0%), indicating the approximate reasoning layer primarily improves interpretability rather than accuracy. This is intentional: we trade marginal gains for transparent decision

paths, enabling inspection of which physiological cues drive predictions and whether errors arise from poor signal quality, genuine ambiguity, or subject-specific fatigue patterns.

B. Ablation Study Insights

Ablations yield three key findings. Removing the prefrontal hemodynamics concept (C_3) and multimodal fusion (C_4) drops accuracy by 1.2 pp and 0.9 pp, respectively, while removing gaze-vigilance (C_2) and oculomotor-fatigue (C_1) slightly increases accuracy (+0.3 pp, +1.3 pp), suggesting redundancy or noise sensitivity in gaze-derived metrics. Łukasiewicz pooling performs best (72.9%), exceeding product operators by 0.9 pp. Finally, learned thresholds improve accuracy by 1.5 pp over fixed $\tau = 0.5$, highlighting the value of adaptive decision boundaries.

C. Limitations

Sample size ($N = 18$) limits generalizability. The labeling strategy conflates fatigue with time-on-task effects. The logic layer provides interpretability without significant accuracy improvement over non-interpretable alternatives. Future work should validate on larger cohorts, explore temporal dynamics, and investigate online adaptation for low-fidelity subjects.

VII. Conclusion

We introduced a neuro-symbolic fatigue detection framework that learns interpretable oculomotor and hemodynamic concepts from eye-tracking and fNIRS and fuses them via differentiable approximate reasoning. Key findings are:

1) Competitive performance: 72.1% $\pm$ 12.3% LOSO-CV accuracy with interpretable decision pathways.
2) Ablations show calibration is most impactful (+5.2 pp), while fNIRS contributes (+1.2 pp) and Łukasiewicz operators outperform product (+0.9 pp).
3) Concept fidelity correlates strongly with accuracy ($r = 0.843$, $p < 10^{-4}$), enabling post-hoc reliability auditing.
4) Rule analysis identifies R_2 as the main discriminator (Cohen's $d = 1.15$) and reveals subject-specific patterns missed by black-box models.

Finally, the framework delivers transparent, expert-auditable fatigue monitoring alongside clearer indicators of prediction trustworthiness.

VIII. Acknowledgement

This research and Javier Fumanal-Idocin were supported by EU Horizon Europe under the Marie Skłodowska-Curie COFUND grant No 101081327 YUFE4Postdocs. This research is supported by UKRI BBSRC project EyeWarn (code: APP37953).

References

[1] K. Kakhi, H. Asgharnezhad, A. Khosravi, R. Alizadehsani, and U. R. Acharya, "A transfer learning-based approach for fatigue detection through the fusion of physiological signals," *IEEE Sensors Journal*, 2025.

[2] Y. Lei, P. Azizinezhad, M. Jamalifard, S. Manohar, M. Wlodarski, T. Foulsham, and J. Andreu-Perez, "Ocular metrics for fatigue assessment: A survey from physiology to machine learning," SSRN, 2024, article no. 5732154.

[3] Y. Albadawi, M. Takruri, and M. Awad, "A review of recent developments in driver drowsiness detection systems," *Sensors*, vol. 22, no. 5, p. 2069, 2022.

[4] Y. Lei, Y. Wang, F. Buchanan, M. Zhao, Y. Sugano, S. He, M. Khamis, and J. Ye, "Quantifying the impact of motion on 2d gaze estimation in real-world mobile interactions," *arXiv preprint arXiv:2502.10570*, 2025.

[5] Y. Lei, S. He, M. Khamis, and J. Ye, "An end-to-end review of gaze estimation and its interactive applications on handheld mobile devices," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–38, 2023.

[6] J. A. Caldwell, J. L. Caldwell, L. A. Thompson, and H. R. Lieberman, "Fatigue and its management in the workplace," *Neuroscience & Biobehavioral Reviews*, vol. 96, pp. 272–289, 2019.

[7] M. Lohani, B. R. Payne, and D. L. Strayer, "A review of psychophysiological measures to assess cognitive states in real-world driving," *Frontiers in human neuroscience*, vol. 13, p. 57, 2019.

[8] O. Ajayi, A. Kurien, K. Djouani, and L. Dieng, "A multimodal systematic review of drivers' fatigue detection methodologies, datasets, and models," *IEEE Access*, 2025.

[9] A. Kashevnik, S. Kovalenko, A. Mamonov, B. Hamoud, A. Bulygin, V. Kuznetsov, I. Shoshina, I. Brak, and G. Kiselev, "Intelligent human operator mental fatigue assessment method based on gaze movement monitoring," *Sensors*, vol. 24, no. 21, p. 6805, 2024.

[10] H. Ayaz, P. A. Shewokis, S. Bunce, K. Izzetoglu, B. Willems, and B. Onaral, "Optical brain monitoring for operator training and mental workload assessment," *Neuroimage*, vol. 59, no. 1, pp. 36–47, 2012.

[11] M. Causse, Z. Chua, V. Peysakhovich, N. Del Campo, and N. Matton, "Mental workload and neural efficiency quantified in the prefrontal cortex using fnirs," *Scientific reports*, vol. 7, no. 1, p. 5222, 2017.

[12] K. M. Hossain, M. A. Islam, S. Hossain, A. Nijholt, and M. A. R. Ahad, "Status of deep learning for eeg-based brain–computer interface applications," *Frontiers in computational neuroscience*, vol. 16, p. 1006763, 2023.

[13] G. Marra, S. Dumančić, R. Manhaeve, and L. De Raedt, "From statistical relational to neurosymbolic artificial intelligence: A survey," *Artificial Intelligence*, vol. 328, p. 104062, 2024.

[14] B. C. Colelough and W. Regli, "Neuro-symbolic ai in 2024: A systematic review," 2025.

[15] L. Serafini, S. Badreddine, I. Donadello, M. Spranger, F. Bianchi *et al.*, "Logic tensor networks: Theory and applications," in *Neuro-Symbolic Artificial Intelligence: The State of the Art*. IOS Press, 2021, pp. 370–394.

[16] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International conference on machine learning*. PMLR, 2020, pp. 5338–5348.

[17] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," *arXiv preprint arXiv:2205.15480*, 2022.

[18] Ü. Kaya, D. Akay, and S. Ş. Ayan, "Eeg-based emotion recognition in neuromarketing using fuzzy linguistic summarization," *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 8, pp. 4248–4259, 2024.

[19] J. Fumanal-Idocin and J. Andreu-Perez, "Ex-fuzzy: A library for symbolic explainable ai through fuzzy logic programming," *Neurocomputing*, vol. 599, p. 128048, 2024.

[20] B. P. Bhuyan, A. Ramdane-Cherif, R. Tomar, and T. Singh, "Neuro-symbolic artificial intelligence: a survey," *Neural Computing and Applications*, vol. 36, no. 21, pp. 12 809–12 844, 2024.

[21] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[23] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[25] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with computers*, vol. 21, no. 1-2, pp. 133–145, 2009.

[26] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.