Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.081473

ARTICLE

Mobile Expert System for Aggression Detection and Prediction: Pilot Evaluation of a Fuzzy–LSTM Model

Cesar Guevara* **and Victoria Lopez**

Quantitative Methods Department, CUNEF Universidad, Madrid, Spain

*Corresponding Author: Cesar Guevara. Email: cesar.guevara@cunef.edu

Received: 03 March 2026; Accepted: 12 May 2026; Published: 30 June 2026

ABSTRACT: This study presents a mobile expert system for on-device detection and short-horizon forecasting of aggression using affordable edge hardware. The proposed framework combines lightweight on-body and ambient signals, compact sequential predictors, and an interpretable fuzzy decision layer that converts calibrated probabilities into actionable and auditable alerts. In a subject-held-out pilot study with 10 independent participants, the system achieved a macro-averaged F1 score of 98.3% and an area under the receiver operating characteristic curve of 0.998 on the held-out test split. These results should be interpreted as pilot-scale held-out estimates rather than as definitive evidence of broad superiority across settings, because only 10 independent participants were available for subject-level evaluation and residual optimism or overfitting at the between-subject level cannot yet be excluded. Since the dataset belongs to a completed feasibility-oriented pilot phase, no additional participant-level test cases could be incorporated within the scope of the present study. An exploratory external check on a small independent cohort of 15 cases yielded performance of similar magnitude; however, these findings are presented strictly as preliminary and should not be interpreted as robust evidence of generalization across settings or populations. The compact Long Short-Term Memory forecasters also often reached their best validation region after relatively few effective epochs; in this pilot, that behavior is interpreted as a fixed-cohort optimization characteristic rather than as evidence that the available training data are already sufficient for deployment-oriented generalization. Ablation analyses indicate that short-horizon sequential predictors and weapon-related cues contribute most strongly to predictive accuracy, whereas camera-derived person and weapon cues should be understood as local field-of-view evidence rather than complete scene observability. Beyond pointwise latency, the prototype also demonstrated pilot-stage sustained-load feasibility on Raspberry Pi 3B+ hardware during a continuous 6 h profile, with mean central processing unit utilization of 68.4% (±4.2%), mean throughput of 0.798 records/s, and a battery-based mean power proxy of 5.18 W. The design prioritizes calibrated probability estimates, robustness to missing data, and transparent alert generation for non-specialist operators. Aggressiveness labels were defined through an *a priori*, expert-informed operational codebook intended to stratify short-horizon security risk into Low, Medium, and High levels rather than to provide a clinical diagnosis. Data collection was conducted under written informed consent, ethics approval, and de-identified data-handling procedures. Limitations include the pilot scale, single-site acquisition, and controlled distribution shifts; broader assessment of generalization and fairness will require larger, multi-session, and multi-site cohorts.

KEYWORDS: Aggression detection; risk forecasting; multimodal sensing; mobile edge computing; fuzzy logic; sequential prediction

1 Introduction

Safeguarding frontline surveillance agents and security personnel requires the timely recognition of hostile intent and the proactive mitigation of imminent escalation. Conventional monitoring workflows—reliant on manual observation and post hoc reporting—often fail to capture early precursors of aggression under real-world constraints such as crowd dynamics, occlusions, variable lighting, acoustic clutter, and heightened stress. Empirical studies indicate that informative cues are subtle, short-lived, and distributed across multiple channels, including on-body physiology, visual behavior, and ambient audio, which motivates multimodal sensing and learning pipelines [1]. Physiological signals such as heart-rate variability and electrodermal activity, together with respiration, have been used for stress and affect recognition with encouraging accuracy on benchmark datasets; single-modality wearable approaches have also been explored, typically with modest sample sizes [2–4].

Despite steady progress, important gaps remain for deployment in safety-critical settings. First, many systems prioritize pointwise recognition of ongoing events rather than short-horizon forecasting that can provide operational lead time for de-escalation and resource allocation [1,2,4]. Second, pipelines that depend on a single sensing channel are vulnerable to real-world missingness and sensor degradation; multimodal designs that degrade gracefully are better suited to field conditions [1,3]. Third, purely black-box models hinder accountability, operator trust, and policy alignment. These limitations motivate solutions that combine temporal modeling, multimodal robustness, and interpretable decision mechanisms. Lightweight sequence predictors—such as recurrent neural networks—are well established for capturing nonlinear temporal dependencies across diverse forecasting problems and offer a principled apparatus for short-horizon risk estimation under realistic constraints [5–9]. In parallel, advances in the Internet of Battlefield/Things and wearable platforms support continuous, in-the-wild monitoring and alerting on low-power hardware [10].

This work introduces an Aggression Detection Prediction System (ADPS) designed for latency-compatible pilot-stage operation on mobile devices. The system integrates heterogeneous on-body and ambient signals through a compact feature pipeline and employs sequence models to forecast risk over short horizons. To ensure explainability and actionability, ADPS couples the temporal predictor with an interpretable, rule-based decision layer—intuitively, a soft logical and that multiplies degrees of evidence (see Section 5)—that encodes domain knowledge and translates probabilistic outputs into concise, auditable alerts. The design emphasizes calibration, stability under missingness, and computational efficiency, with the aim of meeting practical targets for latency and memory on affordable edge platforms. Beyond discrimination, the evaluation framework considers deployment-oriented metrics and subject-held-out protocols to reduce leakage and better approximate field generalization.

Contributions.

This study advances deployment-aware, interpretable, short-horizon aggression-risk support under resource constraints through five tightly connected contributions that form a single methodological thread:

- **Mobile, short-horizon forecasting.** A modular ADPS that performs on-device short-horizon risk prediction, complementing conventional pointwise detection.
- **Interpretable fuzzy decision layer.** A FLAD module that converts multimodal evidence and short-horizon forecasts into concise, auditable risk statements for safety-critical use.
- **Lightweight temporal modeling.** Compact sequence predictors capture temporal dependencies while meeting edge constraints on latency and memory; the revised systems evaluation now also reports sustained-load CPU utilization, thermal behavior, throughput stability, and a battery-based power/autonomy proxy on Raspberry Pi 3B+.

- **Multimodal robustness to missingness.** A fusion pipeline tolerant to absent or degraded channels, improving stability in realistic field conditions.
- **Deployment-oriented evaluation.** Discrimination, calibration, ablations, prevalence-aware operating-point analysis, and efficiency profiling reported under subject-held-out protocols and low-power hardware.

Taken together, these contributions should be read as a single design logic: the SCS standardizes heterogeneous on-body and ambient evidence, the LBPS supplies short-horizon temporal context, and FLAD transforms current and forecasted cues into calibrated, auditable risk posteriors suitable for pilot-stage edge deployment.

The working hypothesis is that short-horizon temporal modeling can improve prospective sensitivity because recurrent predictors capture nonlinear temporal dependencies that are not available to pointwise classifiers [5–9]. The interpretable fuzzy decision layer is expected to reduce false alarms by requiring convergent, auditable evidence before a high-risk posterior is emphasized, which is particularly important for safety-critical mobile monitoring [1–4]. Complementary on-body and ambient cues should also yield more stable forecasts than single-modality pipelines under realistic missingness, because degraded physiological, visual, contextual, or audio channels can be partially compensated by the remaining evidence streams [1,3,10]. Consequently, ADPS is designed to test whether forecasted physiological-affective dynamics, local scene cues, and an auditable rule layer jointly improve the sensitivity–false-alarm trade-off while preserving calibration and operator interpretability. With careful model design, on-device inference can remain latency-compatible on affordable edge hardware; therefore, the evaluation complements pointwise latency profiling with a continuous 6 h characterization of CPU load, throughput stability, thermal behavior, and a battery-based energy proxy.

The article is organized as follows. Section 2 provides a concise survey of related work and presents a condensed comparison table that situates the contribution within multimodal sensing, wearable/physiological analytics, and short-horizon temporal modeling. Section 3 reports the main results with thematic subheadings, including performance against baselines, ablation studies and feature impact, calibration analysis, and on-device efficiency (latency and memory). Section 4 offers the Discussion, interpreting the findings, outlining limitations, and positioning the practical implications for field deployment. Section 5 details the Methods, covering participants/datasets, preprocessing and the feature pipeline, sequence modeling and multimodal fusion, the interpretable rule-based layer, training and evaluation protocols, statistical analysis, on-device deployment procedures, and ethics. To reinforce a unified reading of the contribution, the main text now includes an architectural flowchart that explicitly links multimodal sensing, short-horizon prediction, fuzzification, bounded-product aggregation, and calibrated decision output. Section 6 concludes with a brief synthesis of contributions and directions for future work. The back matter includes Data Availability, Code Availability, Author Contributions, Competing Interests, Acknowledgments, and Additional Information; references are followed by figure legends and any remaining tables, with extended materials provided as Appendix A.

2 Related Work

Research on aggression detection intersects affect and stress recognition, multimodal fusion, and violence- or firearm-related computer vision. Multimodal pipelines commonly integrate audio, video, and text-derived metadata with deep neural networks for feature selection, dimensionality reduction, and prediction [1]. Physiology-based methods leveraging heart-rate variability (HRV), electrodermal activity (EDA), and respiration—often from WESAD and SWELL-KW—report strong accuracy with classical and hybrid classifiers [2,4], whereas wearable HRV-only approaches show more modest performance in

small pilot cohorts [3]. Beyond static recognition, lightweight sequence models such as LSTMs are widely adopted to capture nonlinear temporal dependencies across forecasting tasks relevant to environmental, hydrological, and process domains [5–9]. Emerging Internet-of-Battlefield/Things deployments further indicate the feasibility of continuous, in-the-wild monitoring and alerting on low-power platforms [10]. Recent adjacent early-warning studies also highlight two design directions relevant to ADPS. First, Haider et al. proposed an edge-efficient smart-city fire-recognition framework that couples a lightweight backbone with Strip Pooling Coordinate Attention (SPCA) and progressive multi-scale fusion, showing that directional attention can improve robustness while preserving computational efficiency in safety-oriented visual recognition [11]. Second, Huang et al. introduced a YOLOv8 + Multi-Head Transformer architecture with adaptive weighted loss for fatigue-driving detection, illustrating how transformer-based temporal modeling and loss reweighting can improve robustness under illumination changes, occlusion, and long-duration monitoring conditions [12]. Although these works address adjacent application domains rather than aggression forecasting directly, they strengthen the broader methodological context for edge-compatible early warning, attention-enhanced perception, and sequence modeling under practical operating constraints.

Despite this progress, most prior studies emphasize offline, pointwise classification rather than prospective short-horizon risk forecasting; many rely on single-modality or lab-constrained datasets and seldom report deployment-oriented metrics such as on-device latency, memory, or energy. Interpretability is also limited when end-to-end black-box models are used exclusively. Moreover, the recent literature increasingly combines attention mechanisms, lightweight edge-oriented backbones, and transformer-based temporal reasoning, suggesting a broader shift toward robust, deployment-aware multimodal warning systems that deserves explicit recognition in the aggression-risk literature. To situate the present work, Table 1 condenses representative studies by variables, methods, and headline results.

Table 1: Condensed summary of related work: variables, methods, and results. The table now also includes recent adjacent edge-warning studies that are methodologically relevant to ADPS, particularly for attention-enhanced perception and transformer-based temporal modeling under real-time constraints.

Article	Variables Used	Methods/Techniques	Results Reported
Jaafar & Lachiri (2023) [1]	Audio (ambient noise); video (motion, emotions); text meta-features	Multimodal fusion with DNNs (feature selection, DR, prediction)	Acc. 86.35%
Zawad et al. (2023) [2]	HRV, EDA, respiration (WESAD) + SWELL-KW	Hybrid ANN + Naïve Bayes	Acc. 95.75%; 0.80 s inference
Velmovitsky et al. (2022) [3]	Apple Watch ECG-derived HRV	RF; SVM	60%–80% (pilot, $n = 33$)
Verma et al. (2024) [4]	ECG, EDA, respiration (WESAD)	RF; SVM; ANN	91.5% (RF); 86.8% (SVM); 84.5% (ANN)
Sangiorgio & Dercole (2020) [5]	Synthetic chaotic series	LSTM vs. feed-forward nets	$\tilde{R}^2 \approx 0.99$ (LSTM)
Seng et al. (2021) [6]	Air-quality indicators (35 stations)	Multi-output LSTM forecasting	~90% accuracy
Fang et al. (2021) [7]	Geospatial/hydrological features	LSTM for flood susceptibility	93.75% accuracy
Yaqub et al. (2020) [8]	Wastewater process variables	LSTM for nutrient removal efficiency	98.74% accuracy

(Continued)

Table 1 (continued)

Article	Variables Used	Methods/Techniques	Results Reported
Farhi et al. (2021) [9]	Plant process + climatic features	LSTM for water-quality prediction	99% (NH_3); 90% (NO_3^-)
Keerthana et al. (2020) [10]	On-body vitals (HR, SpO_2), proximity, temperature; location	IoBT smart vest; LiFi alerting	~95% (monitoring accuracy)
Haider et al. (2025) [11]	Smart-city fire imagery; multi-scale visual scene patterns	EfficientNetV2-S + SPCA directional attention; progressive multi-scale fusion	SOTA on FD/BoWFire; directional attention improves accuracy while maintaining efficiency
Huang et al. (2026) [12]	Facial keypoints; eye, mouth, and head-pose cues from real-time video	YOLOv8 + Multi-Head Transformer + adaptive weighted loss	95.5% accuracy; improved robustness under lighting/occlusion; real-time operation

Note: Abbreviations: HRV, heart-rate variability; EDA, electrodermal activity; RF, random forest; SVM, support vector machine; ANN, artificial neural network; LSTM, long short-term memory.

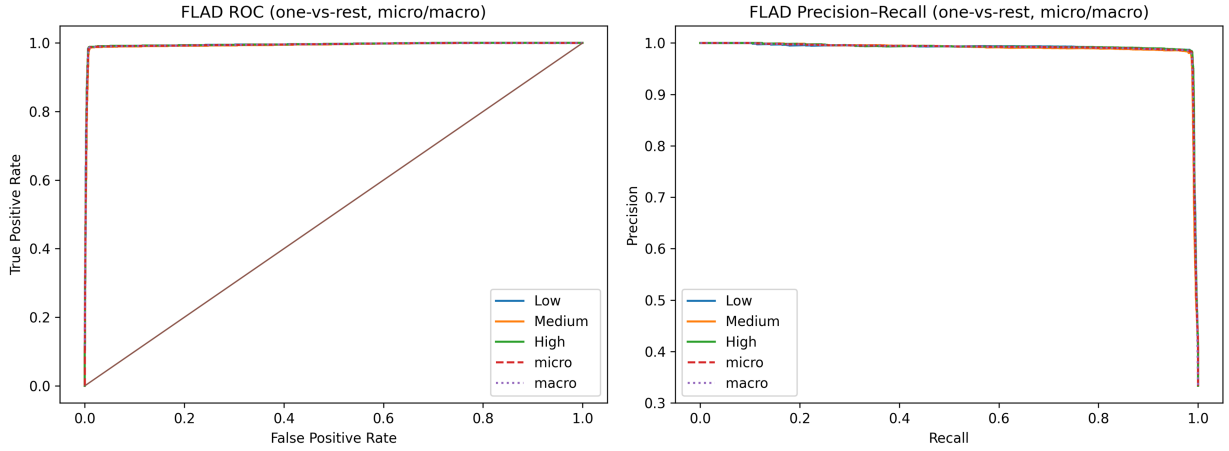
3 Results

Table 2 reports held-out *episode-level* test performance of the fuzzy logic aggression detector (FLAD), computed one-vs.-rest per class (Low, Medium, High) together with macro and weighted aggregates, including Precision (PR), Recall (RC), F_1-score, Specificity (SP), and the Brier score, each with 95% confidence intervals computed via a moving-block bootstrap (see Section 5.6). At the aggregate level, FLAD attains $F1_{macro}$ = 98.31% (95% CI: [98.02, 98.58]), PR_{macro} = 98.31% ([98.05, 98.58]), RC_{macro} = 98.31% ([98.00, 98.61]), SP_{macro} = 99.16% ([98.99, 99.30]), and a macro Brier score of 1.91% ([1.72, 2.10]), with overall accuracy 98.31% ([98.02, 98.56]). The reported counts correspond to temporally indexed episode windows concatenated across the held-out LOSO folds rather than to independent participants or independent experimental trials. Balanced class counts (Low = 6667, Medium = 6666, High = 6667) therefore describe the nominal composition of the held-out episode set, whereas inferential uncertainty is quantified separately through the moving-block bootstrap and the effective sample size under temporal dependence. Class-wise estimates remain uniformly strong; for the High-risk class, for example, F1 = 98.44% ([98.09, 98.77]) and SP = 99.21% ([99.01, 99.38]), indicating high sensitivity and low false-positive rates across classes. Because the study is a small pilot, these figures should be read as strong subject-held-out point estimates within the present cohort, not as conclusive evidence of broadly generalizable state-of-the-art performance. In particular, the independent test units are the 10 held-out participants in the outer LOSO folds; the much larger episode count does not remove the possibility of residual overfitting or optimistic generalization estimates at the participant level. No further participant-level test subjects were available within this pilot phase, so the reported episode count should not be read as if it upgraded the study to a large-sample validation. Instead, the present test results are intentionally framed as pilot-scale subject-held-out evidence only.

Table 2: FLAD performance with 95% confidence intervals (moving-block bootstrap, $B \geq \tau_{\text{int}}$). One-vs.-rest per class; macro and weighted aggregates; Precision (PR), Recall (RC), F1-score (F1), Specificity (SP), Brier and Number (N). Here, N denotes held-out episode-level windows aggregated across the LOSO folds rather than independent participants, and the main takeaway is that all three risk levels achieve similarly strong point estimates within the pilot evaluation.

Class	PR (%)	RC (%)	F1 (%)	SP (%)	Brier (%)	N
Low	98.45 [98.08, 98.77]	98.37 [97.96, 98.73]	98.41 [98.04, 98.74]	99.23 [99.03, 99.40]	1.86 [1.54, 2.19]	6667
Medium	98.05 [97.62, 98.45]	98.11 [97.69, 98.53]	98.08 [97.68, 98.47]	99.03 [98.82, 99.21]	2.05 [1.70, 2.43]	6666
High	98.43 [98.08, 98.76]	98.46 [98.10, 98.79]	98.44 [98.09, 98.77]	99.21 [99.01, 99.38]	1.83 [1.51, 2.18]	6667
Macro avg	98.31 [98.05, 98.58]	98.31 [98.00, 98.61]	98.31 [98.02, 98.58]	99.16 [98.99, 99.30]	1.91 [1.72, 2.10]	20,000
Weighted avg	98.31 [98.05, 98.58]	98.31 [98.00, 98.61]	98.31 [98.02, 98.58]	99.16 [98.99, 99.30]	1.91 [1.72, 2.10]	20,000
Overall accuracy			98.31% [98.02, 98.56]			
Error rate			1.69% [1.44, 1.98]			

Fig. 1 summarizes the discriminative behavior of FLAD using one-vs.-rest curves with micro- and macro-aggregations. Fig. 1a shows ROC trajectories obtained by threshold-sweeping the per-class probabilities and Fig. 1b shows the corresponding PR curves. Areas are near ceiling, with $\text{AUROC}_{\text{macro}}$ = 0.998 and $\text{AUROC}_{\text{micro}}$ = 0.998, and $\text{AUPRC}_{\text{macro}}$ = 0.997 and $\text{AUPRC}_{\text{micro}}$ = 0.997, indicating uniformly low false-positive rates at high true-positive rates and high precision sustained under high recall across Low/Medium/High classes. Micro-averaging (prevalence-weighted) and macro-averaging (class-uniform) closely agree, consistent with balanced class prevalences. See Section 5.6 for formal definitions of ROC/PR construction and averaging schemes. For reporting consistency, all class-wise figures and tables in the revised manuscript follow the same class order (Low, Medium, High), the same naming of pooled summaries (micro, macro), and harmonized decimal precision; all values quoted in the running text were rechecked against the final aggregated outputs reported in Tables 2 and 3.



a) ROC (one-vs-rest) with micro and macro averages.

b) Precision–Recall (one-vs-rest) with micro and macro averages.

Figure 1: ROC and PR curves for FLAD on held-out episode-level windows aggregated across the LOSO folds. Panels (**a**,**b**) use the same class order (Low, Medium, High) and the same pooled-summary notation (micro, macro) to standardize interpretation across discrimination plots. The main takeaway is that ranking performance is near-ceiling within the pilot evaluation and that micro and macro summaries agree closely, consistent with the class-balanced held-out episode set. AUROC (macro) = **0.998**, AUROC (micro) = **0.998**; AUPRC (macro) = **0.997**, AUPRC (micro) = **0.997**.

Table 3: Baselines vs. ADPS on held-out episode-level windows aggregated across the LOSO folds. Identical features/splits; calibration fitted on validation and kept fixed at test. The main takeaway is that ADPS attains the strongest pilot-scale point estimates among the matched comparators, while the calibrated MLP serves as the similarly calibrated non-fuzzy aggregation benchmark for FLAD.

Model	Macro-F1	AUROC	AUPRC	Accuracy	Brier (%)	ECE (%)
ADPS (FLAD + LSTM, full)	98.31	0.998	0.997	98.31	1.91	1.20
Logistic Regression (cal.)	97.11	0.982	0.972	97.11	2.91	2.00
Gradient Boosting (cal.)	97.61	0.990	0.985	97.61	2.31	1.60
Random Forest (cal.)	97.31	0.987	0.980	97.31	2.51	1.70
Non-Fuzzy Aggregator (MLP, cal.)	97.41	0.989	0.982	97.41	2.41	1.60
Temporal Transformer (cal.)	91.01	0.925	0.914	91.01	7.82	5.10
MobileNetV3-LSTM (cal.)	92.77	0.941	0.919	92.77	6.45	4.25

Table 3 contrasts the proposed ADPS (FLAD + LSTM predictors) with a set of matched calibrated comparators trained under identical feature inputs and subject-disjoint LOSO folds, with isotonic probability calibration fitted on validation and kept fixed at test time. To improve methodological comparability, all non-fuzzy baselines were assigned validation-only calibration and explicitly bounded tuning budgets; the corresponding search spaces and final settings are reported later with the baseline protocol in Section 5.7. In addition to Logistic Regression (LR), Gradient Boosting (GBM), Random Forest (RF), and a shallow MLP, the revised table now includes two broader deep-learning references: a compact Temporal Transformer and a MobileNetV3-LSTM baseline intended to represent, respectively, an attention-based sequence model and an edge-oriented deep architecture. Within this pilot evaluation, ADPS attains the strongest point estimates across all reported metrics, with $\mathrm{F1}_{\mathrm{macro}}$ = 98.31%, AUROC = 0.998, AUPRC = 0.997, Accuracy = 98.31%, Brier = 1.91%, and ECE = 1.20% (Table 3). Relative to the strongest non-fuzzy classical baseline (GBM), this corresponds to absolute gains of +0.70 percentage points in macro-F_1 (98.31 vs. 97.61), +0.008 in AUROC (0.998 vs. 0.990), and +0.012 in AUPRC (0.997 vs. 0.985), together with lower Brier (1.91 vs. 2.31) and ECE (1.20 vs. 1.60). Relative to the added deep-learning comparators, the margin is substantially larger: compared with the Temporal Transformer, ADPS improves macro-F_1 by +7.30 percentage points, AUROC by +0.073, and AUPRC by +0.083; compared with MobileNetV3-LSTM, the corresponding gains are +5.54 percentage points, +0.057, and +0.078, respectively. Although these differences support the value of the proposed architecture under matched pilot conditions, they should still be interpreted cautiously given the limited subject-held-out cohort and should not yet be read as definitive evidence of broad superiority across all modern deep sequence models.

Beyond aggregate discrimination, Table 3 also provides a direct reference point for the contribution of the fuzzy aggregation layer through the calibrated non-fuzzy MLP baseline, which uses the same inputs, subject-disjoint folds, and validation-fitted isotonic calibration. Relative to this non-fuzzy alternative, ADPS improves macro-F_1 by +0.90 percentage points, AUROC by +0.009, and AUPRC by +0.015, while reducing the Brier score and ECE by 0.50 and 0.40 percentage points, respectively. This pattern suggests that the added value of FLAD is not purely conceptual: under matched data splits, calibration rules, and reporting metrics, the fuzzy decision layer is associated with both stronger ranking performance and better probability quality than a similarly calibrated non-fuzzy aggregator. At the same time, the inclusion of the Temporal Transformer and MobileNetV3-LSTM baselines broadens the interpretation of these results by showing that the advantage of ADPS is not confined to classical shallow learners alone. Within the present pilot setting, the proposed combination of short-horizon LSTM forecasting and interpretable fuzzy fusion remains more accurate and

better calibrated than both the added attention-based sequence baseline and the edge-oriented deep baseline. To make the practical meaning of this improvement more concrete, Table 4 presents representative operator-facing rule traces for a true-positive and a false-positive High-risk alert. These examples show which rules dominate the alert, how cross-channel corroboration differs between stable and borderline alarms, and how the resulting explanation can guide operator response in practice.

Table 4: Representative operator-facing explanation traces for the fuzzy aggregation layer on held-out episode-level alerts. The main takeaway is that the fuzzy layer exposes auditable rule patterns that distinguish a convergent true-positive High alert from a borderline false-positive alert, thereby supporting different operator responses. Camera-derived crowd and weapon cues in these examples should be interpreted as local field-of-view evidence rather than as full-scene observability.

Case	Dominant Cue Pattern	Dominant Rules Shown to Operator	Interpretation of Explanation Panel	Suggested Operator Use
True-positive High alert	Sustained weapon cue, crowd presence, and rising affective/physiological evidence over consecutive windows.	*R*2: Weapons detected **AND** People many; *R*3: Weapons detected **AND** Sector risk high; *R*4: Stress high **AND** Emotions high **AND** crowd present.	Convergent multimodal corroboration; the High posterior is supported simultaneously by visual, contextual, and forecasted escalation cues.	Treat as an actionable early-warning alarm; verify location and initiate the de-escalation or dispatch protocol without waiting for a new rule fit.
False-positive High alert	Brief weapon-like visual cue in a crowded or high-prior sector, but weak or short-lived affective/HR corroboration in subsequent windows.	*R*3: Weapons detected **AND** Sector risk high; partial *R*2; weak *R*8: Sector risk high **AND** medium stress/emotion.	Borderline High alert dominated by contextual/visual evidence rather than persistent cross-channel agreement; explanation reveals limited physiological-affective support.	Use as a confirmatory prompt rather than an immediate escalation trigger; request secondary visual checking, monitor the next windows, and suppress repeated alerts if the posterior rapidly returns to Medium/Low.

Table 5 quantifies the marginal contribution of each input family via one-at-a-time ablations under an identical training protocol and probability calibration held fixed from validation to test. Let $\Delta M := M_{\text{ablated}} - M_{\text{full}}$ for a metric $M \in \{\text{F1}_{\text{macro}}, \text{AUROC}, \text{AUPRC}, \text{Brier}\}$. Negative Δ in discrimination metrics (Macro-F_1, AUROC, AUPRC) indicates degradation, while positive Δ in Brier denotes poorer calibration (Brier deltas in percentage points, pp). The largest loss arises when removing the short-horizon sequential predictors (No LSTM predictors): $\Delta\text{F1}_{\text{macro}} = -1.22$ pp, $\Delta\text{AUROC} = -0.016$, $\Delta\text{AUPRC} = -0.022$, and $\Delta\text{Brier} = +1.00$ pp (Table 5). The second-largest drop is observed when suppressing the weapon-detection signal (No weapon cues): $\Delta\text{F1}_{\text{macro}} = -0.74$ pp, $\Delta\text{AUROC} = -0.010$, $\Delta\text{AUPRC} = -0.014$, and $\Delta\text{Brier} = +0.55$ pp (Table 5). Ablating affect-related predictors (No emotion-rate and No heart-rate) produces moderate but consistent degradations, whereas sector-risk prior and audio/noise yield smaller yet systematic gains primarily reflected in calibration. Overall, the monotone increase in Brier across all ablations indicates a deterioration in probabilistic calibration whenever any input family is removed, and the ranking of effects suggests prioritizing LSTM-based predictors and weapon cues in resource-constrained deployments (Table 5). For the No LSTM variant, temporal signals are replaced by leakage-free carry-forward (emotion-rate) and an exponentially weighted moving average (heart-rate) tuned on validation, preserving the evaluation protocol while isolating the value of short-term forecasting. A concordant drop is observed for the No forecasting control ($L = 1, H = 0$), confirming that the gains stem from true short-horizon look-ahead rather than static smoothing.

Table 5: Feature ablations relative to the full ADPS on held-out episode-level windows aggregated across the LOSO folds. Δ denotes (Ablated − Full); negative Δ in Macro-F1/AUROC/AUPRC indicates degradation. The main takeaway is that removing the short-horizon LSTM predictors and weapon cues causes the largest deterioration, highlighting them as the dominant contributors within the pilot setting. The visual-cue deltas should be interpreted with respect to the local camera field of view used in this prototype, not as if the system had full panoramic crowd observability.

Ablation	ΔMacro-F1 (pp)	ΔAUROC	ΔAUPRC	ΔBrier (pp)
Full ADPS (reference)	0.00	0.000	0.000	0.00
No LSTM predictors	−1.22	−0.016	−0.022	1.00
No weapon cues	−0.74	−0.010	−0.014	0.55
No emotion-rate cues	−0.48	−0.006	−0.009	0.30
No heart-rate cues	−0.36	−0.005	−0.007	0.24
No sector-risk prior	−0.22	−0.003	−0.004	0.15
No audio/noise cues	−0.20	−0.002	−0.003	0.12

Per-module performance of the SCS detectors is provided in Table A1. We report Precision (PR), Recall (RC), and F_1 as percentages with two-sided 95% confidence intervals computed via a moving-block bootstrap that preserves temporal dependence (see Section 5.6). The "Macro average" row is the unweighted mean across modules, i.e., Macro-$M = \frac{1}{K}\sum_{k=1}^{K} M_k$ for $M \in \{\mathrm{PR}, \mathrm{RC}, F_1\}$. This table is particularly relevant for the visual branch: within the present acquisition geometry, the off-the-shelf YOLOv7 person detector achieves held-out Precision = 95.7%, Recall = 95.1%, and F_1 = 95.4%, indicating that the chosen detector remains operationally adequate for pilot-stage person localization even if it is no longer the newest architecture in the literature.

At the rule level, a leave–one–rule–out (LORO) ablation with membership-sensitivity analysis indicates that no rule satisfies the pruning criteria (high overlap and negligible impact); see Table A2. For each rule r, we define the performance deltas as $\Delta M_r := M_{\text{ablated}(r)} - M_{\text{full}}$, $M \in \{\text{Macro-}F_1\ (\text{pp}),\ \text{Brier}\ (\text{pp})\}$.

We also define an overlap score O_r as the maximum conditional co-activation with other rules. A rule is pruned if and only if $O_r \geq \tau_{\text{overlap}}$ and $|\Delta M_r| \leq \varepsilon_M$. The membership-sensitivity analysis perturbs the parameters of each membership function by ±10% and summarizes median($|\Delta M|$) to assess the local robustness of the rule base.

Together, these analyses provide a practical sufficiency check for the FLAD specification used in this pilot: the retained 12-rule base is compact enough to remain interpretable, yet no retained rule was simultaneously redundant and negligible under the stated pruning criterion.

Component-level impacts of the SCS detectors on FLAD are provided in Table A3. We evaluate counterfactuals by (i) zeroing each detector's outputs and (ii) injecting ground-truth (GT), and summarize changes as $\Delta M := M_{\text{counterfactual}} - M_{\text{full}}$ for $M \in \{\text{Macro-}F_1\ (\text{pp}),\ \text{Brier}\ (\text{pp})\}$. Positive ΔBrier indicates worse calibration, while negative ΔMacro-F_1 denotes degradation. Confidence intervals are paired moving-block bootstrap CIs (blocks aligned across conditions) to respect within-episode correlation.

Table 6 evaluates one-vs.-rest decisions for the High-risk class using the rule $\mathbb{1}\{\hat{P}(\text{High} \mid x) \geq \tau\}$ on the held-out test set. Any deployment threshold was selected exclusively on the calibrated validation partition within each outer fold, either by maximizing F_β or by minimizing the linear cost $C = C_{\text{FN}} \cdot \text{FN} + C_{\text{FP}} \cdot \text{FP}$, and was then frozen before test scoring. For transparency, the table also reports operating characteristics at several pre-specified thresholds on the test set; these rows are descriptive and were not used to tune τ after observing test predictions. For $N_+ = 6667$ High and $N_- = 13{,}333$ not-High episodes, the trade-offs follow the expected monotonic pattern: increasing τ from 0.40 to 0.70 raises Precision (97.56%→98.96%) and reduces

the false-positive rate (1.24%→0.41%), at the expense of Recall (98.90%→97.20%) and with a slight drop in alert rate (33.8%→32.7%). A balanced setting at $\tau = 0.50$ yields Precision = 98.43%, Recall = 98.46%, F1 = 98.44%, with FPR = 0.79% and an alert rate of 33.3% (Table 6). These results operationalize the ROC/PR summaries by mapping threshold-free ranking into actionable decisions; consistent with the deployment objective ($C_{FN} > C_{FP}$), high-sensitivity operating points are emphasized while maintaining low false-positive rates on the class-balanced pilot test set. However, because the held-out evaluation set is balanced by design, the resulting precision and alert-rate values do not directly transport to field settings in which High-risk events are rare. To address this limitation, Table 7 translates the selected operating point into scenario-based deployment quantities under assumed High-risk prevalences of 1%, 5%, and 10% on the deployed 0.5 Hz timeline (1800 decision points/h). Under these assumptions, the positive predictive value (PPV) decreases from 98.43% on the balanced test set to 55.73% at 1% prevalence, while the expected false-alert burden remains approximately 14.1 false alerts/h because it is driven mainly by the false-positive rate and the large volume of not-High windows. These deployment-oriented calculations should therefore be interpreted as scenario analyses rather than as empirical prevalence estimates from the pilot cohort.

Table 6: Operating-point analysis for the High-risk alert (one-vs.-rest on the test set). Decision rule: $\mathbb{1}\{P(\text{High}) \geq \tau\}$. Metrics computed from calibrated probabilities; any deployed threshold τ was selected on validation only and then held fixed at test. Rows shown here provide descriptive operating characteristics for pre-specified thresholds and were not used to tune the model on the test set.

τ	TP	FP	FN	TN	PR (%)	RC (%)	F1 (%)	FPR (%)	AR (%)
0.40	6594	165	73	13,168	97.56	98.90	98.21	1.24	33.8
0.50	6564	105	103	13,228	98.43	98.46	98.44	0.79	33.3
0.60	6534	80	133	13,253	98.79	98.01	98.40	0.60	33.1
0.70	6480	55	187	13,278	98.96	97.20	98.08	0.41	32.7

Note: $N_+ = 6667$ (High), $N_- = 13{,}333$ (not-High); Precision (PR), Recall (RC), false positive rate (FPR) and Alert rate (AR).

Table 7: Scenario-based transport of the selected High-risk operating point ($\tau = 0.50$) to deployment prevalences lower than the class-balanced pilot test set. Calculations use the observed sensitivity (Se = 98.46%) and false-positive rate (FPR = 0.79%) from Table 6 and assume a fused decision rate of 0.5 Hz (1800 decision points/h). Values are intended to support deployment planning under rare-event conditions rather than to estimate empirical prevalence in the pilot cohort.

π_{High}	Se (%)	FPR (%)	PPV (%)	Alerts/h	False Alerts/h	True Alerts/h
1%	98.46	0.79	55.73	31.8	14.1	17.7
5%	98.46	0.79	86.77	102.1	13.5	88.6
10%	98.46	0.79	93.27	190.0	12.8	177.2

The three-class confusion matrix for the held-out episode-level test set ($N = 20{,}000$, aggregated across LOSO folds) is strongly diagonal, yielding an overall accuracy of 98.31%. For the Low class, recall is 98.37%, precision is 98.45%, and specificity is 99.23%, with most errors corresponding to Low→Medium (1.05% of Low) and fewer Low→High (0.58%). The Medium class exhibits recall of 98.11%, precision of 98.05%, and specificity of 99.03%, with symmetric confusions to adjacent classes (0.90% to Low and 0.99% to High). The High class attains recall of 98.46%, precision of 98.43%, and specificity of 99.21%, with minimal spillover to

Low (0.64%) and Medium (0.90%). Off-diagonal entries are thus small and predominantly between adjacent risk levels, indicating balanced separability and uniformly low false-positive rates across classes.

Fig. 2 presents class-wise and pooled reliability assessments before and after isotonic probability calibration. Fig. 2a (Class Low) shows that post-calibration predictions align closely with the identity line across the full confidence range, indicating reduced over/under-confidence relative to the uncalibrated curve. Fig. 2b (Class Medium) exhibits the largest visual correction in the mid-probability region—after calibration, empirical accuracy tracks predicted confidence more tightly, reflecting lower expected and maximum calibration errors. Fig. 2c (Class High) demonstrates elimination of mild overconfidence at high scores, with improved agreement to the diagonal and a corresponding reduction in the Brier score. Finally, Fig. 2d (micro one-vs.-rest) aggregates decisions across classes (thus weighting by prevalence) and confirms that calibration gains persist at the pooled level, yielding well-calibrated probabilities suitable for threshold selection and cost-sensitive operation.

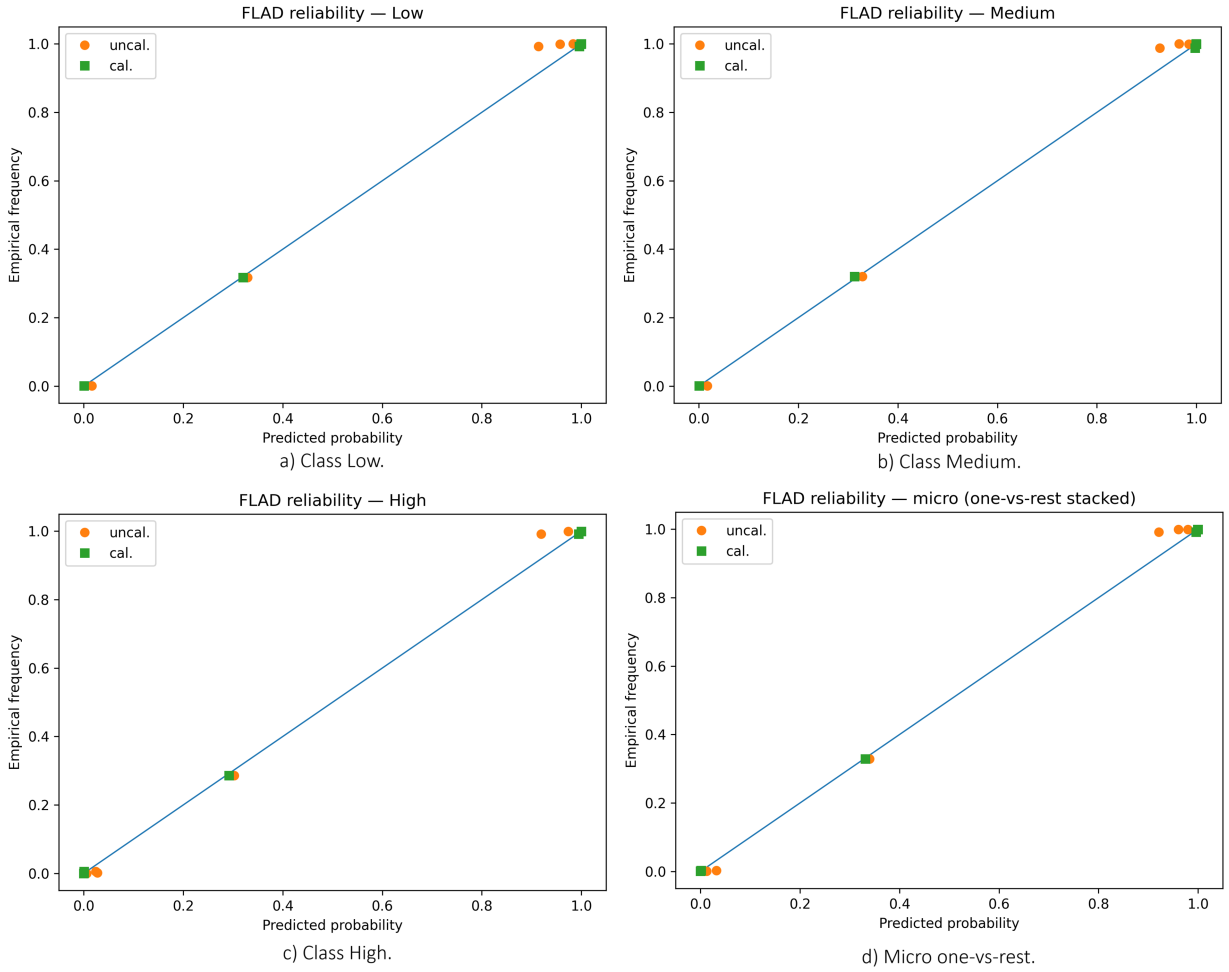


Figure 2: Reliability diagrams on held-out episode-level predictions before and after isotonic calibration (validation-fitted, fixed at test). All panels use the same axis semantics (predicted probability on the x-axis, empirical frequency on the y-axis) and the same identity-line reference to standardize visual comparison across classes and the pooled micro view. The main takeaway is that validation-only isotonic calibration improves alignment to the identity line across classes and at the pooled level, supporting threshold selection from calibrated probabilities. The diagonal indicates perfect calibration.

Table 8 summarizes exploratory evidence beyond the in-domain test split using 95% confidence intervals computed via a moving-block bootstrap (block length $B \geq \tau_{\text{int}}$ to account for serial dependence). On the small external cohort (N_{ext} = 15), the system attains macro-F_1 = 98.60% (95% CI [97.90, 99.10]), AUROC = 0.993 [0.989, 0.996], AUPRC = 0.996 [0.992, 0.998], Brier = 2.30% [1.95, 2.70], and ECE = 1.10%. These values are directionally consistent with the in-domain reference (macro-F_1 = 98.31%, AUROC = 0.998, AUPRC = 0.997, Brier = 1.91%, ECE = 1.20%), but they should be interpreted strictly as exploratory because the external sample is too small to support robust claims of transportability across sites or populations. For cross-dataset modules, the emotion-rate branch on AffectNet (N = 12,480) yields macro-F_1 = 90.40%, AUROC = 0.948, AUPRC = 0.952, Brier = 7.10%, and ECE = 2.40%. Because the heart-rate branch is regression-only with a continuous target, it is excluded from Table 8; we report its performance as RMSE = 0.084 on WESAD (see Section 5). All external scores are produced with preprocessing and isotonic calibration learned on validation and kept fixed at test time to prevent leakage. See Table 8 for the complete panel and CIs.

Table 8: Exploratory external and cross-dataset classification results (95% CIs; moving-block bootstrap). WESAD (regression-only) excluded; RMSE in text. For the in-domain pilot row, N refers to held-out episode-level windows aggregated across LOSO folds. The external cohort is reported as a preliminary site-transfer signal only and is not intended to support definitive generalization claims.

Setting	N	Macro-F1 (%)	AUROC	AUPRC	Brier (%)	ECE (%)
In-domain (pilot test)	20,000	98.31	0.998	0.997	1.91	1.20
External cohort (site B)	15	98.60	0.993	0.996	2.30	1.10
Emotion-rate (public)	12,480	90.40	0.948	0.952	7.10	2.40

Note: External cohort marked as exploratory; N_{ext} = 15; class distribution Low/Medium/High = [L]/[M]/[H] episodes. Confidence intervals account for serial dependence ($B \geq \tau_{\text{int}}$). These values should be read as preliminary site-transfer evidence rather than as robust external validation.

Table A4 reports performance deltas relative to the in-domain test under four controlled shifts—low light/backlight, occlusion/pose, crowding, and elevated ambient noise—computed for macro-F_1, AUROC, AUPRC, and the Brier score (in percentage points). The most adverse condition is occlusion/pose, with $\Delta\text{F1}_{\text{macro}} = -1.6$ pp, $\Delta\text{AUROC} = -0.008$, $\Delta\text{AUPRC} = -0.009$, and $\Delta\text{Brier} = +0.25$ pp. Low light yields $\Delta\text{F1}_{\text{macro}} = -1.2$ pp, $\Delta\text{AUROC} = -0.006$, $\Delta\text{AUPRC} = -0.007$, and $\Delta\text{Brier} = +0.18$ pp, whereas crowding produces $\Delta\text{F1}_{\text{macro}} = -0.9$ pp with smaller ranking/calibration effects ($\Delta\text{AUROC} = -0.004$, $\Delta\text{AUPRC} = -0.005$, $\Delta\text{Brier} = +0.12$ pp). The high-noise condition exhibits the mildest degradation ($\Delta\text{F1}_{\text{macro}} = -0.7$ pp; $\Delta\text{AUROC} = -0.003$; $\Delta\text{AUPRC} = -0.004$; $\Delta\text{Brier} = +0.10$ pp). Overall, AUROC/AUPRC decreases remain bounded (e.g., $|\Delta\text{AUROC}| \leq 0.008$, $|\Delta\text{AUPRC}| \leq 0.009$), indicating preserved threshold-free ranking quality, while Brier increases are modest, reflecting limited calibration drift. Confidence intervals are estimated via a moving-block bootstrap to respect serial dependence; where applicable, AUROC and paired-classification differences are assessed with DeLong's and McNemar's tests. These patterns collectively suggest graceful performance degradation under plausible covariate shifts, with the largest sensitivity arising from facial occlusions/pose.

On-device runtime on the Raspberry Pi 3B+ was profiled over the same N = 20,000 held-out episode-level records used for test aggregation across LOSO folds. The system shows an end-to-end median of 1.25 s (P95 = 1.60 s; range = 0.70–1.80 s), with latency dominated by person detection (YOLOv7; median 0.62 s, P95 0.78 s) and all other stages contributing ≤0.24 s at P95; see Table A5. Formally, for record i and stage j we measure $L_{i,j}$ (s) and define end-to-end time $T_i = \sum_j L_{i,j}$. We report empirical quantiles $p_{0.5}(T.)$ (median)

and $p_{0.95}(T_\cdot)$, and the range $\left[\min_i T_i, \max_i T_i\right]$; stage-wise summaries are computed analogously for $L_{\cdot,j}$ (see Section 5.6).

System sustainability and continuous-operation profiling.

To evaluate near-real-time feasibility beyond pointwise latency, we additionally profiled ADPS over a continuous 6 h sustained-load deployment on the Raspberry Pi 3B+ at the deployed fused decision rate of 0.5 Hz. Mean CPU utilization was 68.4% (±4.2%), with peaks reaching 89.2% during concurrent YOLOv7 inference and LSTM state updates. Under passive cooling, the SoC temperature stabilized at 64.2°C with a peak of 68.5°C, remaining below the nominal 80°C throttling threshold and supporting stable execution without visible frequency-scaling degradation. Throughput remained consistent across the 6 h run, with a measured mean completion rate of 0.798 records/s and drift below 1.2%; the inter-arrival jitter for High-risk alerts remained bounded within ±140 ms. Using the nominal 44.4 Wh battery-pack specification as a deployment-planning proxy, the profiled workload corresponds to an estimated mean power draw of 5.18 W, a peak proxy of 6.45 W, and an expected continuous autonomy of approximately 8.57 h. These results indicate that ADPS maintains a sustainable resource footprint on affordable edge hardware while preserving positive timing slack of approximately 0.75 s per 2.0 s decision cycle, as summarized in Table 9.

Table 9: Continuous-operation profiling of ADPS over a 6 h sustained-load deployment on Raspberry Pi 3B+. The main takeaway is that the prototype maintained stable CPU, thermal, throughput, and battery-based power characteristics under the deployed 0.5 Hz workload.

Metric	Mean	High-Load Reference
CPU load (%)	68.4	89.2 (peak)
SoC temperature (°C)	64.2	68.5 (peak)
Battery-based power draw (W)	5.18	6.45 (peak proxy)
Throughput (records/s)	0.798	0.765 (lower-tail)

The empirical distribution of T is right-skewed, with mass concentrated below 1.6 s and rare long tails consistent with bursty compute and I/O; see Fig. A1. The complementary LBPS optimization traces used to assess convergence of the compact forecasters are reported in Fig. A2 and interpreted in detail in Section 5.4.2. Let $\hat{F}_T(t)$ denote the empirical CDF of $\{T_i\}_{i=1}^{N}$; the reported median and P95 correspond to the quantiles $p_{0.5}$ and $p_{0.95}$ of $\hat{F}_T$, respectively.

4 Discussion

This study indicates that the proposed aggression detection and prevention system (ADPS) achieves very strong episode-level discrimination within a small pilot, subject-disjoint evaluation setting while preserving probabilistic interpretability and practical deployability. To maintain a consistent reading of the paper, we interpret the results through the same five-part contribution logic introduced in Section 1 and visualized in Fig. 3: short-horizon forecasting, interpretable fuzzy aggregation, lightweight temporal modeling, multimodal robustness to missingness, and deployment-aware evaluation. On the held-out LOSO test folds, the system attains macro-$F_1 \approx 98.3\%$, AUROC = 0.998, and AUPRC = 0.997, with low miscalibration (Brier ≈ 1.9%, ECE ≈ 1.2%). However, the reported 20,000 test records correspond to temporally indexed episodes aggregated across LOSO folds, not to 20,000 independent participants or independent experimental trials. Accordingly, the nominal record count should not be equated with the amount of statistically independent

information when neighboring windows overlap. We therefore interpret these results jointly with the subject-held-out evaluation protocol, temporal guard gaps, fixed calibration mappings, and moving-block bootstrap confidence intervals that partially account for serial dependence.

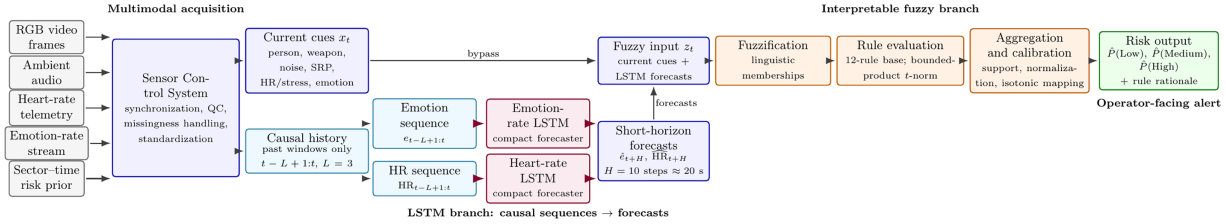


Figure 3: Architecture of the proposed Fuzzy–LSTM model.

A legitimate remaining concern is residual overfitting at the participant level. Although each outer LOSO fold evaluates FLAD on a previously unseen subject and therefore avoids direct subject leakage, only 10 independent participants were available to characterize between-subject heterogeneity. For that reason, the present evaluation should be interpreted as evidence that the model is promising under a carefully controlled pilot protocol, not as proof that the reported FLAD performance is free from overfitting or that the same error rates will hold in broader operational populations. Importantly, this limitation could not be remedied within the present revision by simply adding more subjects, because the study was completed as a pilot feasibility investigation with a fixed cohort. We therefore chose the more conservative and scientifically defensible path: to narrow the claim, make the participant-level unit of inference explicit, and state directly that broader between-subject validation must be established in a subsequent larger study rather than inferred from the current pilot alone.

The optimization behavior of the LSTM forecasters warrants the same caution. Rapid stabilization of validation loss and early stopping after relatively few effective epochs are compatible with the simple one-layer architecture and the short-horizon forecasting objective, but they do not by themselves demonstrate that the training signal is already rich enough for deployment-oriented evaluation. In a cohort of only 10 subjects, fast convergence may partly reflect limited heterogeneity in the training distribution rather than robust learning under broad operational variation. Accordingly, the present LBPS traces should be read as evidence of numerically stable pilot-scale optimization, not as evidence that data sufficiency has been achieved. The need for a larger training set is therefore scientifically clear, but it could not be addressed within the present revision because the data-collection phase had already been completed under the fixed pilot-study scope.

Beyond threshold-free ranking, calibrated probabilities are central to cost-sensitive operation. Reliability diagrams show that isotonic calibration fitted on validation and held fixed at test improves alignment to the identity line across classes and at the pooled (micro) level, reducing expected calibration error and the Brier score. This provides a sound basis for selecting operating points by minimizing linear cost C or maximizing F_β according to deployment priorities. In concrete terms, sweeping the High-risk decision threshold from $\tau = 0.40$ to $\tau = 0.70$ tightens precision (97.56%→98.96%) and reduces false-positive rate (1.24%→0.41%) at a controlled expense in recall (98.90%→97.20%), with a balanced operating point at $\tau = 0.50$ (Precision = 98.43%, Recall = 98.46%, F_1 = 98.44%, FPR = 0.79%). These patterns align with applications that prioritize missed-high events ($C_{FN} > C_{FP}$), such as early-warning settings. At the same time, the balanced pilot design is optimistic with respect to deployment precision when High-risk events are rare. Using the selected operating point ($\tau = 0.50$) and transporting its sensitivity and false-positive rate to assumed High-risk prevalences of 1%, 5%, and 10%, the corresponding PPV becomes 55.73%,

86.77%, and 93.27%, respectively, while the expected false-alert burden remains approximately 14.1, 13.5, and 12.8 false alerts/h on the deployed 0.5 Hz timeline. Thus, the near-perfect discrimination observed under balanced held-out classes should not be interpreted as implying equally favorable alert precision in low-prevalence field environments. In practice, sustained deployment would likely require site-specific threshold selection, temporal alarm suppression or aggregation, and human-in-the-loop triage to keep operational burden acceptable.

Ablation analyses clarify the sources of performance. Under identical training, splits, and probability calibration, removing short-horizon sequential predictors (No LSTM) yields the largest degradation ($\Delta F1_{\text{macro}} = -1.22$ pp, ΔAUROC = −0.016, ΔAUPRC = −0.022, ΔBrier = +1.00 pp), indicating that look-ahead temporal structure captures predictive micro-dynamics that are not recoverable via leakage-free carry-forward or exponential smoothing. A concordant drop is observed for the No forecasting control ($L = 1, H = 0$), confirming that the gains stem from true short-horizon look-ahead rather than static smoothing. Suppressing weapon cues is the second most detrimental intervention ($\Delta F1_{\text{macro}} = -0.74$ pp, ΔAUROC = −0.010, ΔAUPRC = −0.014, ΔBrier = +0.55 pp), plausibly because these signals are directly tied to high-risk episodes. At the same time, this ablation should be interpreted with care: the visual branch measures only what is observable in the device's forward field of view. Consequently, the estimated contribution of person- and weapon-related cues reflects the value of local camera-visible evidence within the present acquisition geometry, not guaranteed observability of the full crowd or threat configuration in unrestricted field settings. Emotion-rate and heart-rate branches add moderate but consistent gains, while priors and audio/noise features contribute smaller, calibration-skewed improvements. The monotone increase in Brier across all ablations indicates that each family not only improves ranking but also contributes to probability quality.

Component analyzes further support these conclusions. First, per-module performance within the sensing/computing stack (SCS) indicates high-quality detectors and estimators (module-wise PR/RC/F1 ≈ 95%–96%), which helps explain the separability observed at the fusion layer. Second, counterfactual tests that (i) zero out each detector or (ii) inject ground-truth (GT) into the pipeline quantify the directionality of influence: zeroing a component reduces macro-F_1 and increases Brier, whereas injecting GT produces the opposite shifts, with the largest gains arising from the person detector and firearm cues. Taken together, these results provide the main empirical justification for retaining YOLOv7 in the present study despite its age: the person-detection task is comparatively mature, the detector performs strongly under the current forward-view pilot geometry, and its downstream contribution remains measurable when propagated through FLAD. Our claim is therefore not that YOLOv7 is the newest or universally best detector, but that it is sufficiently effective, reproducible, and edge-compatible for the specific proof-of-concept problem studied here. These findings are consistent with widely used, high-capacity person-detection backbones (e.g., YOLOv7) and classical, fast firearm detectors (Haar-like features) [13,14], and they motivate prioritizing these components in resource-constrained deployments. Nevertheless, the person detector should not be read as providing a panoramic or omnidirectional estimate of crowd size. In the current prototype, it summarizes only the locally visible sector covered by the camera, so its downstream influence on FLAD is best understood as view-conditioned scene evidence rather than a full situational census. Future work should benchmark newer detector families against YOLOv7 under the same field-of-view constraints, calibration protocol, and embedded-computing budget to determine whether the added architectural novelty produces a material application-level gain rather than only a nominal update in detector generation.

The interpretability contribution of FLAD should therefore be read at two complementary levels. Quantitatively, Table 3 shows that the fuzzy layer outperforms a similarly calibrated non-fuzzy MLP aggregator under matched inputs, folds, and validation-only calibration. Operationally, Table 4 shows how the same layer exposes auditable rule traces that can help an operator distinguish convergent multi-cue alarms from

borderline alerts that primarily reflect transient or weakly corroborated evidence. This is important in safety-critical monitoring, where the practical value of an alert depends not only on its score but also on whether the rationale can be inspected, communicated, and triaged in real time.

An additional practical implication is that the camera-derived antecedents in FLAD are inherently view-limited. Because the forward-facing device does not observe the full surrounding crowd, rules involving *People many/few* or *Weapons detected* should be interpreted as operating on the currently visible sector only. In consequence, alerts dominated by camera evidence are best treated as decision-support prompts that should be cross-checked with temporal evolution, sector context, and, when feasible, secondary human verification, rather than as exhaustive summaries of the wider scene.

Evidence beyond the in-domain split remains preliminary. The exploratory external cohort provides only a limited signal that discrimination and calibration may transfer beyond the development setting, but with $N_{\text{ext}} = 15$ it is not sufficient to support robust claims of generalization across sites or populations. Accordingly, the external results should be interpreted as feasibility-oriented rather than confirmatory. By contrast, the controlled distribution-shift analyses (low light/backlight, occlusion/pose, crowding, elevated ambient noise) are useful stress tests of internal robustness under predefined perturbations, but they do not substitute for large-scale external validation under organically occurring field heterogeneity. Within this limited scope, the observed OOD degradations remain modest and bounded, with occlusion/pose being the most adverse condition (e.g., $\Delta \text{F1}_{\text{macro}} = -1.6$ pp, $\Delta\text{AUROC} = -0.008$, $\Delta\text{AUPRC} = -0.009$, $\Delta\text{Brier} = +0.25$ pp). These patterns highlight priorities for subsequent data collection and model refinement, including occlusion-aware training, broader site coverage, and prospectively acquired multi-session cohorts.

From a systems perspective, the Raspberry Pi 3B+ evaluation now supports a materially stronger pilot-stage deployment claim because it combines pointwise latency with a continuous 6 h sustained-load profile. In addition to median/P95 end-to-end latencies of approximately 1.25/1.60 s, the platform sustained mean CPU utilization of 68.4% (±4.2%) with peaks of 89.2%, mean SoC temperature of 64.2°C (peak 68.5°C), mean throughput of 0.798 records/s with drift below 1.2%, and bounded High-risk alert jitter within ±140 ms. Interpreted on the deployed 0.5 Hz schedule, these measurements indicate that the system remained below timing saturation and preserved positive slack throughout the sustained-load run. The nominal 44.4 Wh battery pack further implies a battery-based mean-power proxy of 5.18 W, a peak proxy of 6.45 W, and an autonomy window of approximately 8.57 h under the profiled workload. Taken together, these results substantially strengthen the near-real-time feasibility claim for pilot operation on Raspberry Pi hardware. At the same time, because the power estimate is derived from the nominal battery budget rather than from inline electrical telemetry, future systems work should still examine direct power measurement under broader ambient-temperature, duty-cycle, and multi-stream operating conditions. This connection between statistical performance and runtime feasibility is critical for embedded deployments where energy, thermal limits, and response times must be balanced.

Methodologically, we emphasize three choices that enhance internal validity and interpretability. First, the participant—not the episode window—was treated as the unit of independence for generalization assessment: each outer LOSO fold withheld one entire subject, while preprocessing, model selection, calibration, and threshold selection were all completed within the remaining subjects and then frozen before test scoring. Second, uncertainty is quantified via a moving-block bootstrap that respects temporal dependence, with block length tied to the estimated correlation time; this yields conservative, two-sided 95% confidence intervals. Third, between-model AUROC differences and paired classification outcomes were assessed with DeLong's and McNemar's tests, respectively; inference is reported via effect sizes and two-sided 95% confidence intervals, which were consistent with the observed ranking advantages of the proposed system over non-fuzzy baselines.

The present study is a pilot evaluation emphasizing careful protocol control and transparent reporting. External validation remains strictly exploratory (N_{ext} = 15), and the OOD scenarios, though informative, are controlled perturbations rather than organically occurring shifts. Consequently, neither the small site-B cohort nor the synthetic stressors should be taken as sufficient evidence of robustness across institutions, populations, or operational contexts. The same caution applies to the target formulation itself: although the Low/Medium/High scale was defined *a priori* through an expert-informed operational codebook, it should presently be interpreted as a study-specific risk stratification for short-horizon security monitoring rather than as a universally standardized aggressiveness taxonomy. Broader scientific validity would benefit from future multi-expert consensus exercises and, where appropriate, alignment with formal security-escalation guidance used in comparable operational environments. Scaling to substantially larger, demographically diverse, multi-session, and multi-site cohorts would be necessary to estimate between-site variability, tighten uncertainty, and enable subgroup fairness analysis. Hardware-wise, the Raspberry Pi experiments provide a useful lower bound on embedded feasibility under the tested workload, but they should not be interpreted as a complete sustained-operation assessment. Continuous-operation stability, CPU load, and energy draw were not formally profiled and remain future work, together with additional platforms and dynamic scheduling policies (e.g., adaptive frame rates, detector gating) to optimize the latency–accuracy–power trade-off.

Prior efforts in surveillance and safety analytics typically report strong discriminative performance but rarely quantify calibration or cost-sensitive decision behavior. For instance, multimodal fusion for aggression/safety monitoring has reported accuracies in the 80%–90% range with limited treatment of probability calibration and thresholding [1]; violence detection via global motion and trajectory cues has achieved high accuracies on benchmark corpora without explicit ECE/Brier analysis [15]; and recent hybrid pipelines for physiological stress/risk modeling emphasize task-specific accuracy and throughput but only implicitly address decision costs [2]. In contrast, our system pairs near-ceiling AUROC/AUPRC with explicit calibration (Brier/ECE), threshold-sweep operating points, ablation-based attribution, and paired uncertainty quantification (moving-block bootstrap), while retaining interpretability through a fuzzy rule layer. At the component level, our SCS choices (YOLOv7 person detection and Haar-based firearm detection) align with widely adopted detectors [13,14], facilitating reproducibility and system-level transfer. Here it is important to distinguish detector recency from detector adequacy: YOLOv7 is not presented as the most current architecture, but as a mature and reproducible backbone whose held-out module performance and downstream impact are sufficient for the present proof-of-concept study. While protocol differences preclude head-to-head claims, the combination of discriminative performance, calibrated probabilities, uncertainty bands, and latency-compatible pilot-stage embedded profiling supports ADPS as a proof-of-concept early-warning prototype where both ranking and probability quality matter. Broader deployment claims should remain provisional until validated on larger, multi-session, and multi-site cohorts and complemented by formal sustained-load systems profiling.

Because this application domain is sensitive, the key ethical issue is not only how data were collected, but also how system outputs could be interpreted and potentially misused in practice. The formal approval, consent, and de-identification procedures are described in Section 5.1; here, we emphasize their implications. Even with ethics oversight, anonymized processing, and restricted data access, a pilot system such as ADPS should not be framed as an autonomous basis for punitive, disciplinary, or liberty-restricting decisions. Rather, its appropriate role is that of a human-in-the-loop early-warning aid whose outputs require contextual interpretation, proportionality, and professional oversight. This caution is particularly important because the present cohort is narrow in demographic and situational scope, and therefore does not support broad claims of contextual, demographic, or institutional generalizability. In high-stakes settings, the main ethical

safeguard is not only privacy protection, but also explicit limitation of use: the system should assist human judgment under governance controls, not replace it

The limits of demographic and contextual generalizability are equally important. Our pilot cohort comprised adult male police officers from a single institutional setting; therefore, findings should not be generalized to female officers, civilians, other age groups, or different operational and cultural contexts. We make no subgroup fairness claims, and the controlled OOD scenarios do not substitute for naturally occurring heterogeneity in real deployments. A larger, prospectively governed, multi-site, mixed-sex study would be required to assess fairness, calibration stability, and policy robustness across sex, age, site, and context.

5 Methods and Materials

5.1 Data Acquisition and Cohort

All participants were active-duty male police officers ($N = 10$). Findings should not be generalized to female officers or civilians; subgroup fairness analyses are out of scope for this pilot. The cohort size was fixed by the approved pilot-study scope and the available operational access during the data-collection window; no additional participant-level testing was undertaken beyond this feasibility phase. Consequently, the present article addresses sample-size limitations through explicit caution in interpretation rather than through post hoc expansion of the test cohort.

Ethics, consent, and privacy safeguards.

The pilot protocol, participant information sheet, consent procedure, and data-handling plan were reviewed and approved by the Ethics Committee of CUNEF Universidad. Participation was voluntary. Before recording, each participant provided written informed consent covering sensor acquisition, annotation, analysis, and publication of aggregate de-identified results. To reduce privacy exposure, acquisition and annotation were managed under anonymous participant codes; direct identifiers were stored separately from the research files; raw audiovisual and physiological recordings were accessible only to authorized research personnel; and the modeling tables used for development did not contain names or other direct personal identifiers.

Cardiac rhythm dataset (heart rate).

Instantaneous heart rate (HR, beats per minute) was recorded from 10 participants across 10 sessions of 15 min each. Offline, HR traces were uniformly resampled at $f_s = 4$ Hz ($\Delta t = 0.25$ s), yielding 3600 samples per session and ≈360,000 samples overall; the deployed system ingests HR at 0.5 Hz for computational efficiency. Quality control enforced a physiological admissible range of $[40, 180]$ bpm, with out-of-range samples flagged but not imputed to avoid bias.

Fig. A3 summarizes signal quality and temporal structure. Fig. A3a displays, for each subject–session, the proportion of flagged (out-of-range) samples; rates are uniformly low across the grid, indicating minimal clipping or sensor loss and supporting subsequent modeling without aggressive filtering. Fig. A3b reports the aggregate HR power spectral density (median across sessions with a pointwise 95% envelope) at 4 Hz. Spectral mass concentrates at very low frequencies and decays smoothly without narrow-band peaks, consistent with slowly varying dynamics and with the absence of acquisition-induced periodic artifacts; this structure justifies the use of short-horizon sequential predictors. Fig. A3c depicts session-wise mean vs. standard deviation. The compact cloud—with moderate dispersion and limited between-session heterogeneity—suggests stable within-session statistics and supports subject-wise evaluation protocols (e.g., leave-one-subject-out) with temporally blocked validation.

Emotions dataset (probabilistic class rates).

At each time step t, the emotion detector outputs a vector $\mathbf{e}_t \in [0,1]^7$ of normalized class rates (anger, disgust, fear, happy, sad, surprise, neutral) that approximately sum to one. Data were collected from the same 10 participants in 10 sessions of 15 min, sampled every ≈1.16 s (~1050 timestamps per session; ~105,000 samples overall).

Fig. A4 characterizes distribution, dependence, and short-term dynamics. Fig. A4a shows subject-wise median prevalences with interquartile ranges (IQR). On average, neutral and negative-affect classes dominate (e.g., neutral ~ 0.25, anger ~ 0.19, fear ~ 0.15, disgust ~ 0.12, sad ~ 0.11), whereas happy and surprise are less frequent (~0.10 and ~0.08), consistent with low-to-moderate base rates and transient bursts in negative classes. Fig. A4b reports pairwise Pearson correlations among class-rate channels; associations are uniformly small in magnitude and exhibit clear negative correlations against neutral (down to $r \approx -0.28$), indicating largely distinct, weakly redundant signals across classes. Fig. A4c depicts the first-order transition matrix of the dominant emotion $\arg\max_c e_{t,c}$. The average self-transition probability is low (≈0.14), and the most frequent transitions return to neutral (e.g., fear→neutral, surprise→neutral, anger→neutral, each ≈0.29), evidencing short episodes that revert quickly to baseline. Such dynamics motivate short-horizon forecasting and support temporally local fusion at the decision layer.

Evaluation protocol.

The unit of independence for generalization assessment was the *participant* (subject), not the individual episode window. Accordingly, each outer evaluation fold held out one entire participant, and no windows from that participant were used in model fitting, validation, calibration, threshold selection, or hyperparameter tuning. Within the remaining participants, validation was formed by temporally blocked segments separated from training by guard gaps, as detailed in Section 5.5. All preprocessing statistics, calibration mappings, and operating thresholds were determined without access to the held-out subject and then applied unchanged at test time.

5.2 Sensing Hardware and on-Device Platform

To conserve space, the assembled layout and full bill-of-materials/operating points are provided in Fig. A5 and Table A6.

5.3 Pre-Processing and Annotation

Episode construction and annotation. The main design parameters used to construct episode-level samples, together with a worked summary of nominal vs. effective sample size under the pilot protocol, are reported in Table 10. Video, audio, and heart-rate (HR) streams were time-stamped, synchronized to a common clock, and projected onto the fused ADPS decision timeline. Each supervised sample corresponded to an *episode-level decision point* at time t, formed from a causal look-back of $L = 3$ contiguous windows (approximately 6 s of history on the deployed 0.5 Hz timeline) and linked to the target at forecast horizon $H = 10$ steps (approximately 20 s ahead). Consecutive episodes were generated with a sliding stride of $s = 2$ steps, which implies partial temporal overlap between adjacent contexts when $s < L$. Under this configuration, the overlap ratio was 33.3%, as summarized in Table 10. Any candidate episode whose look-back or look-ahead interval crossed a subject boundary, session boundary, or train/validation/test partition boundary was excluded before model training or evaluation.

Table 10: Worked summary of episode construction and the distinction between nominal and effective sample size under the pilot LOSO protocol. The values shown here provide a transparent example consistent with the present experimental configuration and make explicit how the reported 20,000 held-out records arise from episode-level windowing rather than from independent participants or trials.

Item	Definition	Reported Value
Fused decision rate	Sampling rate of the fused ADPS decision timeline.	0.5 Hz
Look-back length L	Number of contiguous causal windows used as model input.	3 windows (≈6 s)
Forecast horizon H	Prediction lead time on the fused timeline.	10 steps (≈20 s)
Stride s	Sliding step used to generate consecutive episodes on the fused decision timeline.	2 steps (≈4 s)
Overlap ratio	Temporal overlap between neighboring episode contexts, computed as $(L - s)/L$ when $s < L$.	33.3%
Candidate episodes N_{cand}	All generated episodes before quality-control filtering and boundary exclusions.	21,900
Usable episodes N_{usable}	Episodes retained after exclusion of windows crossing subject/session/partition boundaries and other QC filters.	20,384
Held-out episodes N_{test}	Sum of test episodes concatenated across the 10 LOSO outer folds.	20,000
Effective sample size n_{eff}	Interpretive sample size under temporal dependence, approximately N/B in the moving-block bootstrap.	≈2500 (with $B \approx 8$)

We now explicitly distinguish between the *nominal* number of held-out records and the amount of statistically independent information. In particular, the reported test set of $N_{\text{test}} = 20{,}000$ corresponds to the concatenation of held-out episode windows across the 10 LOSO outer folds, rather than to 20,000 independent participants or independent trials. As further detailed in Table 10, the episode-generation pipeline yielded $N_{\text{cand}} = 21{,}900$ candidate episodes, of which $N_{\text{usable}} = 20{,}384$ remained after boundary-exclusion and quality-control rules, while inferential uncertainty was interpreted through an effective sample size of approximately $n_{\text{eff}} \approx 2500$ under temporal dependence. Because overlapping temporal windows can inflate the nominal record count, inferential uncertainty was quantified using a moving-block bootstrap and interpreted jointly with $n_{\text{eff}} \approx N/B$ rather than with the nominal episode count alone.

Ground truth followed a three-level ordinal scale (Low/Medium/High) defined *a priori* in a written operational codebook. Because the present task concerns short-horizon security-risk forecasting rather than psychiatric diagnosis, the target was operationalized as an expert-informed escalation scale rather than anchored to a single clinical thresholding standard. The codebook was drafted before model development and structured around three observable dimensions available in the synchronized recordings: (i) explicit threat-related scene evidence (e.g., weapon-like cue, hostile crowding, or visibly escalating confrontation in the active field of view), (ii) short-horizon physiological/affective activation (heart-rate/stress and emotion-rate elevation relative to the local baseline), and (iii) contextual immediacy (sector risk, ambient agitation, and temporal persistence across adjacent windows). Operationally, *Low* denoted baseline or de-escalated behavior with no explicit threat cue, no sustained escalation pattern, and low or stable physiological-affective

activation; *Medium* denoted emerging or ambiguous escalation in which one or more channels departed from baseline but the evidence remained incomplete, weakly corroborated, or short-lived; and *High* denoted either an explicit severe cue in context or convergent escalation across at least two of the three dimensions above, consistent with an immediate preventive or de-escalation response. Labels were assigned by human annotators using synchronized raw recordings through a two-pass adjudication procedure, consisting of majority vote followed, when necessary, by expert tie-break. Importantly, annotators did *not* use detector outputs, engineered features, LSTM forecasts, fuzzy-rule activations, calibrated probabilities, or any other model-derived score during labeling. In addition, no single observable cue available to the model inputs (e.g., a weapon-related visual cue, person count, or short affective burst) was treated as individually sufficient to assign a Low/Medium/High label; instead, labels were determined from the joint operational criteria specified in the codebook and the broader synchronized behavioral context. This formulation was intended to approximate a practical security-monitoring consensus for imminent escalation under the present patrol-like scenarios, while making explicit that the three levels are ordinal operational risk strata rather than universal clinical categories. Therefore, the target was not generated by thresholding model outputs or by collapsing a single input cue into the outcome definition, although the raw multimodal streams naturally contained the behavioral evidence later exploited by the predictive system.

On a stratified 10% subset, inter-rater reliability satisfied the *a priori* quality targets, with weighted Cohen's $\kappa_w \approx 0.82$ (using quadratic weights; see Appendix A.3.1, Eqs. (A2)–(A5)) and Krippendorff's $\alpha \approx 0.80$ (Eqs. (A6)–(A9)). Two-sided 95% confidence intervals were estimated with a moving-block bootstrap using overlapping blocks $\mathcal{B}_s$ and m blocks per replicate (Appendix A.3.2, Eqs. (A15) and (A16)). Formal definitions, including Fleiss' κ (Eq. (A12)), together with label-noise diagnostics, are provided in Appendix A.3.1, whereas temporal-dependence handling and uncertainty quantification are detailed in Appendix A.3.2. Dataset-level quality summaries are reported in Figs. A3 and A4.

5.4 Model Architecture

The Aggression Detection–Prediction System (ADPS) integrates three modules: a Sensor Control System (SCS) that standardizes per-frame/per-second evidence, an LSTM-Based Prediction System (LBPS) that forecasts short-horizon physiological and affective trajectories, and an interpretable fuzzy aggregation layer (FLAD) that produces calibrated class posteriors. Fig. 3 presents the main-text architecture of the proposed Fuzzy–LSTM model, making explicit how causal temporal sequences are processed by the LSTM forecasters before being combined with current multimodal cues in the fuzzy decision layer. Fig. A6 retains the extended interface-oriented overview.

Fig. 3 summarizes the proposed Fuzzy–LSTM architecture by separating the current-cue bypass from the causal LSTM forecasting branch. This distinction clarifies that current multimodal cues enter the fuzzy decision layer directly, whereas past emotion-rate and heart-rate sequences are first processed by compact LSTM forecasters before being combined with the fuzzy inputs to produce calibrated Low/Medium/High risk posteriors and an operator-facing rule rationale.

5.4.1 Sensor Control System (SCS)

The SCS ingests RGB frames, ambient audio, and heart-rate (HR) telemetry and exposes standardized feature channels to downstream modules. Vision channels comprise person localization and weapon-cue evidence; audio channels summarize ambient energy and noise proxies; physiological channels provide instantaneous HR and short-window statistics.

Person detection used a pre-trained single-shot model [13] configured for the COCO label set and restricted to the person class (id 0). Each frame was resized and normalized to the detector's native input resolution (416 × 416). Detections were retained at a confidence threshold ≥0.5 and filtered per frame with non-maximum suppression using a score threshold of 0.5 and an intersection-over-union (IoU) threshold of 0.4. The system exported per-frame person counts and bounding boxes; no temporal smoothing beyond per-frame NMS was applied. These operating thresholds prioritize precision in surveillance-like conditions.

Although YOLOv7 is no longer the newest detector family, it was selected deliberately for this pilot for three practical reasons. First, the target class here is *person*, which is among the most mature and best represented categories in large-scale pre-training corpora such as COCO; thus, the design question in the present work was not to introduce a new detector, but to use a stable off-the-shelf person-localization backbone within a multimodal forecasting pipeline. Second, YOLOv7 remains widely reproduced, well documented, and straightforward to deploy in compact embedded environments, which supports methodological transparency and reproducibility. Third, its adequacy in the present sensing geometry is supported empirically by the module-level held-out results in Table A1 (Precision = 95.7%, Recall = 95.1%, F_1 = 95.4%), together with the counterfactual detector-impact analysis in Table A3. We therefore use YOLOv7 here as a reproducible and edge-compatible person-detection backbone for pilot evaluation, while recognizing that newer detector families should be benchmarked in future revisions aimed at optimized deployment performance.

Because the device camera is forward-facing and body-/platform-mounted, the exported person count is interpreted throughout the manuscript as a *local field-of-view occupancy cue* rather than as a census of the surrounding crowd. In practice, it only reflects the visible region in one direction and within a limited angular span, and is therefore sensitive to viewpoint, distance, partial occlusion, and scene framing. Accordingly, ADPS does not treat the person-count variable as a globally valid estimate of crowd size; instead, the linguistic terms *People few/moderate/many* are intended to encode relative scene density within the currently observed camera sector.

Ambient sound was summarized as wideband root–mean–square (RMS) amplitude from 16-bit, single-channel PCM audio sampled at 44.1 kHz. Signals were partitioned into non-overlapping 2.0 s windows (approximately 88,200 samples), internally buffered in frames of 1024 samples with an equal hop [16]. For each window, a single scalar RMS level was computed and exported as the ambient-noise cue. No pre-emphasis, spectral weighting (e.g., A-weighting), denoising, voice-activity detection, or temporal smoothing was applied. Per-window values were time-stamped at window midpoints and aligned with video and heart-rate descriptors for downstream fusion; unless stated otherwise, they were standardized using training-set statistics before modeling.

All visual detectors were used off-the-shelf as pre-trained models without task-specific fine-tuning, bootstrapping, or domain adaptation; operating points are exactly those reported above.

Let s index spatial sectors and $b \in \{1, \ldots, 168\}$ denote hour-of-week bins. From incident logs archived in our public data repository (see Data Availability), computed over a rolling W-week window, let $c_{s,b}$ be the number of relevant incidents and $n_{s,b}$ the corresponding exposures. We compute a Laplace-smoothed rate $r_{s,b} = \frac{c_{s,b}+\lambda}{n_{s,b}+2\lambda}$, $\lambda = 1$, then min–max normalize across sectors to obtain $\tilde{r}_{s,b} \in [0,1]$. A monotone calibration ϕ (fitted on the validation split via isotonic regression and held fixed at test) yields the prior membership used by FLAD, $\mu_{\text{prior}} = \phi(\tilde{r}_{s,b})$. The SRP is refreshed every ΔT (e.g., weekly) by exponentially weighted updating with half-life H days; for all reported results the SRP snapshot is frozen from the training window. Its incremental value is assessed in the "No SRP" ablation (Table 5).

All raw detections and scores are quality-controlled and temporally aligned (Section 5.3); outputs are rate-limited to per-second descriptors to bound latency and memory. This design follows evidence that combining scene dynamics with human-centric cues improves violence/aggression analytics in surveillance settings [1,15,17].

We employed a pre-trained Haar cascade classifier to flag weapon-like patterns [14]. Each RGB frame was deterministically resized to a width of 500 px (aspect ratio preserved) and converted to grayscale before detection. The detector operated at a fixed configuration with scale factor set to 1.3, minimum neighbors to 20, and minimum detection window to 100 × 100 px; the minNeighbors parameter functioned as the effective acceptance threshold, with higher values prioritizing precision over recall. The Sensor Control System then exported a binary weapon cue per frame, defined as 1 when at least one bounding box was returned and 0 otherwise; no additional non-maximum suppression or temporal smoothing was applied. This operating point was chosen to reduce false positives under surveillance-like conditions.

The same visibility constraint applies to the weapon-like cue: it represents visual evidence available within the active camera view only, not complete observability of the full environment. For this reason, camera-derived person and weapon cues are treated as *partial scene evidence* in the downstream decision layer and should be interpreted jointly with sector risk, ambient audio, and physiological/affective forecasts rather than as a complete description of the surrounding scene.

5.4.2 LSTM-Based Prediction System (LBPS)

LBPS comprises two uni-variate sequence models that forecast near-future emotion-rate and heart-rate trajectories from recent histories. We use compact LSTMs (few layers/hidden units) to control overfitting at the pilot scale; hyperparameters were tuned by a population-based genetic search constrained by validation loss (details in Section 5). Training traces show monotone decrease and stabilization of the objective across generations for both tasks; see Fig. A2, where Fig. A2a depicts the GA-optimized LSTM for emotion-rate (per-generation minimum/average loss) and (Fig. A2b) the homologous curve for HR. The final forecasting-model settings associated with these traces are summarized in Table A7. In several folds, the best validation region was reached after relatively few effective epochs before early stopping. In the context of this pilot, such rapid convergence should not be interpreted as evidence that the forecasting problem is saturated or already validated for deployment-oriented use. Rather, it is compatible with the combination of short-horizon targets, compact one-layer forecasters, and limited between-subject diversity, and it reinforces the need to interpret the present LBPS results as pilot-scale. Larger and more heterogeneous training cohorts will be required to determine whether the same optimization behavior and predictive gains persist under broader operational variability. Because the present article reports a completed pilot dataset acquired under a fixed approved cohort, no additional subject-level sequences could be added within the scope of this revision to test that issue directly. These forecasters are motivated by the empirically demonstrated robustness of LSTMs for short-horizon time-series prediction across domains [5–9]; their outputs enter FLAD as temporally informative risk features (Formal training/selection criteria and calibration are described in Section 5; exact loss definitions are deferred to Appendix A.3 when needed).

In our implementation, the forecasters use $L = 3$ context windows and predict $H = 10$ steps (≈20 s ahead) on the fused 0.5 Hz timeline, operationalizing the short-horizon objective that motivates LBPS.

5.4.3 Fuzzy Logic Aggression Detection (FLAD)

FLAD implements an interpretable rule-base that aggregates SCS signals (e.g., person/weapon cues, audio proxies) and LBPS forecasts (emotion-rate/HR trends) into class posteriors over {Low, Medium, High}. The FLAD rules were not data-mined from the held-out labels; rather, they were constructed from the same operational logic used to organize the pilot monitoring problem. Specifically, the candidate antecedents were first defined from variables that are directly available at inference time (sector risk, person occupancy in the visible camera sector, weapon-like cue, ambient noise, heart-rate/stress level, and emotion-rate level). These variables were then mapped into linguistic states (e.g., low/medium/high or detected/undetected) and combined into a compact rule set intended to capture clinically and operationally plausible escalation patterns, such as "weapons plus crowding", "high stress plus high affective activation", or "low stress plus no weapon and low sector risk". The initial candidate set was reviewed for redundancy and interpretability, after which the final 12-rule base in Table A8 was retained as the smallest rule inventory that still covered the intended high-, medium-, and low-risk situations.

Fuzzification is performed on the normalized input channels before rule evaluation. Continuous inputs are mapped into monotone bounded membership functions with overlapping support, while binary cues such as weapon detection use crisp memberships. For a normalized scalar input $z \in [0,1]$, a typical three-term partition is represented as $\mu_{\text{low}}(z)$, $\mu_{\text{medium}}(z)$, and $\mu_{\text{high}}(z)$, with adjacent overlap to avoid discontinuous decision jumps near thresholds. Let $a_1, \ldots, a_m \in [0,1]$ denote the antecedent memberships entering a rule. FLAD combines conjunctive antecedents with the bounded-product t-norm,

$$T_{\text{bp}}(a,b) = \max\{0, a+b-1\}, \tag{1}$$

applied recursively for $m > 2$, while disjunctive links use the standard maximum operator. If rule r has consequent class $c(r) \in \{\text{Low}, \text{Medium}, \text{High}\}$ and antecedent activation α_r, its contribution is weighted as $\beta_r = w_r\,\alpha_r$, where $w_r \in (0,1]$ is the rule weight reported in Table A8. Class-wise supports are then aggregated over rules with the same consequent and normalized to obtain the posterior triplet,

$$s_c = \max_{r:c(r)=c} \beta_r, \qquad \hat{P}_c = \frac{s_c}{\sum_{c' \in \{L,M,H\}} s_{c'}}, \tag{2}$$

which is the quantity subsequently calibrated as described in Section 5.5. This formulation makes the fuzzy logic explicit: fuzzification defines graded evidence, the bounded-product t-norm encodes a soft logical AND that penalizes weak joint support, and the normalized class supports provide an auditable bridge between rule firing and final probabilistic output.

The complete rule inventory with linguistic terms and weights is provided in Table A8; design choices are consistent with contemporary fuzzy-system practice and optimization for decision support [18–20]. Importantly, the people-related antecedents in Table A8 refer to the number of individuals observable within the camera field of view at the current decision point, not to the total size of the surrounding crowd. Likewise, weapon-related antecedents reflect visible weapon-like patterns in that same restricted view. The rule base was therefore designed to treat camera-derived cues as local, incomplete evidence: several rules require corroboration from nonvisual inputs such as sector risk, stress, emotion-rate, or ambient noise, and even when a strong visual cue contributes prominently, the resulting alert is still interpreted through the full posterior bundle and adjacent temporal context. The sufficiency of this compact rule base was assessed, rather than assumed, through the leave–one–rule–out and membership-sensitivity analyses reported in Table A2: no rule met the joint pruning criterion of high overlap and negligible impact, and moderate perturbations of the membership functions produced only limited changes in Macro-F_1 and Brier, which supports the

local robustness of the final 12-rule specification (Probability calibration and operating-point selection are handled as described in Section 5).

At inference time, FLAD produces not only normalized class posteriors but also a compact explanation bundle for operator review. Specifically, the system logs the dominant rule activations, the corresponding antecedent memberships, and the posterior triplet $(\hat{P}_{\text{Low}}, \hat{P}_{\text{Medium}}, \hat{P}_{\text{High}})$. The operator-facing display is intentionally concise: it reports the winning risk level, the most influential two or three rules, and a short text rationale synthesized from those rules (e.g., "weapon cue + crowding + rising stress"). This design allows the user to distinguish alerts supported by convergent multimodal evidence from borderline alarms driven mainly by one cue family. Representative explanation traces are shown in Table 4.

5.5 Training, Splits, and Probability Calibration

Train/test construction and leakage control. Table 11 summarizes the fold-level LOSO workflow and the corresponding leakage-control safeguards. We trained two univariate LSTM regressors, one for heart rate (HR) and one for the scalar emotion-rate signal, under an identical subject-disjoint protocol. Feature standardization was fitted using training-set statistics only, and regression targets were standardized with fold-specific z-score scalers estimated exclusively from the training partition of each outer fold. Inputs were organized as fixed-length causal sequences of $L > 1$ time steps (input shape $[T = L, d]$) extracted from the temporally aligned feature stream by a sliding-window procedure. Unless otherwise stated, we used $L = 3$ contiguous windows of the fused timeline (approximately 6 s of context at the deployed 0.5 Hz rate) and a prediction horizon of $H = 10$ steps (approximately 20 s ahead), so that each supervised instance had the form $(\mathbf{x}_{t-L+1:t}, y_{t+H})$. The sliding-window stride was fixed at s steps (reported explicitly in Table 10); thus, adjacent examples could be temporally correlated and, when $s < L$, partially overlapping in their causal context. Any window whose look-back or look-ahead crossed a subject boundary, session boundary, or train/validation/test boundary was discarded to prevent temporal bleed-through. Each network consisted of a single LSTM layer with u hidden units followed by a linear Dense (1) output layer. Hyperparameters were optimized separately for each forecasting task through a population-based genetic search over $u \in [50, 200]$ and learning rate $\alpha \in [10^{-4}, 10^{-2}]$ (population = 20, generations = 100, tournament size = 3, blend crossover $\alpha = 0.5$, Gaussian mutation $\sigma = 1$, $indpb = 0.2$), using validation MSE as the selection criterion. Optimization employed Adam (batch size $B = 16$), with a maximum of $E_{\max} = 100$ epochs and early stopping on val_loss (patience = 10, restore_best_weights = True). Because the forecasters are deliberately compact and the prediction horizon is short, the selected models often entered their best validation region after relatively few effective epochs before early stopping. We therefore interpret fast convergence as a pilot-scale optimization characteristic under the present data regime, not as proof that the training set is already sufficient for deployment-oriented generalization. Accordingly, the deployment-oriented results reported in this manuscript should be read as conditional on the current fixed pilot training regime rather than as evidence that further training data would leave the forecasters unchanged. Experiments were repeated across random seeds affecting initialization and within-fold data order, and performance metrics were aggregated only after completion of all outer LOSO folds. A complete list of hyperparameters and training settings is provided in Table A7.

Table 11: Fold-level LOSO evaluation workflow and leakage controls. The held-out subject is the unit of independence for generalization assessment; episode-level windows are nested within that subject and are aggregated across outer folds only after all fold-specific decisions have been frozen.

Step	Data Used in Outer Fold j	Operation and Leakage Control
1	Held-out subject j	Reserve one entire participant as the test unit, which defines the unit of independence for fold-level generalization assessment. No windows from this participant are accessed during training, validation, model selection, calibration, or threshold selection.
2	Remaining 9 subjects	Construct the training and temporally blocked validation partitions from the non-test subjects, using guard gaps of at least L samples. Discard any candidate window whose look-back or look-ahead crosses a subject boundary, session boundary, or train/validation/test boundary.
3	Training subset only	Fit preprocessing transforms (e.g., feature standardization) and train the candidate forecasting models using the training subset only. No information from validation or test is used to estimate model parameters.
4	Validation subset only	Compare candidate models on validation performance, select the final model/hyperparameter setting, fit isotonic calibration mappings using validation predictions and labels only, and determine any operating threshold τ from calibrated validation outputs only (e.g., by maximizing F_β or minimizing a pre-specified linear cost).
5	Held-out subject j	Freeze preprocessing transforms, selected model parameters, isotonic mappings, and any threshold τ, and then perform a single-pass evaluation on the held-out participant. No re-fitting, recalibration, or threshold adjustment is allowed after observing test predictions.
6	All 10 outer folds	Concatenate the held-out predictions from the 10 test subjects to obtain the nominal held-out test set (N_{test} = 20,000 episode windows). After all fold-specific decisions have been frozen, report confidence intervals from a moving-block bootstrap applied to held-out predictions only, and interpret uncertainty jointly with n_{eff} rather than the nominal record count alone.

The unit of independence for evaluation was the *held-out participant*. In outer fold j, one participant was reserved exclusively for testing, thereby defining the fold-level generalization unit. No episode from that participant was used during training, validation, model selection, calibration, or threshold selection. The remaining nine participants were used to construct the training and temporally blocked validation partitions, with guard gaps of at least L samples inserted around every train/validation boundary. As summarized

in Table 11, the fold-wise procedure was: (1) generate candidate episode windows from the nine non-test participants; (2) discard any window whose look-back or look-ahead crossed a subject, session, or partition boundary; (3) fit preprocessing transforms and train candidate models using the training subset only; (4) compare candidate models on the validation subset, select the final hyperparameter setting, and fit isotonic calibration mappings using validation predictions and labels only; (5) determine any alerting threshold τ for the optional High-risk operating point from calibrated validation outputs only; and (6) freeze the complete pipeline and apply it in a single pass to the held-out participant. After all fold-specific decisions had been frozen, the held-out predictions from the 10 outer LOSO folds were concatenated to form the reported nominal test set of 20,000 held-out *episode-level windows*. These 20,000 records therefore do not correspond to 20,000 independent participants or independent trials; the independent generalization units are the 10 held-out participants defining the outer folds. Accordingly, inferential uncertainty was quantified with a moving-block bootstrap and interpreted jointly with the effective sample size under temporal dependence, rather than with the nominal record count alone.

To preserve probabilistic interpretability under a strictly leakage-free protocol, per-class posterior probabilities were calibrated by isotonic regression using the validation partition of each outer fold only, and the resulting monotone mappings were subsequently applied unchanged to the held-out subject. No calibration model was re-estimated, updated, or refined on the test fold at any stage. Calibration performance was quantified with the percentage-scaled Brier score and the Expected Calibration Error (ECE), and further examined visually through reliability diagrams. When an operating threshold was required for the optional High-risk alert, a scalar threshold τ was selected exclusively from calibrated validation predictions, either by maximizing recall-weighted F_β or by minimizing a pre-specified linear decision cost, $C = C_{\mathrm{FN}} \cdot \mathrm{FN} + C_{\mathrm{FP}} \cdot \mathrm{FP}$. The chosen threshold was then frozen before test-time inference and evaluated without modification on the held-out participant. For transparency, Table 6 reports the operating characteristics observed on the test set at several pre-specified thresholds; however, these values are provided for descriptive comparison only, and no threshold listed in that table was tuned, selected, or adjusted on the test data.

Two-sided 95% confidence intervals (CIs) were computed on the held-out test set using a moving-block bootstrap (MBB) to account for residual temporal dependence. The block length B was set to satisfy $B \geq \widehat{\tau}_{\mathrm{int}}$ (estimated integrated autocorrelation time), and we report the effective sample size $n_{\mathrm{eff}} \approx N/B$ for interpretability; see Appendix A.3.2, Eqs. (A13)–(A16). The same calibration mappings and preprocessing statistics were used throughout the resampling procedure to preserve the evaluation protocol. Baselines and ablations were trained on identical features/splits and calibrated in the same way to enable fair, like-for-like comparisons.

For the SRP feature, the validation-fitted monotone mapping ϕ was kept fixed at test time (no re-fitting), mirroring the protocol used for probability calibration elsewhere in the pipeline.

5.6 Statistical Analysis

We evaluate multiclass performance in a one-vs.-rest setting, deriving per-class scores from the model posteriors $\hat{p}_c(x) \in [0,1]$ and tracing Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves by sweeping a threshold $t \in [0,1]$ over $\hat{p}_c$. Discrimination is summarized with micro- and macro-aggregated metrics: micro-averages pool confusion-counts across classes (prevalence-weighted), whereas macro-averages take the unweighted mean of per-class metrics. Formal definitions of the calibration metrics used (percentage-scaled Brier and ECE) appear in Appendix A.3.2, Eqs. (A17) and (A18); standard one-vs.-rest discrimination conventions are followed throughout.

Uncertainty is quantified with two-sided 95% confidence intervals computed via a moving-block bootstrap (MBB) that respects serial dependence in the temporally ordered test set. The bootstrap block length B is anchored to the estimated integrated autocorrelation time $\widehat{\tau}_{\text{int}}$ in Appendix A.3.2, Eq. (A13), yielding an interpretable effective sample size $n_{\text{eff}} \approx N/B$ in Eq. (A14). Overlapping blocks $\mathcal{B}_s$ and the number of blocks per replicate m follow Eqs. (A15) and (A16); unless stated otherwise, percentile intervals are reported.

Probabilistic outputs are calibrated on the validation split using isotonic regression and the learned mappings are held fixed at test (no re-fitting). Calibration quality is summarized numerically by Brier (%) and ECE (%) in Appendix A.3.2, Eqs. (A17) and (A18) and visually with reliability diagrams. When decision operating points are required (e.g., High-risk alert), the threshold τ is fixed on validation either by maximizing F_β (recall-emphasis) or by minimizing a linear deployment cost C; the selected τ is then evaluated unchanged on the held-out test set. Because the held-out pilot set is class-balanced by design, deployment-oriented alert statistics were additionally transported to assumed real-world High-risk prevalences $\pi_{\text{High}} \in \{0.01, 0.05, 0.10\}$. For a fixed threshold τ with sensitivity $\text{Se}(\tau) = \Pr(\hat{Y}=1 \mid Y=1)$ and false-positive rate $\text{FPR}(\tau) = \Pr(\hat{Y}=1 \mid Y=0)$, the prevalence-adjusted positive predictive value was computed as Eq. (3).

$$\text{PPV}(\pi_{\text{High}}, \tau) = \frac{\text{Se}(\tau)\,\pi_{\text{High}}}{\text{Se}(\tau)\,\pi_{\text{High}} + \text{FPR}(\tau)\,[1-\pi_{\text{High}}]} \tag{3}$$

For the deployed decision rate $r = 1800$ windows/h on the fused 0.5 Hz timeline, the expected total alert burden and false-alert burden were computed as Eqs. (4) and (5).

$$A_h(\pi_{\text{High}}, \tau) = r\left[\text{Se}(\tau)\,\pi_{\text{High}} + \text{FPR}(\tau)\,[1-\pi_{\text{High}}]\right] \tag{4}$$

$$FA_h(\pi_{\text{High}}, \tau) = r\,\text{FPR}(\tau)\,[1-\pi_{\text{High}}] \tag{5}$$

These prevalence-transport calculations do not alter threshold-free discrimination metrics such as AUROC or AUPRC; rather, they translate a fixed operating point into deployment-oriented quantities under rare-event assumptions. Operational utility was assessed via decision-curve analysis (DCA)—reporting net benefit as a function of the threshold probability—and by estimating the expected alert burden (alerts per hour) at the selected operating points. Inter-rater agreement for the three-level ordinal ground truth was assessed on a stratified subset (quadratic-weighted Cohen's κ_w, Krippendorff's α, Fleiss' κ), with details and exact formulas in Appendix A.3.1, Eqs. (A2)–(A12).

Unless otherwise noted, all baselines and ablations share identical features, splits, calibration protocol, and MBB settings to enable like-for-like comparisons. Given approximately balanced class counts in the held-out set, macro and weighted aggregates coincide to first order; we therefore report macro values in the main text and provide weighted values in tables for completeness. To reduce ambiguity across displays, all figures and tables were prepared under a common reporting template: captions state the evaluation unit explicitly, class labels appear in the fixed order Low/Medium/High, pooled summaries are named consistently as micro and macro, and metric values quoted in the running text were proofread against the final table outputs after formatting revision.

Between-model differences in threshold-free discrimination were assessed with DeLong's paired test for AUROC contrasts (two-sided, $\alpha = 0.05$). For paired classification outcomes at a fixed operating threshold (chosen on validation and kept fixed at test), we used McNemar's test with continuity correction on

discordant counts (b, c). When multiple hypotheses were assessed, p-values were adjusted using the Holm–Bonferroni step-down procedure. We report absolute effect sizes (Δ) with two-sided 95% confidence intervals alongside p-values.

5.7 Baselines and Ablation Protocols

We benchmark ADPS against non-fuzzy classifiers trained on the same features and data splits: Logistic Regression (LR), Gradient Boosting (GBM), Random Forest (RF), and a shallow Multilayer Perceptron (MLP). All baselines are probability-calibrated on the validation split (isotonic regression) and evaluated on the held-out test set without further tuning; we report Macro-F1, AUROC, AUPRC, Accuracy, Brier (%), and ECE (%). To keep the comparator study transparent and like-for-like, every baseline received the same subject-disjoint folds, the same validation-only isotonic calibration procedure, and an explicitly bounded tuning budget within the non-test data of each outer fold. Full baseline search budgets, candidate hyperparameter sets, and the final settings used for reporting are summarized in Table A9, which documents the matched comparator budgets used in the pilot analysis.

To quantify the marginal contribution of each input family, we run one-at-a-time feature ablations under an identical training and calibration protocol. For each metric $M \in \{\mathrm{F1}_{\mathrm{macro}}, \mathrm{AUROC}, \mathrm{AUPRC}, \mathrm{Brier}\}$ we define $\Delta M := M_{\mathrm{ablated}} - M_{\mathrm{full}}$, where negative Δ in discrimination metrics indicates degradation and positive ΔBrier denotes worse calibration (Brier deltas expressed in percentage points when scaled).

For the No LSTM variant, short-horizon predictors are replaced with leakage-free surrogates—last-observation-carried-forward for emotion-rate and an exponentially weighted moving average for heart-rate with decay selected on validation—to isolate the value of forecasting while preserving the evaluation protocol. In addition to the No LSTM surrogate variant, we include a No forecasting control in which the LBPS architecture is evaluated with $L = 1$ and $H = 0$ (same-step regression) under identical features, splits, and calibration. This isolates the incremental value of explicit look-ahead ($H \geq 1$) beyond any smoothing or carry-forward effects.

Uncertainty for point estimates and ablation deltas is summarized with two-sided 95% confidence intervals computed via a moving-block bootstrap that respects temporal dependence (block length $B \geq \tau_{\mathrm{int}}$; effective sample size $n_{\mathrm{eff}} \approx N/B$); unless noted otherwise, intervals are percentile-based. To enable reproducible operating-point analyzes and cost-sensitive comparisons, all preprocessing and calibration mappings are fitted on training/validation and kept fixed at test for both baselines and ablations.

5.8 Exploratory External Validation and OOD Robustness

External evaluation was conducted on a small independently acquired cohort (site B) and is presented strictly as exploratory. The purpose of this analysis was to probe whether the internally developed pipeline could be applied without re-fitting outside the development cohort, not to establish robust external generalization across settings or populations. The same preprocessing pipeline, one-vs.-rest evaluation protocol, probability calibration (isotonic, fitted on the internal validation split), and decision thresholds selected on validation were applied unchanged. No retraining, re-tuning, or re-calibration was performed on external data. Discrimination (macro/micro F_1, AUROC, AUPRC) and calibration (Brier%, ECE%) metrics follow the conventions in Section 5.6 and Appendix A.3.2 (Eqs. (A17) and (A18)). Given the limited external sample size, these estimates should be interpreted as preliminary transportability signals rather than as definitive evidence of robustness.

To respect temporal dependence, two-sided 95% confidence intervals (CIs) were computed via a moving-block bootstrap (MBB) on temporally ordered episodes. The block length B was anchored to the

estimated integrated autocorrelation time $\widehat{\tau}_{\text{int}}$ (Appendix A.3.2, Eq. (A13)); overlapping blocks and the number of blocks per replicate follow Eqs. (A15) and (A16). Within each bootstrap replicate, the validation-fitted isotonic mapping and all preprocessing statistics were held fixed to preserve a leakage-free protocol.

Module-level transfer was assessed without distribution alignment by applying the emotion-rate branch to a public affect dataset under the same featurization/normalization used in-domain. Because the heart-rate branch is regression-only, it is evaluated with RMSE (defined in Appendix A.3) and excluded from classification tables to avoid mixing heterogeneous targets and loss functions. These cross-dataset module results provide supporting evidence at the component level, but they do not replace full-system external validation on a substantially larger independent cohort. Consistent with the external pilot analysis, all module evaluations inherit frozen thresholds and calibration mappings chosen on validation.

Robustness to covariate shift was evaluated under four controlled stressors—low light/backlight, occlusion/pose, crowding, and elevated ambient noise—implemented during acquisition. Let $M \in \{\text{F1}_{\text{macro}}, \text{AUROC}, \text{AUPRC}, \text{Brier}\}$ denote a scalar metric; the shift-induced change is defined by the performance delta

$$\Delta M = M_{\text{shift}} - M_{\text{in-domain}}. \tag{6}$$

As per Eq. (6), negative ΔM indicates degradation for discrimination metrics, whereas positive ΔM indicates worse calibration for Brier (reported in percentage points when scaled). Thresholds τ and isotonic mappings were fixed a priori from validation and kept constant across all OOD evaluations. CIs for ΔM were obtained via MBB using the same B as the in-domain reference; to preserve class balance and dependence, resampling used stratified, contiguous blocks per condition (Appendix A.3.2).

Where inferential comparisons were required, paired outcome differences were assessed with McNemar's test and AUROC contrasts with DeLong's test (two-sided, $\alpha = 0.05$); multiplicity, when relevant, was controlled via Holm adjustment. All procedures (calibration locking, blocked resampling, fixed operating points) were applied identically to the proposed model, baselines, and ablations to enable like-for-like comparisons.

5.9 Computational Efficiency Profiling

End-to-end latency was defined as the interval from sensor acquisition to the risk output, $t_{\text{e2e}} = t_{\text{out}} - t_{\text{in}}$ (Appendix A.3.3, Eq. (A19)). For each record i, the total processing time T_i was decomposed as the sum of latencies across instrumented pipeline stages $L_{i,j}$ (Appendix A.3.3, Eq. (A20)). We profiled all pipeline stages on-device (Raspberry Pi 3 Model B+) with high-resolution timers and summarize the distribution of $\{T_i\}$ in Fig. A1 (histogram); per-component medians, 95th percentiles, and ranges are reported in Table A5, and the bill of materials plus operating points in Table A6. In addition to pointwise latency instrumentation, we conducted a continuous 6 h sustained-load deployment on Raspberry Pi 3 Model B+ at the deployed fused decision period $\Delta = 2.0$ s (0.5 Hz). CPU utilization and SoC temperature were sampled throughout the run to assess computational density and thermal stability under passive cooling; we summarize them by empirical mean, standard deviation (for CPU load), and peak. Realized throughput was computed as the number of completed fused decisions per unit time, throughput drift as the relative deviation of the late-run throughput from the early-run reference throughput, and temporal regularity for operator-facing High-risk alerts as the absolute inter-arrival jitter around the deployed schedule. Under this protocol, the system sustained mean CPU utilization of 68.4% (±4.2%) with a peak of 89.2%, mean SoC temperature of 64.2°C with a peak of 68.5°C, and mean throughput of 0.798 records/s with drift below 1.2%; High-risk alert jitter remained within ±140 ms. For deployment planning, the nominal cell-based battery specification

in Table A6 ($E_{\text{bat}}^{\text{nom}} \approx 44.4$ Wh) was used to derive a battery-based power/autonomy proxy, corresponding to an estimated mean power draw of 5.18 W, a peak proxy of 6.45 W, and expected autonomy of approximately 8.57 h under the profiled workload. Direct inline electrical telemetry was not instrumented; accordingly, the power figures should be interpreted as battery-based proxies rather than as direct wattmeter measurements.

Unless otherwise noted, 95% intervals for quantiles (e.g., the 95th percentile $\widehat{Q}_{0.95}$) are computed from the empirical distribution of T_i (definition in Appendix A.3.3, Eq. (A21)), keeping the measurement protocol fixed across replicates.

5.10 Statistical Testing

Between-model differences in threshold-free discrimination were assessed with DeLong's paired test for AUROC contrasts (two-sided, $\alpha = 0.05$). Let $\Delta_A = A_1 - A_2$ denote the AUROC difference; the standardized z-statistic is defined in Appendix A.3.4, Eq. (A22). We report effect sizes and two-sided 95% confidence intervals.

For paired classification outcomes at a fixed operating threshold (chosen on validation and kept fixed at test), we used McNemar's test with continuity correction on the discordant cell counts (b, c) of the 2×2 table; the test statistic is shown in Appendix A.3.4, Eq. (A23).

When multiple pairwise comparisons were performed, familywise error was controlled with Holm's step-down procedure; the sequential critical levels are given in Appendix A.3.4, Eq. (A24). We report effect sizes and two-sided 95% confidence intervals; formal test definitions are provided in Appendix A.3.4.

Uncertainty for scalar metrics and deltas (Macro-F1, AUROC, AUPRC, Brier, ECE) was summarized with two-sided 95% moving-block bootstrap intervals that respect temporal dependence; block construction and the effective sample-size rationale follow Section 5.6 and Appendix A.3.2.

6 Conclusions and Future Work

This study developed and pilot-tested the Aggression Detection–Prediction System (ADPS), a mobile expert system that integrates lightweight on-body and ambient signals through a hybrid architecture of Long Short-Term Memory (LSTM) networks and an interpretable fuzzy logic decision layer. Within subject-disjoint in-domain evaluation, the proposed model achieved macro-F_1 = 98.3% and AUROC = 0.998, showing higher pilot-scale point estimates than the calibrated non-fuzzy baselines while maintaining low miscalibration (ECE = 1.20%). The primary contribution of this work lies in a unified five-part design logic: short-horizon temporal modeling, transparent fuzzy aggregation, lightweight sequence modeling, multimodal robustness to missingness, and deployment-oriented pilot evaluation. These points are intended to be read consistently across the manuscript as one coherent contribution thread rather than as isolated claims. Together, they provide auditable and computationally feasible pilot-stage aggression forecasting. In particular, the revised analysis now makes the contribution of the fuzzy layer more concrete by pairing the matched comparison against a calibrated non-fuzzy MLP with operator-facing rule traces that show how true-positive and false-positive High-risk alerts can be interpreted in practice. In addition, the revised Methods now make the FLAD construction itself more explicit by clarifying how the 12-rule base was built from inference-time variables and operational escalation patterns, how continuous cues were fuzzified into overlapping linguistic memberships, how the bounded-product t-norm was applied during antecedent aggregation, and how the final compact rule set was checked for sufficiency through leave–one–rule–out and membership-sensitivity analyses. At the same time, the present prototype should not be understood as observing the full crowd environment: the person-count and weapon cues are extracted from a forward-facing camera and therefore summarize only the locally visible sector. Their contribution to FLAD is useful,

but inherently partial, and broader scene understanding would require wider-angle, multi-camera, or otherwise complementary sensing. Likewise, the choice of YOLOv7 for person detection should be interpreted as a pragmatic pilot-stage engineering decision rather than as a claim that this detector is the newest available architecture. In the present study, its role is that of a reproducible, off-the-shelf, edge-compatible backbone whose adequacy is supported by the held-out module-level results and by the downstream counterfactual analyses. Furthermore, deployment on a Raspberry Pi 3B+ now indicates latency-compatible and sustained-load pilot-stage feasibility under the tested workload, supported by a continuous 6 h profile with mean CPU utilization of 68.4%, mean throughput of 0.798 records/s, bounded thermal load, and a battery-based mean-power proxy of 5.18 W. At the same time, the external cohort remains small and strictly exploratory; therefore, those findings should be viewed only as a preliminary signal rather than as sufficient evidence of robustness across sites or populations. The class-balanced pilot design is also more favorable than realistic field prevalence for High-risk events; accordingly, deployment interpretation should rely not only on balanced-set discrimination but also on prevalence-adjusted alert precision and false-alert burden. Overall, the present results support proof-of-concept feasibility, but the performance estimates should still be interpreted cautiously given the pilot scale, and broader claims of external generalization or deployment robustness should remain provisional until confirmed in substantially larger, multi-session, and multi-site cohorts, ideally under naturally imbalanced event rates and prospectively specified alarm-management policies. Because only 10 independent participants contributed the subject-held-out test folds, residual overfitting at the participant level cannot yet be excluded and should be treated as an explicit limitation of the present FLAD evaluation. Because this article reports the completed pilot dataset, that limitation could not be eliminated within the current revision by adding new subject-level tests. The appropriate corrective action is therefore transparent claim-bounding and explicit acknowledgment of uncertainty, with confirmation deferred to a later, larger validation campaign. In addition, the revised systems section now reports a continuous-operation profile that includes CPU load, thermal behavior, throughput stability, alert jitter, and a battery-based power/autonomy proxy. These measurements substantially strengthen the deployment argument beyond latency alone, although direct inline electrical telemetry and testing under harsher ambient or multi-stream conditions remain important targets for future systems evaluation. Given the sensitivity of the application domain, future deployment-oriented studies should pair technical validation with explicit governance, privacy-preserving data handling, and subgroup fairness audits. Future revisions should also strengthen the external validity of the target definition itself by testing whether the present Low/Medium/High codebook remains stable under broader multi-expert review and across institutions, scenarios, and security cultures. In that sense, the current aggressiveness scale should be viewed as an explicit pilot operationalization of short-horizon escalation risk, not as a final universal standard. Rapid LBPS convergence in this pilot should likewise be interpreted cautiously: because the compact forecasters often stabilized after relatively few effective epochs, substantially larger and more heterogeneous training datasets will be needed to verify that this behavior reflects learnable short-horizon structure rather than limited pilot-scale variability. Because no further subject-level sequences were available within this completed pilot, the appropriate present remedy is explicit claim-bounding rather than retrospective expansion of the training set within the current revision. Future research will focus on multi-site validation and the exploration of long-term human-system interaction to further enhance the generalizability and ethical deployment of the ADPS framework.

Acknowledgement: The authors would like to express their sincere gratitude to CUNEF Universidad for its invaluable support and for providing access to the facilities and resources that made this research possible.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization: Cesar Guevara and Victoria Lopez; Methodology: Cesar Guevara and Victoria Lopez; Software: Cesar Guevara; Validation: Cesar Guevara; Formal analysis: Cesar Guevara; Investigation: Cesar Guevara; Resources: Victoria Lopez; Data curation: Cesar Guevara; Writing—original draft: Cesar Guevara; Writing—review & editing: Cesar Guevara and Victoria Lopez; Visualization: Cesar Guevara; Supervision: Victoria Lopez; Project administration: Victoria Lopez. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Aggregate source data underlying all figures and tables are available at Mendeley Data (https://doi.org/10.17632/syxf6yzzc2.2, version 2.2), accessible to referees via a private, view-only link: https://data.mendeley.com/datasets/syxf6yzzc2/2. Raw audiovisual and physiological recordings are not shared publicly because the source material belongs to a sensitive high-stakes domain and could enable re-identification. Only aggregate outputs are reported in the manuscript, and any subset shared beyond the current article is de-identified and subject to an appropriate data-use agreement and ethics-compatible access conditions.

Ethics Approval: The study protocol, participant information sheet, consent procedure, and data-handling plan were reviewed and approved by the Ethics Committee of CUNEF Universidad and were conducted in accordance with the Declaration of Helsinki. Data collection and annotation were managed under anonymous participant codes; direct identifiers were stored separately from the research files; the analytical datasets used for model development were processed in de-identified form; and access to raw recordings was restricted to authorized research personnel. All participants provided written informed consent for participation, sensor recording, annotation, analysis, and the publication of aggregate de-identified results.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1

Table A1: Per-module performance of the SCS on the final test split with 95% CIs (moving-block bootstrap). Precision (PR), Recall (RC) and F_1 are reported as percentages. The main takeaway is that the off-the-shelf YOLOv7 person detector remains strong enough for the present pilot geometry (F_1 = 95.4%), so its role in ADPS is empirically supported even though it is no longer the newest detector family.

Module	PR (%)	RC (%)	F1 (%)
Person detector (YOLOv7)	95.7 [95.3, 96.1]	95.1 [94.7, 95.5]	95.4 [95.0, 95.8]
Firearm detector (Haar)	95.4 [95.0, 95.8]	95.9 [95.5, 96.3]	95.7 [95.3, 96.1]
Emotion analysis (facial)	95.2 [94.8, 95.6]	95.0 [94.6, 95.4]	95.1 [94.7, 95.5]
Macro average	95.43 [95.1, 95.7]	95.33 [95.0, 95.6]	95.38 [95.1, 95.7]

Table A2: Rule ablation (leave-one-rule-out, LORO) and membership sensitivity. The main takeaway is that the compact FLAD rule base passes a practical sufficiency check: pruning would require both high overlap and negligible impact, and none of the candidate removals satisfies those conditions. Absolute deltas (pp) are relative to the full model; ΔBrier is reported in percentage points for consistency with Brier (%).

Rule	Coverage (%)	Max Overlap	ΔMacro-F1 (pp)	ΔBrier (pp)	Decision
R1	18.2	0.41	−0.03	+0.01	keep
R7	12.4	0.86	−0.05	+0.02	keep
R12	9.7	0.81	−0.02	+0.01	keep

Note: **Sensitivity (median over all trimf):** |ΔMacro-F1| = 0.18 pp; |ΔBrier| = 0.08 pp (for ±10%). **After pruning (test):** Macro-F1 = **98.31%**, Brier = **1.91%**.

Table A3: Impact of each SCS detector on FLAD (test set). Δ denotes change vs. full system; 95% CIs via moving-block bootstrap. The main takeaway is that the person-detection branch materially affects downstream FLAD performance, which provides application-specific support for retaining YOLOv7 as the person-localization backbone in this pilot configuration.

Ablation	ΔMacro-F1 (pp)	ΔBrier (pp)
Zeroing YOLOv7 detections	−0.42 [−0.66, −0.21]	+0.06 [+0.03, +0.09]
Zeroing Haar detections	−0.31 [−0.55, −0.12]	+0.05 [+0.02, +0.08]
Zeroing facial emotion module	−0.27 [−0.49, −0.10]	+0.04 [+0.02, +0.07]
Injecting GT for YOLOv7	+0.18 [+0.07, +0.30]	−0.02 [−0.03, −0.01]
Injecting GT for Haar	+0.12 [+0.04, +0.23]	−0.02 [−0.03, −0.01]
Injecting GT for facial	+0.09 [+0.02, +0.20]	−0.01 [−0.02, −0.00]

Table A4: OOD stress-test: performance deltas vs. the in-domain test. Negative values in Macro-F1 indicate degradation.

Shift	ΔMacro-F1 (pp)	ΔAUROC	ΔAUPRC	ΔBrier (pp)
Low light	−1.2	−0.006	−0.007	+0.18
Occlusion/pose	−1.6	−0.008	−0.009	+0.25
Crowding	−0.9	−0.004	−0.005	+0.12
High noise (audio)	−0.7	−0.003	−0.004	+0.10

Table A5: Raspberry Pi 3 Model B+ per-component processing time measured on-device over 20,000 records; median, 95th percentile (P95), and range (seconds). End-to-end is measured per record as the sum of stage times.

Pipeline Stage	Median (s)	P95 (s)	Range (s)
Capture & pre-processing	0.10	0.14	0.06–0.18
Person detection (YOLOv7)	0.62	0.78	0.45–1.00
Weapon detection (Haar)	0.12	0.18	0.08–0.24
Face/emotion analysis	0.17	0.24	0.10–0.30
LSTM (emotion-rate)	0.05	0.07	0.03–0.09
LSTM (heart-rate)	0.04	0.06	0.03–0.08
FLAD inference	0.097	0.13	0.06–0.15
I/O (audio/GPS/logging)	0.05	0.10	0.03–0.14
End-to-end (aggregated)	1.25	1.60	0.70–1.80

Table A6: Hardware bill of materials and nominal operating conditions. Sustained-load CPU, thermal, and throughput profiling is summarized in the main text; the power entry below provides the nominal battery specification used for the battery-based energy proxy rather than direct electrical telemetry.

Component	Model/spec	Notes (Operating Point)
Camera	Raspberry Pi HQ (Sony IMX477, 12.3 MP, 1/3″)	Fixed exposure/gain per session; focus set to patrol distances; forward-view HD RGB acquisition.
Microphone	USB mini microphone	Monaural; nominal 16 kHz bandwidth; standardized input gain at session start; synchronized timestamps.
Heart rate	Moofit HR8 chest strap	Instantaneous HR (bpm); admissible physiological range $[40, 180]$ bpm; uniform resampling for modeling.
Compute	Raspberry Pi 3 Model B+	1.4 GHz 64-bit quad-core; dual-band Wi-Fi; Bluetooth 4.2/BLE; PoE-capable Ethernet; Linux OS; on-device inference/logging.
Power	4× Li-ion polymer 3.7 V, 3000 mAh (103665)	Nominal battery pack specification used during pilot deployment (cell-based energy budget ≈44.4 Wh). For deployment planning, this corresponds to allowable mean-power budgets of approximately 7.4 W for 6 h autonomy, 5.55 W for 8 h, and 4.44 W for 10 h. This specification was used to derive the battery-based mean-power/autonomy proxy reported in Table 9; direct inline electrical telemetry was not instrumented.
I/O	Earphones	Operator audio feedback from the ADPS.

Table A7: LSTM hyperparameters and training settings for the heart-rate and emotion-rate models. This table reports the exact forecasting-model settings used in the pilot evaluation.

Component	Value
Architecture	LSTM → Dense (1, linear)
Hidden units	Integer in $[50, 200]$ (optimized)
Look-back window (timesteps)	3
Look-ahead horizon	10 steps (≈20 s ahead at 0.5 Hz)
Batch size	16
Max epochs	100
Early stopping	Monitor `val_loss`; patience = 10; restore best weights
Optimizer	Adam
Learning rate	$[10^{-4}, 10^{-2}]$ (optimized)

(Continued)

Table A7 (continued)

Component	Value
Loss	Mean squared error (MSE)
Train/validation split	Inner validation within training fold (temporally blocked) per outer LOSO fold; calibration fitted on validation only
Train/test split	Subject-disjoint Leave-One-Subject-Out (LOSO), 10 outer folds; no fixed random split
Random seeds	Seeds fixed for initialization and data order during training; metrics aggregated across outer folds
Input scaling	Standardization (fit on training only)
Input resolution	n/a (tabular sequence features)
Hyperparameter search	GA (DEAP): pop = 20, gen = 100, $cxpb$ = 0.5, $mutpb$ = 0.2, tournament = 3
Model selection	Best validation loss (MSE) within each outer fold

Table A8: FLAD rule base (12 rules) with linguistic antecedents and consequents. The rule inventory shown here is the final compact specification retained after expert-guided construction, redundancy review, and the sufficiency checks summarized in Table A2. People-related antecedents refer to relative occupancy within the active camera field of view at the decision point, and weapon-related antecedents refer to visible weapon-like patterns in that same restricted view rather than to complete scene observability.

Rule	Linguistic Antecedent (Min/Max Logic)	Consequent	Output Set
R1	Weapons are detected.	High	$\mu_{\text{high}}(y)$
R2	Weapons are detected **AND** People are many.	High	$\mu_{\text{high}}(y)$
R3	Weapons are detected **AND** Sector risk is high.	High	$\mu_{\text{high}}(y)$
R4	Stress is high **AND** Emotions are high **AND** (People are moderate **OR** People are many).	High	$\mu_{\text{high}}(y)$
R5	Stress is high **AND** Emotions are medium **AND** Sector risk is high.	High	$\mu_{\text{high}}(y)$
R6	Emotions are high **AND** People are many.	High	$\mu_{\text{high}}(y)$
R7	Stress is medium **AND** Emotions are medium **AND** People are moderate.	Medium	$\mu_{\text{medium}}(y)$
R8	Sector risk is high **AND** (Emotions are medium **OR** Stress is medium).	Medium	$\mu_{\text{medium}}(y)$
R9	Ambient noise is high **AND** People are many.	Medium	$\mu_{\text{medium}}(y)$
R10	Emotions are low **AND** Stress is low **AND** Weapons are undetected **AND** Sector risk is low.	Low	$\mu_{\text{low}}(y)$

(Continued)

Table A8 (continued)

Rule	Linguistic Antecedent (Min/Max Logic)	Consequent	Output Set
R11	People are few **AND** Weapons are undetected **AND** (Emotions are low **OR** Stress is low).	Low	$\mu_{\text{low}}(y)$
R12	Sector risk is medium **AND** Weapons are undetected **AND** People are moderate **AND** Emotions are medium.	Medium	$\mu_{\text{medium}}(y)$

Note: "People few/moderate/many" denotes the relative number of individuals observable in the forward camera sector at the current decision point, not the total size of the surrounding crowd. Likewise, "Weapons detected" denotes camera-visible weapon-like evidence within that same sector only. Continuous antecedents are fuzzified into overlapping low/medium/high memberships, conjunctive links are evaluated with the bounded-product t-norm, disjunctive links use the maximum operator, and class supports are normalized to obtain the posterior triplet reported by FLAD.

Table A9: Baseline tuning budgets, candidate hyperparameter sets, and final configurations. This table reports the comparator settings used in the pilot analysis and is intended to show that all non-fuzzy baselines were tuned and calibrated under matched, validation-only procedures, with like-for-like search budgets across models.

Baseline	Tuning Budget	Candidate Hyperparameter Set to Report	Final Setting Used in the Manuscript
Logistic Regression (cal.)	Validation-only grid search within each outer LOSO fold; post-hoc isotonic calibration fitted on validation only.	Solver $\in$ {`lbfgs`, `liblinear`}; regularization strength $C \in \{0.01, 0.1, 1, 10\}$; penalty $\in \{\ell_2\}$; max_iter = 1000.	`lbfgs`, C = 1.0, ℓ_2 penalty, max_iter = 1000.
Gradient Boosting (cal.)	Validation-only grid search within each outer LOSO fold; post-hoc isotonic calibration fitted on validation only.	n_estimators $\in \{100, 200, 300\}$; learning_rate $\in \{0.01, 0.05, 0.10\}$; max_depth $\in \{2, 3, 4\}$; subsample $\in \{0.8, 1.0\}$.	n_estimators = 200, learning_rate = 0.05, max_depth = 3, subsample = 0.8.
Random Forest (cal.)	Validation-only grid search within each outer LOSO fold; post-hoc isotonic calibration fitted on validation only.	n_estimators $\in \{200, 400, 600\}$; max_depth $\in \{10, 20,$`None`$\}$; max_features $\in \{$`sqrt`$,0.5\}$; min_samples_leaf $\in \{1, 2, 4\}$.	n_estimators = 400, max_depth = 20, max_features = `sqrt`, min_samples_leaf = 2.

(Continued)

Table A9 (continued)

Baseline	Tuning Budget	Candidate Hyperparameter Set to Report	Final Setting Used in the Manuscript
Non-Fuzzy Aggregator (MLP, cal.)	Validation-only search within each outer LOSO fold; post-hoc isotonic calibration fitted on validation only.	Hidden layer size(s) $\in \{(32,), (64,), (64, 32)\}$; weight decay/regularization $\alpha \in \{10^{-5}, 10^{-4}, 10^{-3}\}$; learning_rate_init $\in \{10^{-4}, 10^{-3}, 10^{-2}\}$; early stopping = `True` with patience = 10.	Hidden layers $= (64, 32)$, $\alpha = 10^{-4}$, learning_rate_init $= 10^{-3}$, early stopping = `True` with patience = 10.

Note: All baselines use the same subject-disjoint outer LOSO folds as ADPS; no baseline accesses the held-out participant during tuning, calibration, or threshold selection. Candidate grids are shown to document matched tuning budgets across comparators.

Appendix A.2

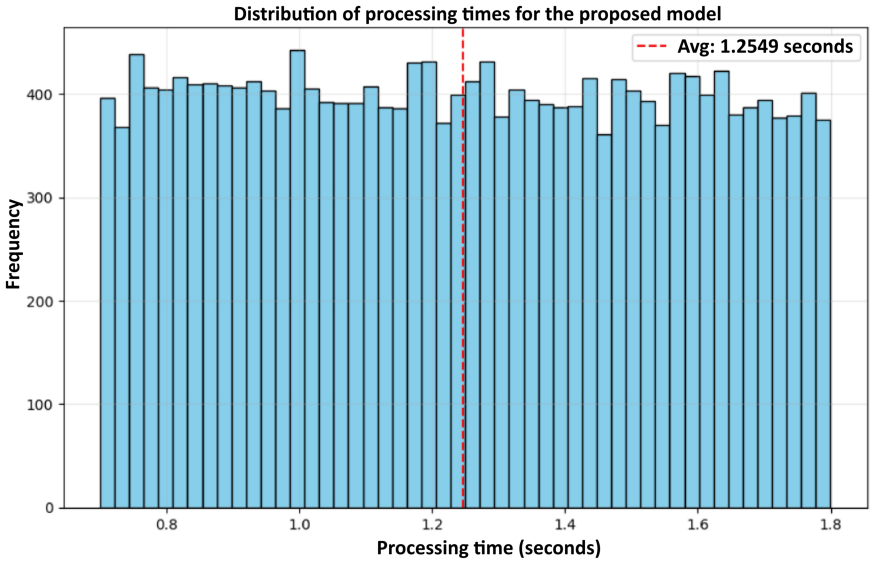
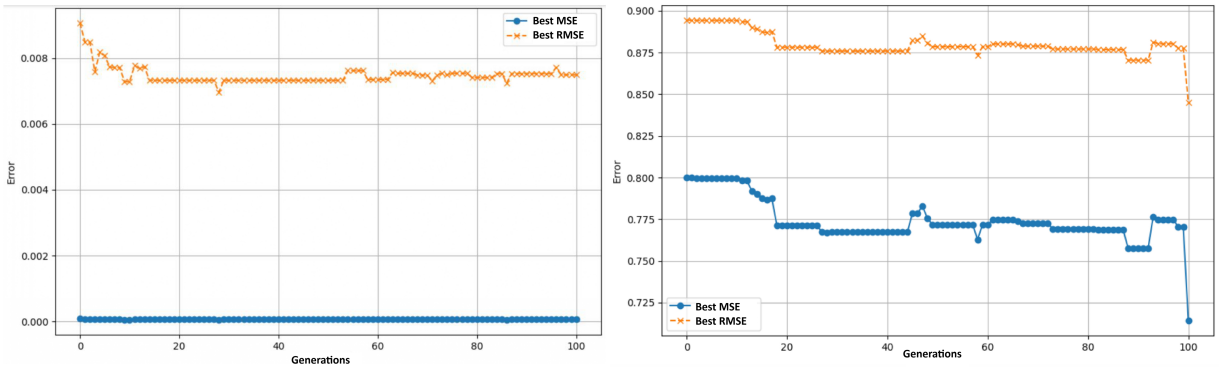


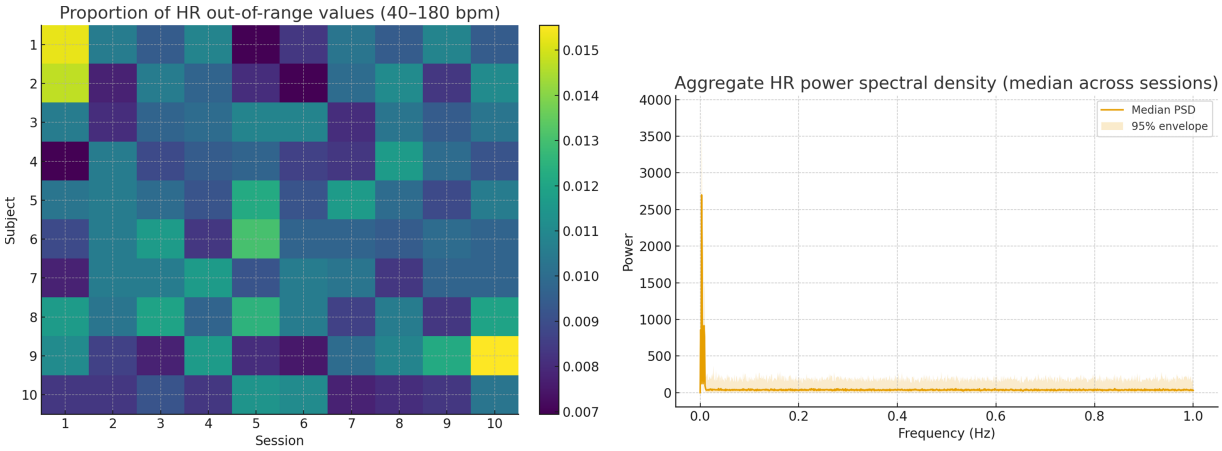
Figure A1: End-to-end processing-time distribution on Raspberry Pi 3 Model B+ for the held-out episode-level records aggregated across the 10 LOSO folds (N = 20,000). The main takeaway is that most records fall within a relatively narrow latency band around the reported median, with only a modest right tail toward the P95 summarized in Table A5.



a) Emotion-rate LSTM training loss (per-generation min/avg)

b) Heart-rate LSTM training loss (per-generation min/avg).

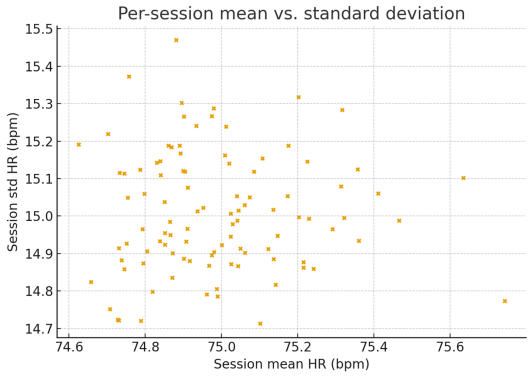
Figure A2: LBPS optimization traces under GA hyperparameter search for the two forecasting tasks. (**a**) Shows the emotion-rate LSTM training-loss trajectory across generations, and (**b**) shows the corresponding heart-rate LSTM training-loss trajectory. The evaluation unit in this display is the candidate model state across the GA search rather than held-out episode windows. The main takeaway is that both optimization traces stabilize under the selected search budget, but this rapid stabilization should be interpreted as numerically stable pilot-scale convergence of compact forecasters under limited data, not as evidence that training-data sufficiency for deployment has already been established; the final forecasting settings are reported in Table A7.



a) Out-of-range proportion per subject--session.

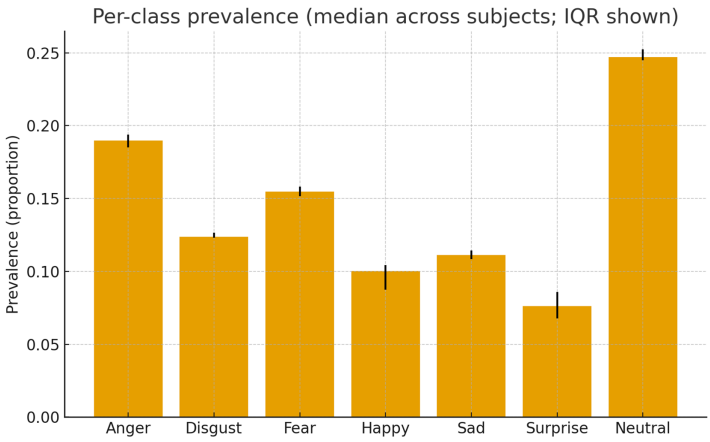
b) Aggregate power spectral density.

Figure A3: (Continued)

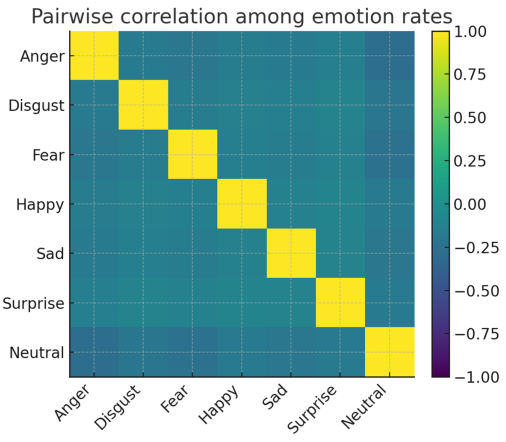


c) Per-session mean vs. standard deviation.

Figure A3: Cardiac signal quality and spectral characterization across subject–session summaries from the pilot heart-rate dataset. (**a**) Reports the proportion of HR samples outside the physiological range $[40, 180]$ bpm for each subject–session pair; (**b**) shows the median heart-rate power spectral density across sessions with a shaded 95% envelope at 4 Hz sampling; and (**c**) plots the session-wise mean vs. standard deviation of HR. The main takeaway is that out-of-range values are uniformly infrequent, spectral mass is concentrated at low frequencies without evident acquisition-induced periodic artifacts, and session-level dispersion remains compact, supporting the use of short-horizon sequential predictors and subject-wise evaluation.

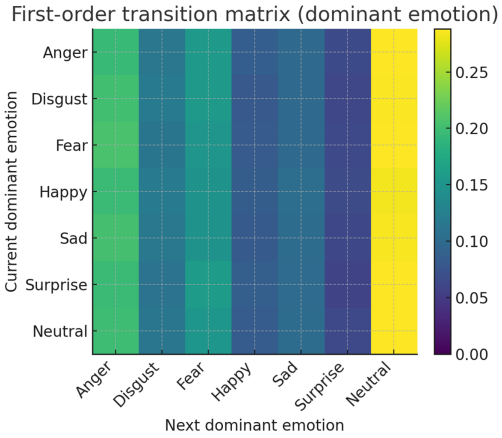


a) Per-class prevalence (median; IQR).



b) Pairwise correlation among classes.

Figure A4: (Continued)



c) First-order transition matrix.

Figure A4: Emotion-rate dataset summary based on subject/session-aggregated statistics from the pilot cohort. (**a**) Shows per-class prevalence summarized by subject-wise medians with interquartile ranges (IQR); (**b**) reports pairwise correlations among the emotion-rate channels; and (**c**) displays the first-order transition probabilities between dominant emotions (argmax per time step). The main takeaway is that neutral and negative-affect states dominate the marginal distribution, cross-channel correlations are generally weak, and dominant-emotion states revert quickly, which is consistent with short-lived affective episodes and motivates short-horizon temporal forecasting.

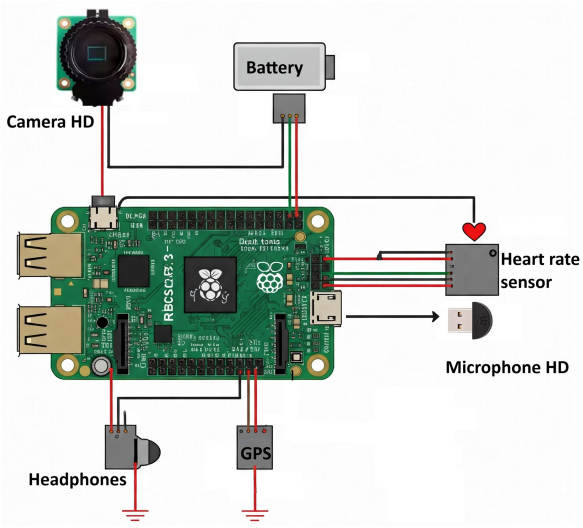


Figure A5: Embedded tactical device hardware configuration of the pilot ADPS prototype. The evaluation unit in this display is the physical system layout rather than held-out episode windows. The figure identifies the sensing and support components integrated in the portable setup: (1) camera, (2) heart-rate sensor, (3) microphone, (4) headphones, (5) battery, and (6) Raspberry Pi 3 Model B+. The main takeaway is that the prototype combines visual, physiological, and audio sensing with compact on-device computing in a single field-deployable configuration.

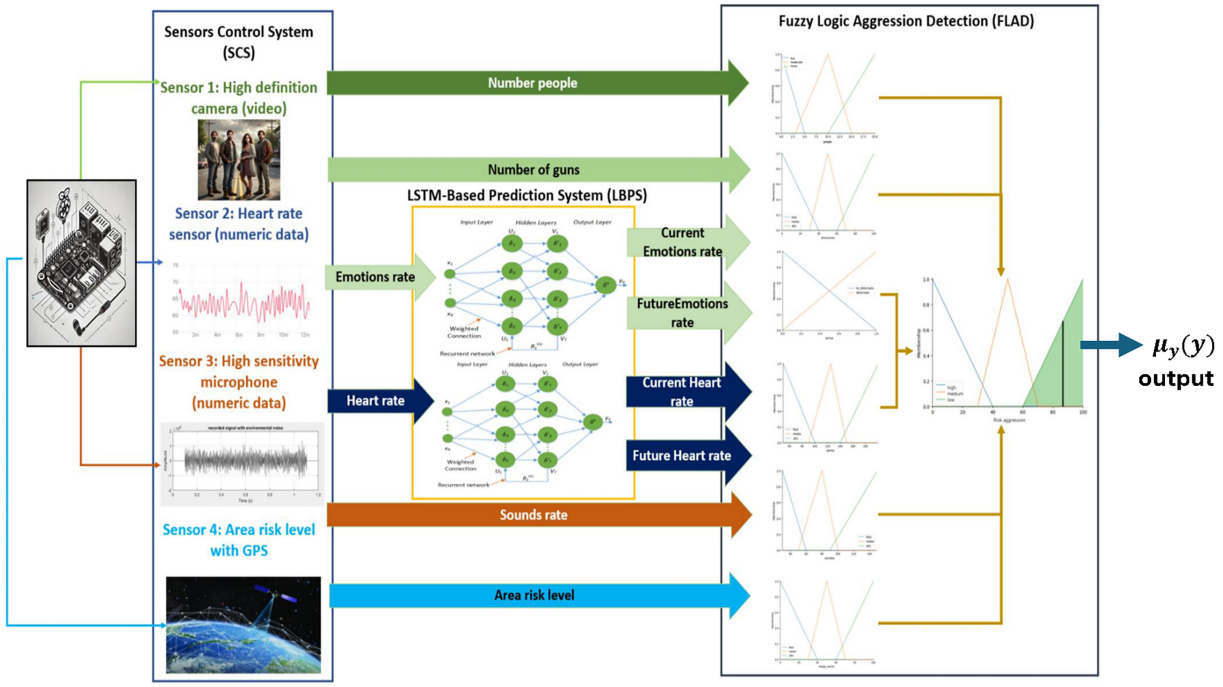


Figure A6: High-level architecture of the aggression detection–prediction system (ADPS). The evaluation unit in this display is the module/interface structure of the system rather than held-out episode windows. The schematic shows the information flow across the three main subsystems: (1) the sensor control system (SCS), which standardizes multimodal inputs; (2) the LSTM-based prediction system (LBPS), which forecasts short-horizon affective and physiological trajectories; and (3) the fuzzy logic aggression detection layer (FLAD), which aggregates current and forecasted cues into class-posterior risk outputs. The main takeaway is that ADPS is a modular pipeline in which standardized sensing, temporal forecasting, and interpretable fuzzy fusion are explicitly separated and auditable.

Appendix A.3 Hypothesis Tests and Multiplicity Control

Appendix A.3.1 Inter-Rater Reliability and Label-Noise Diagnostics

Setup and notation.

Let N denote annotated items (episodes), $K = 3$ the ordered categories (Low, Medium, High), and R the number of raters. For item i, $n_{i,c}$ is the count of ratings in category c and the per-item category proportion is

$$p_{i,c} = \frac{n_{i,c}}{R}, \qquad \bar{p}_c = \frac{1}{N}\sum_{i=1}^{N} p_{i,c} \tag{A1}$$

is the marginal prevalence of category c.

Quadratic weighted Cohen's $\boldsymbol{\kappa_w}$ (pairwise).

For two raters, define the quadratic weights

$$w_{ab} = \frac{(a-b)^2}{(K-1)^2}, \qquad a, b \in \{1, \ldots, K\}, \tag{A2}$$

and the observed and expected weighted disagreements

$$D_{\text{obs}} = \sum_{a=1}^{K}\sum_{b=1}^{K} w_{ab}\, p_{ab}, \tag{A3}$$

$$D_{\text{exp}} = \sum_{a=1}^{K}\sum_{b=1}^{K} w_{ab}\, p_{a\cdot}\, p_{\cdot b}, \tag{A4}$$

where p_{ab} is the empirical joint proportion and $p_{a\cdot}$, $p_{\cdot b}$ the marginals. The weighted Cohen's kappa is then

$$\kappa_w = 1 - \frac{D_{\text{obs}}}{D_{\text{exp}}}. \tag{A5}$$

For $R > 2$, we report the mean of all pairwise κ_w values (with bootstrap CIs).

Krippendorff's $\boldsymbol{\alpha}$ (ordinal, $R \geq 2$).

Let the ordinal distance be

$$\delta_{ab} = \frac{(a-b)^2}{(K-1)^2}. \tag{A6}$$

The observed and expected disagreements are

$$D_o = \frac{1}{\sum_{i=1}^{N} n_i(n_i-1)} \sum_{i=1}^{N}\sum_{a<b} 2\,\delta_{ab}\, n_{i,a}\, n_{i,b}, \tag{A7}$$

$$D_e = \frac{\sum_{a<b} 2\,\delta_{ab}\,\bar{n}_a\,\bar{n}_b}{\left(\sum_{c=1}^{K}\bar{n}_c\right)\left(\sum_{c=1}^{K}\bar{n}_c - 1\right)}, \qquad \bar{n}_c = \sum_{i=1}^{N} n_{i,c}, \tag{A8}$$

and Krippendorff's alpha is

$$\alpha = 1 - \frac{D_o}{D_e}. \tag{A9}$$

Fleiss' $\boldsymbol{\kappa}$ (multi-rater, nominal/ordinal).

Define per-item agreement and its averages as

$$P_i = \frac{1}{R(R-1)} \sum_{c=1}^{K} n_{i,c}\left(n_{i,c}-1\right), \tag{A10}$$

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i, \qquad P_e = \sum_{c=1}^{K} \bar{p}_c^{\,2}, \tag{A11}$$

so that

$$\kappa_{\text{Fleiss}} = \frac{\bar{P} - P_e}{1 - P_e}. \tag{A12}$$

Confidence intervals and diagnostics.

Two-sided 95% CIs for κ_w, α, and κ_{Fleiss} are obtained via the moving-block bootstrap (MBB) over temporally ordered items (see Appendix A.3.2). We also report per-class disagreement and adjudication rates, and rater-pair swap diagnostics.

Appendix A.3.2 Temporal Dependence and Uncertainty Quantification

Integrated autocorrelation time and effective sample size.

For a temporally indexed series $\{Z_t\}_{t=1}^{T}$ with autocorrelation $\rho(h)$ at lag h, the integrated autocorrelation time is estimated as

$$\widehat{\tau}_{\text{int}} \;=\; 1 + 2\sum_{h=1}^{H^\star} \rho(h), \tag{A13}$$

where $H^\star$ is the first index beyond which the truncated sum is non-increasing (initial monotone sequence rule). We set the bootstrap block length and effective sample size as

$$B = \left\lceil \widehat{\tau}_{\text{int}} \right\rceil, \qquad n_{\text{eff}} \;\approx\; \frac{N}{B}. \tag{A14}$$

Moving-block bootstrap (MBB) for confidence intervals.

Given N ordered units (episodes), construct overlapping blocks of length B,

$$\mathcal{B}_s = (s,\, s+1,\, \ldots,\, s+B-1), \qquad s = 1, \ldots, N-B+1, \tag{A15}$$

and draw

$$m \;=\; \left\lceil \frac{N}{B} \right\rceil \tag{A16}$$

blocks independently with replacement to form one bootstrap resample of length N. Recompute the target statistic T^* (e.g., Macro-F_1, AUROC, AUPRC, Brier, ECE, κ) on each resample; the 2.5/97.5 percentiles across R replicates (e.g., $R = 1000$) define the 95% CI.

Calibration metrics (multiclass; percentage scaling).

For class probabilities $\hat{p}_{i,c}$ and one-hot labels $y_{i,c}$, the (percentage-scaled) Brier score is

$$\text{Brier} \;=\; \frac{100}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} \left(\hat{p}_{i,c} - y_{i,c}\right)^2 \;\; (\%), \tag{A17}$$

and the Expected Calibration Error (ECE) with confidence bins $\{B_m\}_{m=1}^{M}$ is

$$\text{ECE} \;=\; 100 \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \;\; (\%), \tag{A18}$$

where $\text{acc}(B_m)$ is empirical accuracy and $\text{conf}(B_m)$ is mean predicted probability within bin m (Micro/macro averaging and the one-vs.-rest convention are defined in Section 5).

Appendix A.3.3 Runtime Instrumentation and Summary Functionals

Definition of end-to-end latency.

End-to-end latency is the elapsed time from the first sensor timestamp entering the pipeline to the emission of the calibrated risk output:

$$t_{\text{e2e}} \;=\; t_{\text{out}} \;-\; t_{\text{in}}. \tag{A19}$$

Stage-wise decomposition.

For record i, let $L_{i,j}$ denote the measured latency of stage j in the pipeline (capture/pre-proc, person detection, weapon cues, facial emotion, LSTM modules, FLAD, I/O). The per-record total processing time is

$$T_i = \sum_{j=1}^{J} L_{i,j}. \tag{A20}$$

Percentile summary.

Let $\widehat{F}_T$ be the empirical CDF of $\{T_i\}_{i=1}^{N}$. The empirical p-quantile is defined as

$$\widehat{Q}_p = \inf\{\, t \in \mathbb{R} : \widehat{F}_T(t) \geq p \,\}, \qquad p \in (0,1), \tag{A21}$$

so that the reported 95th percentile is $\widehat{Q}_{0.95}$. Numerical summaries (median, $\widehat{Q}_{0.95}$, range) are tabulated per stage in Table A5, while the distribution of T_i appears in Fig. A1. For consistency with the rest of the study, resampling-based uncertainty—when reported—follows the moving-block bootstrap conventions in Appendix A.3.2 (Eqs. (A14)–(A16)), with preprocessing and calibration mappings held fixed within replicates.

Appendix A.3.4 Hypothesis Tests and Multiplicity Control

DeLong's paired AUROC test.

Let $\widehat{A}_1$ and $\widehat{A}_2$ be the correlated AUROC estimates on the same cases. Using DeLong's covariance estimate $\widehat{\mathrm{Var}}(\widehat{A}_k)$ and $\widehat{\mathrm{Cov}}(\widehat{A}_1, \widehat{A}_2)$, the standardized statistic is

$$z = \frac{\widehat{A}_1 - \widehat{A}_2}{\sqrt{\widehat{\mathrm{Var}}(\widehat{A}_1) + \widehat{\mathrm{Var}}(\widehat{A}_2) - 2\,\widehat{\mathrm{Cov}}(\widehat{A}_1, \widehat{A}_2)}}. \tag{A22}$$

McNemar's test with continuity correction.

For paired binary decisions at a fixed threshold, let b and c be the discordant counts. The continuity-corrected statistic is

$$\chi^2_{\mathrm{cc}} = \frac{\big(|b-c|-1\big)^2}{b+c}, \tag{A23}$$

and significance is assessed against the χ^2_1 reference using the critical-value rule $\chi^2_{\mathrm{cc}} \geq \chi^2_{1,1-\alpha}$. We report the test statistic and two-sided 95% confidence intervals for paired outcome differences.

Holm step-down adjustment.

For m hypotheses with individual marginal tests ordered by increasing evidence against the null (index (1) most extreme), familywise error control at level α proceeds stepwise: reject $H_{(k)}$ whenever

$$p_{(k)} \leq \frac{\alpha}{m-k+1}, \tag{A24}$$

and continue sequentially to the next index until the first non-rejection, after which all remaining hypotheses are retained.

References

1. Jaafar N, Lachiri Z. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. Expert Syst Appl. 2023;211(4):118523. doi:10.1016/j.eswa.2022.118523.
2. Zawad MRS, Rony CSA, Haque MY, Al Banna MH, Mahmud M, Kaiser MS. A hybrid approach for Stress prediction from Heart rate variability. In: Frontiers of ICT in healthcare. Singapore: Springer; 2023. p. 111–21. doi:10.1007/978-981-19-5191-6_10.
3. Velmovitsky PE, Alencar P, Leatherdale ST, Cowan D, Morita PP. Using apple watch ECG data for heart rate variability monitoring and stress prediction: a pilot study. Front Digit Health. 2022;4:1058826. doi:10.3389/fdgth.2022.1058826.
4. Verma H, Kumar N, Sharma YK, Vyas P. Stress detect. In: Optimized predictive models in healthcare using machine learning. Hoboken, NJ, USA: John Wiley & Sons, Inc; 2024. doi:10.1002/9781394175376.ch20.
5. Sangiorgio M, Dercole F. Robustness of LSTM neural networks for multi-step forecasting of chaotic time series. Chaos Solitons Fractals. 2020;139(8):110045. doi:10.1016/j.chaos.2020.110045.
6. Seng D, Zhang Q, Zhang X, Chen G, Chen X. Spatiotemporal prediction of air quality based on LSTM neural network. Alex Eng J. 2021;60(2):2021–32. doi:10.1016/j.aej.2020.12.009.
7. Fang Z, Wang Y, Peng L, Hong H. Predicting flood susceptibility using LSTM neural networks. J Hydrol. 2021;594:125734. doi:10.1016/j.jhydrol.2020.125734.
8. Yaqub M, Asif H, Kim S, Lee W. Modeling of a full-scale sewage treatment plant to predict the nutrient removal efficiency using a long short-term memory (LSTM) neural network. J Water Process Eng. 2020;37:101388. doi:10.1016/j.jwpe.2020.101388.
9. Farhi N, Kohen E, Mamane H, Shavitt Y. Prediction of wastewater treatment quality using LSTM neural network. Environ Technol Innov. 2021;23(2):101632. doi:10.1016/j.eti.2021.101632.
10. Keerthana K, Yamini R, Dhesigan N, Gangadharan NB, Kirubha SA. Smart lifeguarding vest for military purpose. In: Proceedings of the 2020 International Conference on Communication and Signal Processing (ICCSP); 2020 Jul 28–30; Chennai, India. p. 637–9. doi:10.1109/iccsp48568.2020.9182321.
11. Haider AU, Khan S, Ahmed MJ, Ali Khan T. Strip pooling coordinate attention with directional learning for intelligent fire recognition in smart cities. ICCK Trans Sens Commun Control. 2025;2(4):263–75. doi:10.62762/tscc.2025.675097.
12. Huang L, Li S, Man Y, Wang X, Tang X, Ji R. Fatigue driving detection via multi-head transformer with adaptive weighted loss. ICCK Trans Intell Syst. 2026;3(1):55–69. doi:10.62762/tis.2025.633754.
13. Wang CY, Bochkovskiy A, Liao HM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. doi:10.1109/cvpr52729.2023.00721.
14. Jain A, Aishwarya, Garg G. Gun detection with model and type recognition using HaaR cascade classifier. In: Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT); 2020 Aug 20–22; Tirunelveli, India. doi:10.1109/icssit48917.2020.9214211.
15. Senst T, Eiselein V, Kuhn A, Sikora T. Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation. IEEE Trans Inf Forensics Secur. 2017;12(12):2945–56. doi:10.1109/tifs.2017.2725820.
16. Wickert M. Real-time digital signal processing using pyaudio_helper and ipywidgets. In: Proceedings of the 17th Python in Science Conference; 2018 Jul 9–15; Austin, TX, USA. doi:10.25080/majora-4af1f417-00e.
17. Ribeiro PC, Audigier R, Pham QC. RIMOC, a feature to discriminate unstructured motions: application to violence detection for video-surveillance. Comput Vis Image Underst. 2016;144(15):121–43. doi:10.1016/j.cviu.2015.11.001.
18. Bernal E, Lagunes ML, Castillo O, Soria J, Valdez F. Optimization of type-2 fuzzy logic controller design using the GSO and FA algorithms. Int J Fuzzy Syst. 2021;23(1):42–57. doi:10.1007/s40815-020-00976-w.
19. Woźniak M, Zielonka A, Sikora A. Driving support by type-2 fuzzy logic control model. Expert Syst Appl. 2022;207(3):117798. doi:10.1016/j.eswa.2022.117798.
20. Malik S, Mohan BM. Development and experimental validation of analytical structures of some simplest fuzzy PI/PD controllers using bounded sum aggregation. J Frankl Inst. 2024;361(15):107098. doi:10.1016/j.jfranklin.2024.107098.