Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A confidence-convergence decision layer for reliability-aware sequential recognition under limited observations

Guangyu Luo

luoguangyu@whut.edu.cn

Article

Keywords: Sequential recognition, Confidence convergence, Underwater sensing, Limited observations, Reliability-aware decision making, Uncertainty

Posted Date: July 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10101351/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A confidence-convergence decision layer for reliability-aware sequential recognition under limited observations

Guangyu. Luo.[1,*]

[1]School of Information, Wuhan University of Technology, Wuhan, China

*luoguangyu@whut.edu.cn

Abstract

Underwater moving target behavior recognition is challenging when observations arrive sequentially, are uncertain, and cover only a limited time window. Many methods optimize final classification accuracy, but they rarely assess whether an intermediate prediction is reliable enough to report. This study proposes a confidence-convergence-based sequential recognition method for underwater moving target behaviors under limited observations. The method updates the behavior-type confidence distribution as each observation arrives and accepts a decision only when dominance, temporal stability, and inter-class separation are jointly satisfied. Otherwise, it continues observing or reports non-convergence when evidence is insufficient. A simulation-based dataset was constructed with three representative behavior types, multiple observation-window lengths, and uncertain sensing conditions, including noise, missing data, and delayed measurements. Under the standard setting, the method improved accepted-decision accuracy by 15.4 and 9.2 percentage points over the static classifier and threshold-only baseline, respectively, and reduced the wrong accepted-decision rate by 16.3 and 9.0 percentage points. It also outperformed the strongest abstention-capable baseline in accepted-decision accuracy and risk. These results support confidence convergence, rather than confidence magnitude alone, as a reliable basis for sequential underwater behavior recognition.

Keywords: Sequential recognition; Confidence convergence; Underwater sensing; Limited observations; Reliability-aware decision making; Uncertainty

1 Introduction

Sequential recognition under uncertain observations is a broad problem in engineering sensing systems. A sensor stream often arrives gradually, is affected by missing or delayed measurements, and may need to support decisions before a complete observation record is available. Underwater sensing is a representative setting because acoustic and motion-related observations are affected by propagation uncertainty, multipath effects, sensor noise,

limited bandwidth, and temporary data loss [1, 2, 3, 4]. In these conditions, a recognition system must infer the current behavior and decide whether the evidence is reliable enough for output.

Many underwater recognition models are designed for fixed-window classification. They process an acoustic, motion, or time-frequency segment and output a final label after the input window has been observed [5, 6, 7, 8, 9, 10]. Recent studies have improved this final-label setting through frequency-channel optimization, compact one-dimensional networks, pre-trained transformers, convolution-transformer cooperation, two-stream temporal-frequency modelling, few-shot learning with unlabeled samples, spectral attention, and Wasserstein spectral features [11, 12, 13, 14, 15, 16, 17, 18]. These advances strengthen recognition after a chosen input segment has been processed, but they do not by themselves determine whether an intermediate prediction should be accepted. Sequential testing, hidden Markov models, partially observable decision processes, and selective prediction address related questions in broader statistical learning and decision theory [19, 20, 21, 22, 23, 24]. However, a single posterior or softmax confidence value can still be misleading when observations are noisy or poorly separated.

Confidence magnitude alone is an incomplete reliability criterion. Modern classifiers can be poorly calibrated, especially under class imbalance or distribution shift, and high-confidence predictions can occur before the underlying evidence has stabilized [25, 26, 27, 28]. In a sequential stream, a temporary confidence peak may come from a short-lived observation fluctuation, a missing feature, a delayed measurement, or weak separation between two behavior types. Accepting such a prediction can produce a wrong accepted decision even when a later observation would have clarified the behavior. A reliability-aware system therefore needs to consider how confidence evolves, not only how large it is at one step.

This study uses confidence convergence as a decision condition for sequential recognition under limited observations. The method updates a behavior-type confidence distribution as observations arrive and accepts a behavior label only when three conditions are jointly satisfied: dominance of the leading behavior type, temporal stability of the confidence trajectory, and inter-class separation from the second most likely behavior type. If these criteria are not met, the model continues observing or reports non-convergence at the end of the available window. This non-convergence output is treated as a reliability state, not as a forced classification error.

This work contributes a reliability-aware decision layer that can be placed above statistical or learned evidence models. It provides a limited-observation sequential recognition formulation, a confidence-convergence decision rule combining dominance, temporal stability, and inter-class separation, and a reproducible simulation-based evaluation. The evaluation uses controlled underwater sensing scenarios with noise, missing observations, and bounded measurement delay. It reports accepted-decision accuracy, wrong accepted-decision rate, decision time, coverage-risk behavior, ablation results, robustness, and sensitivity tests. Detailed settings and full numerical tables are provided in the Supplementary Information.

2 Results

2.1 Overall performance under limited observations

Under the standard limited-observation setting, the proposed method improved accepted-decision reliability while avoiding full-window waiting. Figure 1 gives the method overview, and Fig. 2 summarizes the main comparison across forced-output, selective, temporal, and confidence-convergence baselines. The standard setting used three behavior types, a complete horizon of $T = 100$, an available observation window of $T_w = 40$, six observation features, medium observation noise, a 10% missing-observation ratio, and a maximum delay of one time step. Full numerical values are reported in Supplementary Table S5.

Accepted-decision accuracy reached 91.2% for the proposed method. This was 15.4 percentage points higher than the static classifier and 9.2 percentage points higher than the threshold-only sequential baseline. The method also reduced the wrong accepted-decision rate to 7.9%, compared with 24.2% for the static classifier and 16.9% for the threshold-only baseline. Relative to full-window forced recognition, the mean decision time was shortened by 26.5% while maintaining a non-convergence state for insufficient evidence.

The strongest abstention-capable baseline was the early-exit GRU. The proposed method improved accepted-decision accuracy by 3.3 percentage points over this baseline and reduced accepted-decision risk by 3.2 percentage points. Overall accuracy was slightly lower than the forced-output GRU because non-converged cases were not assigned a forced label. This trade-off is expected for reliability-aware recognition: the method rejects or postpones ambiguous cases rather than treating all samples as accepted outputs.

The reliability-timeliness panel in Fig. 2 shows that the threshold-only baseline made the earliest decisions, but at higher risk. The proposed method made decisions later than the threshold-only baseline, slightly earlier than early-exit GRU, and earlier than full-window forced baselines. Its operating point combined lower wrong accepted-decision rate with partial-window decision time.

2.2 Coverage-risk analysis

The coverage-risk analysis tested whether the proposed method simply improved accuracy by rejecting more samples. Figure 3 shows accepted-decision risk as acceptance strictness was varied. Coverage is the proportion of test episodes with an accepted recognition output, and accepted-decision risk is $100 - A_{\mathrm{lw}}$, where A_{lw} is accepted-decision accuracy.

As coverage increased, all abstention-capable methods accepted more difficult samples and their accepted-decision risk rose. Across comparable coverage values, the proposed method remained below the threshold-only, calibrated selective, and early-exit GRU curves. At the standard operating point near 90% coverage, the accepted-decision risk was 8.8%, below the strongest abstention-capable baseline at similar coverage. This indicates that the convergence criteria did more than reject uncertain samples; they selected accepted decisions with lower empirical risk.

2.3 Effect of observation-window length

Longer observation windows improved reliability, but the benefit diminished once enough evidence had accumulated. Figure 4 shows the effect of the observation-window ratio $\gamma = T_w/T$ across $\gamma \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. Accepted-decision accuracy increased from 85.0% at $\gamma = 0.2$ to 96.1% at $\gamma = 1.0$. The increase was largest between $\gamma = 0.2$ and $\gamma = 0.4$, whereas the gain from $\gamma = 0.8$ to $\gamma = 1.0$ was smaller.

The wrong accepted-decision rate decreased from 9.3% at $\gamma = 0.2$ to 3.9% at $\gamma = 1.0$. The non-convergence rate also decreased from 38.0% to 1.0%. These trends show that additional observations reduce ambiguity, but they also delay decisions. Even when the full horizon was available, the mean decision time was 50.8, well below $T = 100$. The algorithm therefore did not simply wait until the end of the sequence; it stopped when the confidence distribution satisfied the convergence conditions. Full window-length results are given in Supplementary Table S6.

2.4 Confidence evolution and convergence behavior

Representative confidence trajectories explain how the decision layer behaves in different evidence conditions. Figure 5 shows correct convergence, delayed convergence, and non-convergence cases, together with binary maps for dominance, temporal stability, inter-class separation, label consistency, and the all-criteria trigger.

In the correct-convergence case, the true behavior type became dominant and stable by $\tau = 21$. The confidence margin over the second-ranked behavior type also exceeded the separation threshold, so the model accepted the correct label before the available window ended. In the delayed-convergence case, the leading behavior type changed during the early and middle steps. Although maximum confidence temporarily exceeded the dominance threshold, stability and separation were not satisfied together. The model continued observing and accepted the correct label at $\tau = 32$.

The non-convergence case showed persistent ambiguity. Two behavior types remained close, the dominant label switched repeatedly, and the separation margin did not remain above the threshold. The model therefore reported non-convergence at $T_w = 40$. This behavior illustrates the central distinction between confidence magnitude and confidence convergence. A large confidence value at one step was not enough; acceptance required a stable and separated confidence trajectory.

2.5 Ablation analysis

The ablation study tested whether dominance, temporal stability, and inter-class separation each contributed to reliability. Figure 6 compares dominance-only recognition, dominance plus stability, dominance plus separation, and the full criterion set with label consistency. Full numerical values are reported in Supplementary Table S7.

Dominance-only recognition made the earliest decisions, with a mean decision time of 22.8, but produced the highest wrong accepted-decision rate of 16.9%. Adding temporal stability reduced this rate to 10.6%, and adding inter-class separation reduced it to 11.2%. The full method achieved the highest accepted-decision accuracy, 91.2%, and the lowest

wrong accepted-decision rate, 7.9%.

The full method increased mean decision time from 22.8 to 29.4 relative to the dominance-only variant. This cost was moderate compared with the 9.0 percentage-point reduction in wrong accepted-decision rate. Expected calibration error also decreased from 0.086 for the dominance-only variant to 0.044 for the full method. These results support the use of complementary convergence criteria rather than a single confidence threshold.

2.6 Robustness and sensitivity analyses

The method remained more reliable than the threshold-only baseline under degraded sensing conditions. Figure 7 summarizes robustness under observation noise, missing observations, and measurement delay. Under high noise with $\sigma_o = 0.10$, the proposed method reduced the wrong accepted-decision rate from 23.5% to 12.6%. When 30% of observations were missing, accepted-decision accuracy increased from 72.5% to 83.5%, and the wrong accepted-decision rate decreased from 27.4% to 15.1%. When the maximum delay was three time steps, the wrong accepted-decision rate decreased from 26.0% to 14.9%.

The non-convergence rate increased under severe degradation, which is consistent with the reliability-aware design. Samples with unstable or weakly separated confidence distributions were not forced into accepted labels. Additional sensitivity tests are provided in Supplementary Table S8. Laplace noise with matched variance, mixed Gaussian outliers, and imbalanced class priors all reduced performance relative to the Gaussian reference, but wrong accepted-decision rates remained below the corresponding threshold-only behavior.

3 Discussion

The experiments suggest that confidence convergence can improve reliability in sequential recognition under uncertain observations. The method helped because it evaluated the trajectory of the confidence distribution rather than a single confidence value. Dominance required sufficient support for one behavior type, temporal stability reduced the effect of transient confidence peaks, and inter-class separation reduced acceptance when competing behaviors remained close. The ablation results show that each condition reduced wrong accepted decisions, and the full criterion set gave the best balance between reliability and timeliness.

This decision layer differs from confidence thresholding and forced classification. A threshold-only method accepts a label once maximum confidence is high enough, even if that confidence is temporary or weakly separated. A forced classifier always outputs a label at the end of the available window, even when the evidence remains ambiguous. The proposed method can recognize, continue observing, or report non-convergence. In controlled underwater sensing scenarios, this output structure reduced wrong accepted-decision rate while retaining partial-window decision capability.

Non-convergence should be interpreted as a reliability-preserving output rather than as a simple failure. In a time-constrained sensing system, an unsupported label can be more harmful than an explicit uncertain state. The coverage-risk analysis supports this interpretation because the method achieved lower accepted-decision risk than abstention-

capable baselines at comparable coverage. This suggests that temporal stability and inter-class separation provide information beyond a calibrated single-step confidence threshold.

Several limitations remain. The evaluation is simulation based and demonstrates methodological feasibility rather than operational readiness. Real underwater sequences may involve more complex acoustic propagation, sensor placement effects, target interactions, environmental variation, and data association errors. The simulated dataset used three abstract behavior types, which is useful for controlled comparison but narrower than real marine behavior. The main protocol used balanced classes and Gaussian observation disturbances, although sensitivity tests with non-Gaussian noise, mixed outliers, and imbalanced priors partly address this limitation. The convergence thresholds were fixed after validation and may need adaptation across missions, sensing environments, and risk preferences.

Future work should test the method on real underwater acoustic or motion sequences, extend it to non-Gaussian and non-stationary disturbances, and examine open-set, multi-target, and adaptive-calibration settings. The confidence-convergence layer could also be combined with stronger temporal representation models, provided that the output remains interpretable as a reliability-aware decision state. These results indicate that confidence convergence can provide a practical reliability layer for sequential recognition under uncertain sensing, but validation on real underwater sequences is required before operational deployment.

4 Methods

4.1 Problem formulation

The task was formulated as limited-observation sequential recognition. Let the behavior-type set be

$$\mathcal{C} = \{c_1, c_2, \ldots, c_K\}, \tag{1}$$

with $K = 3$ in the simulation study. At time step t, the model receives an observation vector $o_t \in \mathbb{R}^M$ and has access only to the partial sequence

$$O_{1:t} = \{o_1, o_2, \ldots, o_t\}, \quad 1 \le t \le T_w, \tag{2}$$

where T_w is the available observation-window length and T is the complete horizon. The model estimates a confidence distribution over behavior types,

$$\mathbf{b}_t = [b_t(c_1), b_t(c_2), \ldots, b_t(c_K)], \tag{3}$$

where $b_t(c_i) = P(c_i \mid O_{1:t})$ and $\sum_i b_t(c_i) = 1$. The decision state is

$$d_t \in \{\text{recognize}, \text{continue}, \text{non-convergence}\}. \tag{4}$$

4.2 Confidence updating model

The initial confidence distribution was set to a uniform prior when no class-prior information was available. At time step t, the observation evidence for behavior type c_i was denoted by

$$\ell_t(c_i) = p(o_t \mid c_i). \tag{5}$$

The likelihood can be obtained from a statistical observation model, calibrated classifier, or learned temporal model. The simulation used a Gaussian evidence model unless otherwise stated:

$$\ell_t(c_i) = \frac{1}{\sqrt{(2\pi)^M |\Sigma_i|}} \exp\left[-\frac{1}{2}(o_t - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (o_t - \mu_i)\right], \tag{6}$$

where μ_i and Σ_i are the mean vector and covariance matrix associated with behavior type c_i. The preliminary Bayesian update was

$$\bar{b}_t(c_i) = \frac{\ell_t(c_i) b_{t-1}(c_i)}{\sum_{j=1}^{K} \ell_t(c_j) b_{t-1}(c_j)}. \tag{7}$$

Observation reliability was represented by $\rho_t \in [0, 1]$. The final confidence update was

$$b_t(c_i) = \frac{[\bar{b}_t(c_i)]^{\rho_t} [b_{t-1}(c_i)]^{1-\rho_t}}{\sum_{j=1}^{K} [\bar{b}_t(c_j)]^{\rho_t} [b_{t-1}(c_j)]^{1-\rho_t}}. \tag{8}$$

This update reduces abrupt confidence shifts caused by low-quality observations.

4.3 Confidence convergence criteria

The dominant behavior type and second-ranked behavior type were

$$c_t^* = \arg\max_{c_i \in \mathcal{C}} b_t(c_i), \qquad c_t^{(2)} = \arg\max_{c_i \in \mathcal{C} \setminus \{c_t^*\}} b_t(c_i). \tag{9}$$

Dominance was defined as

$$D_t = b_t(c_t^*), \tag{10}$$

and was satisfied when $D_t \geq \theta_D$. Temporal stability was measured over a window of length L:

$$S_t = \frac{1}{L-1} \sum_{r=t-L+2}^{t} |b_r(c_t^*) - b_{r-1}(c_t^*)|. \tag{11}$$

The stability condition was $S_t \leq \epsilon_S$, together with label consistency across the same window. Inter-class separation was

$$G_t = b_t(c_t^*) - b_t(c_t^{(2)}), \tag{12}$$

and was satisfied when $G_t \geq \theta_G$. The convergence indicator was

$$I_t = \mathbb{I}[D_t \geq \theta_D]\, \mathbb{I}[S_t \leq \epsilon_S]\, \mathbb{I}[G_t \geq \theta_G]\, \mathbb{I}[\mathcal{L}_t = 1], \tag{13}$$

where $\mathcal{L}_t$ denotes label consistency. The standard parameters were $\theta_D = 0.75$, $\epsilon_S = 0.03$, $\theta_G = 0.18$, and $L = 5$.

4.4 Sequential decision algorithm

At each time step, the algorithm updated $\mathbf{b}_t$, evaluated I_t, and determined the decision state. If $I_t = 1$, the model output $\hat{c} = c_t^*$ and recorded the decision time $\tau = t$. If $I_t = 0$ and $t < T_w$, the decision state was continue. If $I_t = 0$ and $t = T_w$, the decision state was non-convergence.

With diagonal covariance likelihoods, the computational cost over a window was $O(T_w(KM + K + L))$. With full covariance Gaussian likelihoods and precomputed covariance inverses, it was $O(T_w(KM^2 + K + L))$. If a neural model supplies the evidence score, the dominant cost is the model forward pass. The memory cost was $O(KL)$ because the algorithm stores the current confidence distribution and a short confidence history.

4.5 Simulation data generation

The simulation generated controlled underwater sensing scenarios with three abstract behavior types: stable motion, search-like variation, and intermittent correction. These classes were not platform-specific target models. They differed in maneuver smoothness, heading-change variability, stage-transition frequency, and intermittent correction behavior. The latent state was

$$s_t = [p_t^x, p_t^y, v_t, \psi_t, a_t, \omega_t], \tag{14}$$

where p_t^x and p_t^y are position components, v_t is speed, ψ_t is heading, a_t is acceleration, and ω_t is heading-change rate. The behavior-dependent state transition was

$$s_{t+1} = g_k(s_t) + \eta_t, \quad c_k \in \mathcal{C}. \tag{15}$$

The observation vector had six normalized indicators:

$$o_t = [q_t^v, q_t^\psi, q_t^\beta, q_t^m, q_t^s, q_t^r], \tag{16}$$

where the terms represent speed trend, heading-change intensity, bearing-rate variation, motion consistency, stage transition, and observation quality. The observation model was

$$o_t = h(s_t) + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_o^2 I). \tag{17}$$

Missing observations followed a Bernoulli feature mask with missing ratio r_m. Missing values were replaced by the most recent available or interpolated estimate. Measurement delay was modelled as $\tilde{o}_t = o_{t-\delta_t}$ with $\delta_t \in \{0, 1, \ldots, \delta_{\max}\}$. To avoid undefined normalization under no delay, the normalized delay term was set to zero when $\delta_{\max} = 0$ and to $\delta_t/\delta_{\max}$ otherwise. The reliability coefficient combined missing data, delay, and noise:

$$\rho_t = \mathrm{clip}(1 - \alpha_m r_t^m - \alpha_d \tilde{\delta}_t - \alpha_n \sigma_o, 0, 1). \tag{18}$$

Each random trial generated 900 class-balanced episodes, with 300 episodes per behavior type. The split was 60% training, 20% validation, and 20% testing, giving 180 test episodes per trial. Results were averaged over 30 independent random trials, corresponding to 5400 test episodes across trials. The complete horizon was $T = 100$, and observation-window ratios were $\gamma = T_w/T \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. The standard setting used $T_w = 40$, $M = 6$, $\sigma_o = 0.06$, $r_m = 0.1$, and $\delta_{\max} = 1$. Detailed settings are listed in Supplementary Tables S1 and S2.

4.6 Baseline methods

The baselines included a static classifier, Bayesian sequential recognition without convergence assessment, threshold-only recognition, selective posterior prediction, calibrated selective prediction, early-exit GRU recognition, HMM-based recognition, and GRU-based recognition. Hyperparameters and acceptance thresholds were selected on validation data within each trial. The baseline definitions are summarized in Supplementary Table S3. The HMM and GRU baselines represent classical probabilistic temporal modelling and learning-based temporal modelling, respectively, while the early-exit baseline reflects recent work on uncertainty-aware early stopping in neural classifiers [20, 29, 30, 31, 32].

4.7 Evaluation metrics

The main metrics were accepted-decision accuracy A_{lw}, all-episode accuracy A_{all}, mean decision time T_{dec}, wrong accepted-decision rate R_{wad}, non-convergence rate R_{nc}, coverage, accepted-decision risk, expected calibration error, and robustness degradation. Accepted-decision accuracy was computed over accepted recognition outputs. Wrong accepted-decision rate measured incorrect labels among accepted outputs. For forced-output baselines, every sample was treated as accepted at the final available step. Non-convergence rate measured the proportion of insufficient-evidence outputs. Metric definitions are listed in Supplementary Table S4.

4.8 Statistical analysis and reproducibility workflow

Reported values are means over 30 independent random trials. Error bars and shaded intervals in figures denote 95% confidence intervals computed as $\bar{x} \pm 1.96 s/\sqrt{N_{\mathrm{trial}}}$, where s is the across-trial standard deviation and $N_{\mathrm{trial}} = 30$. Because all methods were evaluated on matched generated splits within each trial, the analysis emphasizes effect sizes, confidence intervals, and consistency across degradation settings. No significance-star notation was used.

The figure source data are stored as CSV files in `figure_source_data/`. The main figures can be regenerated with `scripts/make_figures.py`; figure-source data can be audited with `scripts/audit_figure_data.py`. Specifically, `overall_comparison.csv` supports Fig. 2, `coverage_risk_curve.csv` supports Fig. 3, `window_length_effect.csv` supports Fig. 4, `confidence_evolution.csv` and `confidence_evolution_criteria.csv` support Fig. 5, `ablation_results.csv` supports Fig. 6, and `robustness_summary.csv` supports Fig. 7. The sensitivity tests are stored in `sensitivity_tests.csv`. The figure-source audit and figure-regeneration workflow were verified with Python 3.12.13 on Windows 11, NumPy 2.3.5, pandas 3.0.1, and Matplotlib 3.11.0.

4.9 Ethics statement

This study did not involve human participants, human data, animals, or animal-derived materials.

4.10 AI-assisted language and formatting

During manuscript preparation, AI-assisted tools were used only for language polishing and LaTeX formatting. The author reviewed and edited all content and takes full responsibility for the manuscript.

Data availability

The data supporting the findings of this study are publicly available in Mendeley Data under the dataset title "A confidence-convergence decision layer for reliability-aware sequential recognition under limited observations", version 1, at https://doi.org/10.17632/2hxbs4gfxz.1. The deposited record contains the figure source data tables, sensitivity-test data, publication-ready figure files, and Supplementary Information files. The source data are provided as CSV files and are mapped to the corresponding main figures and sensitivity analyses.

Code availability

The Python scripts used for figure generation and data auditing are publicly available in the same Mendeley Data record, "A confidence-convergence decision layer for reliability-aware sequential recognition under limited observations", version 1, at https://doi.org/10.17632/2hxbs4gfxz.1. The archived code includes `scripts/make_figures.py`, `scripts/audit_figure_data.py`, and the dependency file `requirements.txt`.

References

[1] Luo, J., Han, Y. & Fan, L. Underwater acoustic target tracking: A review. *Sensors* **18**, 112 (2018).

[2] Stojanovic, M. & Preisig, J. Underwater acoustic communication channels: Propagation models and statistical characterization. *IEEE Communications Magazine* **47**, 84–89 (2009).

[3] Akyildiz, I. F., Pompili, D. & Melodia, T. Underwater acoustic sensor networks: Research challenges. *Ad Hoc Networks* **3**, 257–279 (2005).

[4] Kilfoyle, D. B. & Baggeroer, A. B. The state of the art in underwater acoustic telemetry. *IEEE Journal of Oceanic Engineering* **25**, 4–27 (2000).

[5] Feng, S., Ma, S., Zhu, X. & Yan, M. Artificial intelligence-based underwater acoustic target recognition: A survey. *Remote Sensing* **16**, 3333 (2024).

[6] Chen, J., Han, B., Ma, X. & Zhang, J. Underwater target recognition based on multi-decision lofar spectrum enhancement: A deep-learning approach. *Future Internet* **13**, 265 (2021).

[7] Han, X. C., Ren, C., Wang, L. & Bai, Y. Underwater acoustic target recognition method based on a joint neural network. *PLOS ONE* **17**, e0266425 (2022).

[8] Liu, C., Hong, F., Feng, H. & Hu, M. Underwater acoustic target recognition based on dual attention networks and multiresolution convolutional neural networks. In *OCEANS 2021: San Diego – Porto*, 1–5 (2021).

[9] Xie, Y., Ren, J. & Xu, J. Unraveling complex data diversity in underwater acoustic target recognition through convolution-based mixture of experts. *Expert Systems with Applications* **250**, 123431 (2024).

[10] Xue, L., Zeng, X. & Jin, A. A novel deep-learning method with channel attention mechanism for underwater target recognition. *Sensors* **22**, 5492 (2022).

[11] Zeng, D., Yan, S., Yang, J. & Pan, X. An efficient deep learning approach with frequency and channel optimization for underwater acoustic target recognition. *Scientific Reports* **15**, 27369 (2025). URL https://doi.org/10.1038/s41598-025-12452-2.

[12] Huang, Q. *et al.* Efficient one-dimensional network design method for underwater acoustic target recognition. *Journal of Marine Science and Engineering* **13**, 599 (2025). URL https://doi.org/10.3390/jmse13030599.

[13] Tang, J., Ma, E., Qu, Y., Gao, W., Zhang, Y. & Gan, L. UAPT: an underwater acoustic target recognition method based on pre-trained Transformer. *Multimedia Systems* **31**, 50 (2025). URL https://doi.org/10.1007/s00530-024-01614-3.

[14] Jin, A., Yang, S., Lei, M., Zeng, X. & Wang, H. An effective convolutional and transformer cooperation network for underwater acoustic target recognition. *Engineering Applications of Artificial Intelligence* **159**, 111791 (2025). URL https://doi.org/10.1016/j.engappai.2025.111791.

[15] Bi, D., Zhang, L. & Chen, J. A new Two-Stream Temporal-Frequency transformer network for underwater acoustic target recognition. *Signal Processing* **231**, 109891 (2025). URL https://doi.org/10.1016/j.sigpro.2025.109891.

[16] He, Z. *et al.* Unlabeled samples improve few-shot underwater acoustic target recognition. *IEEE Transactions on Geoscience and Remote Sensing* **63**, 1–16 (2025). URL https://doi.org/10.1109/TGRS.2025.3598380.

[17] Gao, W., Chen, D., Liu, Y., Zhang, X. & Zhang, J. An underwater acoustic target recognition model based on heterogeneous spectral attention feature fusion. *Signal Processing* **246**, 110601 (2026). URL https://doi.org/10.1016/j.sigpro.2026.110601.

[18] Zhang, M. & Lei, B. The wasserstein vector spectrum for underwater acoustic target recognition. *Applied Acoustics* **251**, 111342 (2026). URL https://doi.org/10.1016/j.apacoust.2026.111342.

[19] Wald, A. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics* **16**, 117–186 (1945).

[20] Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286 (1989).

[21] Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* **101**, 99–134 (1998).

[22] Ji, S. & Carin, L. Cost-sensitive feature acquisition and classification. *Pattern Recognition* **40**, 1474–1485 (2007).

[23] Chow, C. K. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* **16**, 41–46 (1970).

[24] Geifman, Y. & El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).

[25] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 1321–1330 (2017). URL https://proceedings.mlr.press/v70/guo17a.html.

[26] Naeini, M. P., Cooper, G. F. & Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29 (2015).

[27] Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).

[28] Dong, J. *et al.* A survey on confidence calibration of deep learning-based classification models under class imbalance data. *IEEE Transactions on Neural Networks and Learning Systems* **36**, 15664–15684 (2025). URL https://doi.org/10.1109/TNNLS.2025.3565159.

[29] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Computation* **9**, 1735–1780 (1997).

[30] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

[31] Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).

[32] Kurian, T. & Thangam, S. Enhancing early exit performance with uncertainty-aware training in convolutional neural networks for image classification. *IEEE Access* **13**, 91671–91681 (2025). URL https://doi.org/10.1109/ACCESS.2025.3572415.

Acknowledgements

The author received no specific assistance requiring acknowledgement.

Author contributions

G.L. conceived the study, developed the method, implemented the simulations, analysed the results, prepared the figures, and wrote and revised the manuscript.

Competing interests

The author declares no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Figure legends

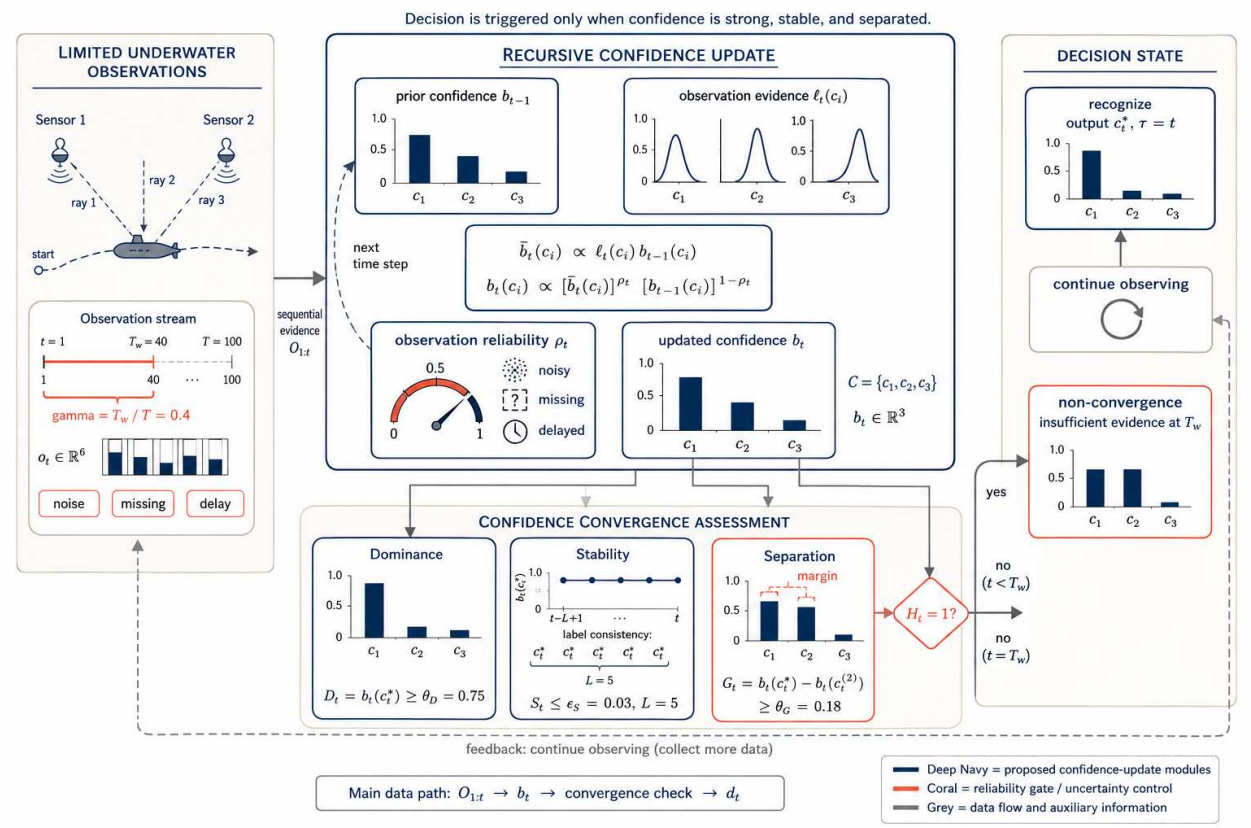


Figure 1: **Confidence-convergence framework for sequential recognition.** The figure shows limited underwater observations entering a recursive confidence update, followed by convergence assessment and one of three decision states: recognize, continue observing, or non-convergence. The confidence distribution $\mathbf{b}_t$ is updated at each time step from the observation evidence and the previous confidence state. Convergence is accepted only when dominance, temporal stability, inter-class separation, and label consistency are jointly satisfied.

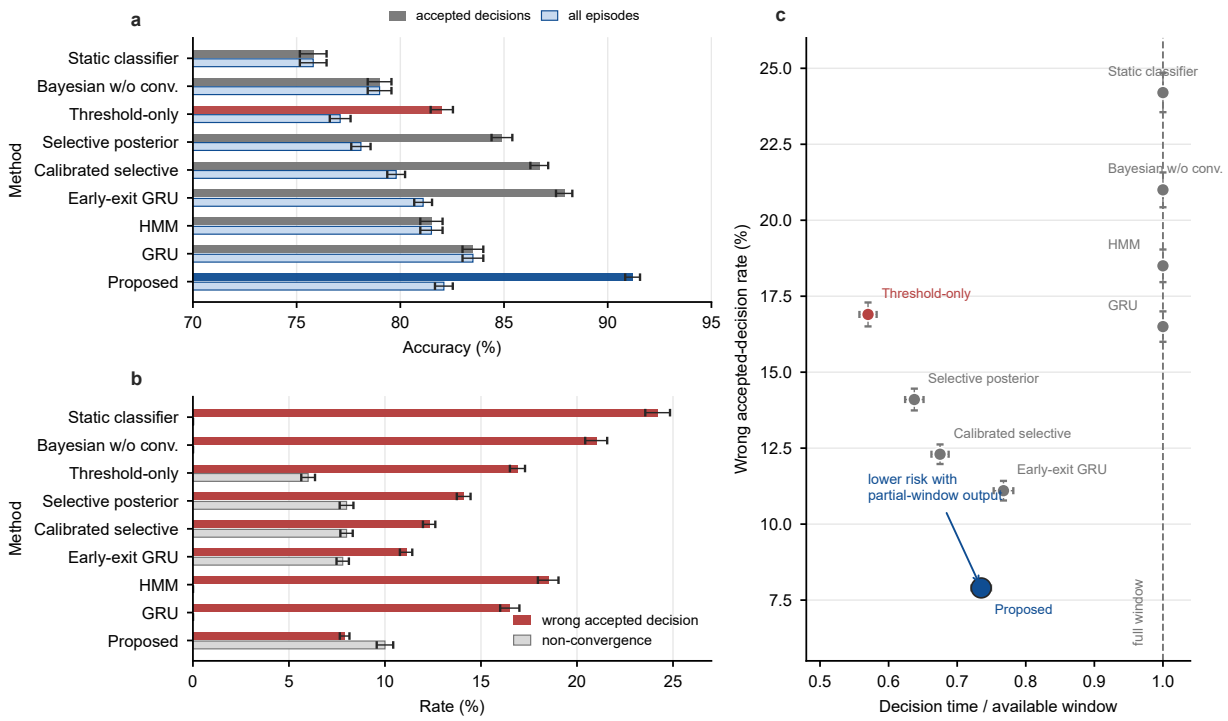


Figure 2: **Overall performance and reliability-timeliness comparison.** Panel a compares accepted-decision accuracy and all-episode accuracy. Panel b compares wrong accepted-decision rate and non-convergence or abstention rate. Panel c shows the reliability-timeliness trade-off, with decision time normalized by the available observation window. Error bars denote 95% confidence intervals over 30 random trials. Forced-output baselines have no non-convergence state by design. Source data are provided in the figure source data files.

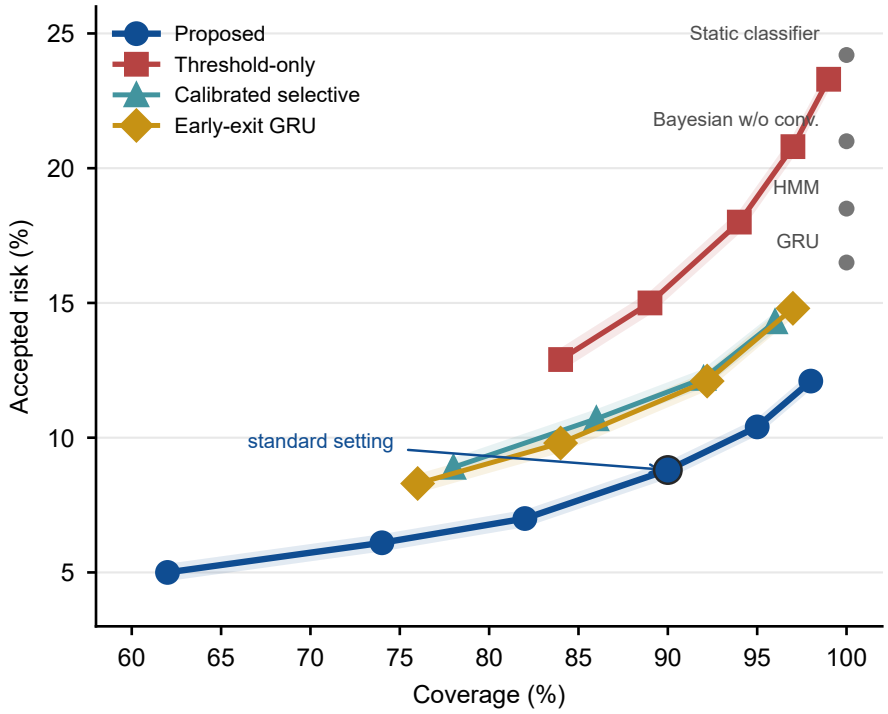


Figure 3: **Coverage-risk behaviour under different acceptance strictness levels.** Coverage is the percentage of test episodes with an accepted recognition output. Accepted-decision risk is $100 - A_{\mathrm{lw}}$, where A_{lw} is accepted-decision accuracy. Curves show how risk changes as acceptance becomes more or less strict. Forced-output baselines are shown as reference points at 100% coverage. Shaded bands denote 95% confidence intervals over 30 random trials. Source data are provided in the figure source data files.

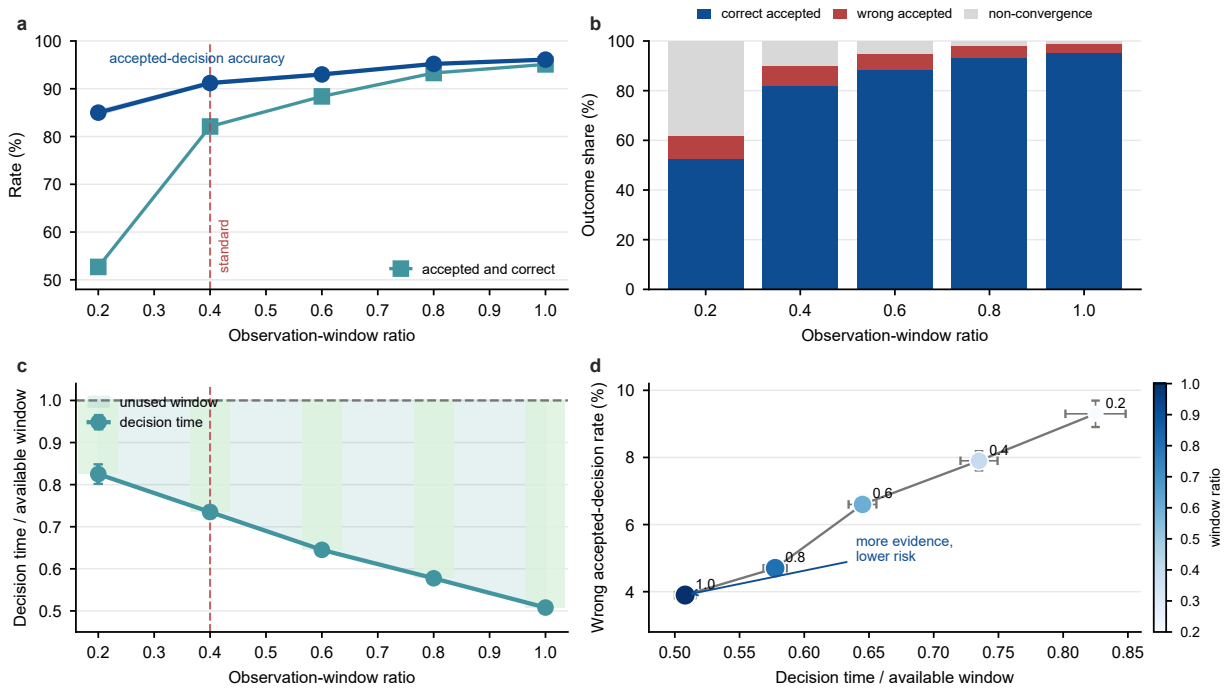


Figure 4: **Effect of observation-window length on reliability and timeliness.** Panel a shows accepted-decision accuracy conditioned on accepted outputs and the correct accepted share over all test episodes. Panel b shows the decision-outcome composition. Panel c shows mean decision time relative to the available window. Panel d shows the reliability-timeliness trajectory as the observation-window ratio increases. Error bars and shaded bands denote 95% confidence intervals over 30 random trials where applicable. Source data are provided in the figure source data files.

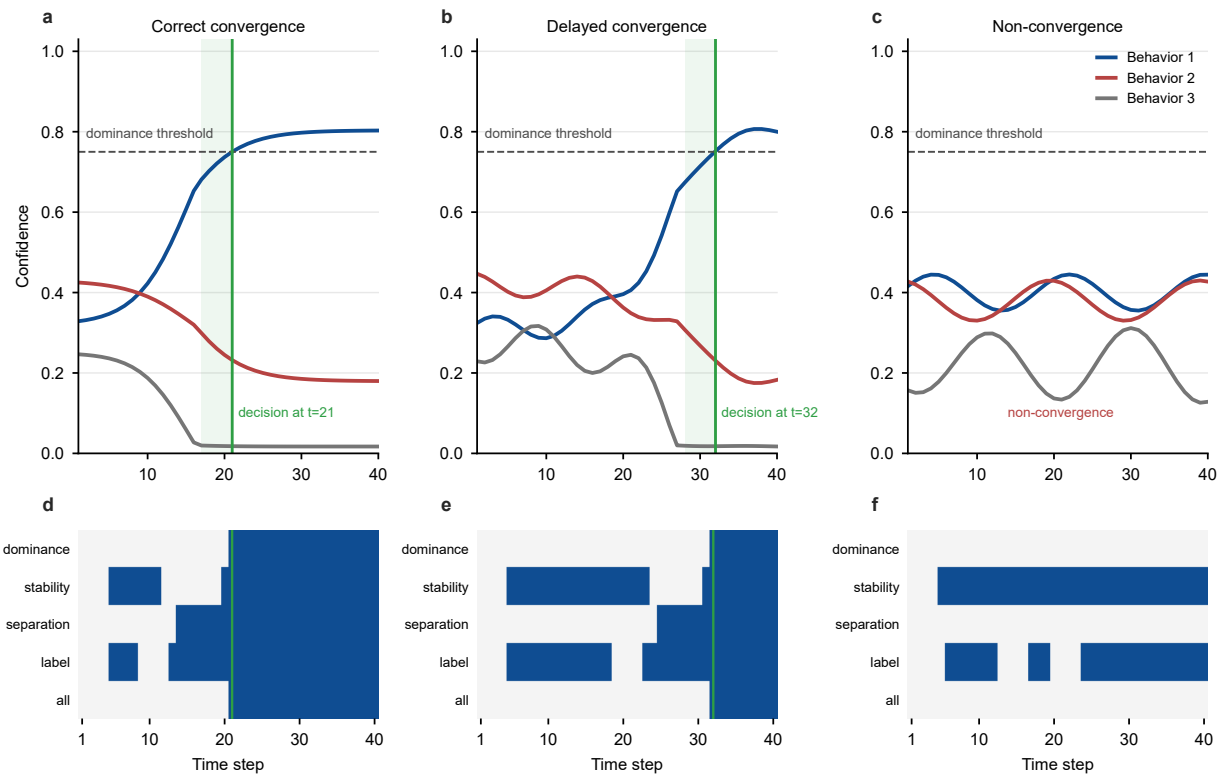


Figure 5: **Representative confidence evolution and convergence-criterion maps.** Panels a to c show confidence trajectories for correct convergence, delayed convergence, and non-convergence cases. Panels d to f show binary satisfaction maps for dominance, temporal stability, inter-class separation, label consistency, and the all-satisfied trigger. The vertical green line marks the first time step at which all convergence conditions are satisfied. Source data are provided in the figure source data files.

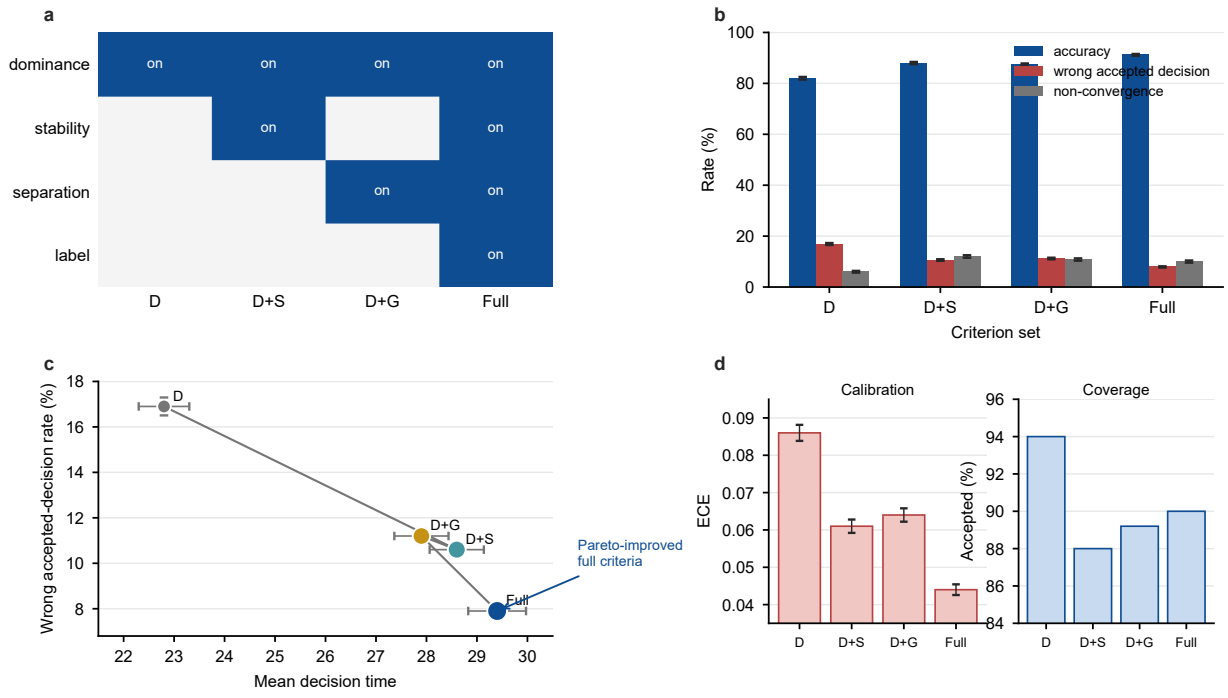


Figure 6: **Ablation analysis of confidence-convergence criteria.** Panel a shows the active criteria for each variant, including label consistency. Panel b compares accepted-decision accuracy, wrong accepted-decision rate, and non-convergence rate. Panel c shows reliability and timeliness. Panel d reports calibration error and accepted coverage. Error bars denote 95% confidence intervals over 30 random trials where applicable. Source data are provided in the figure source data files.

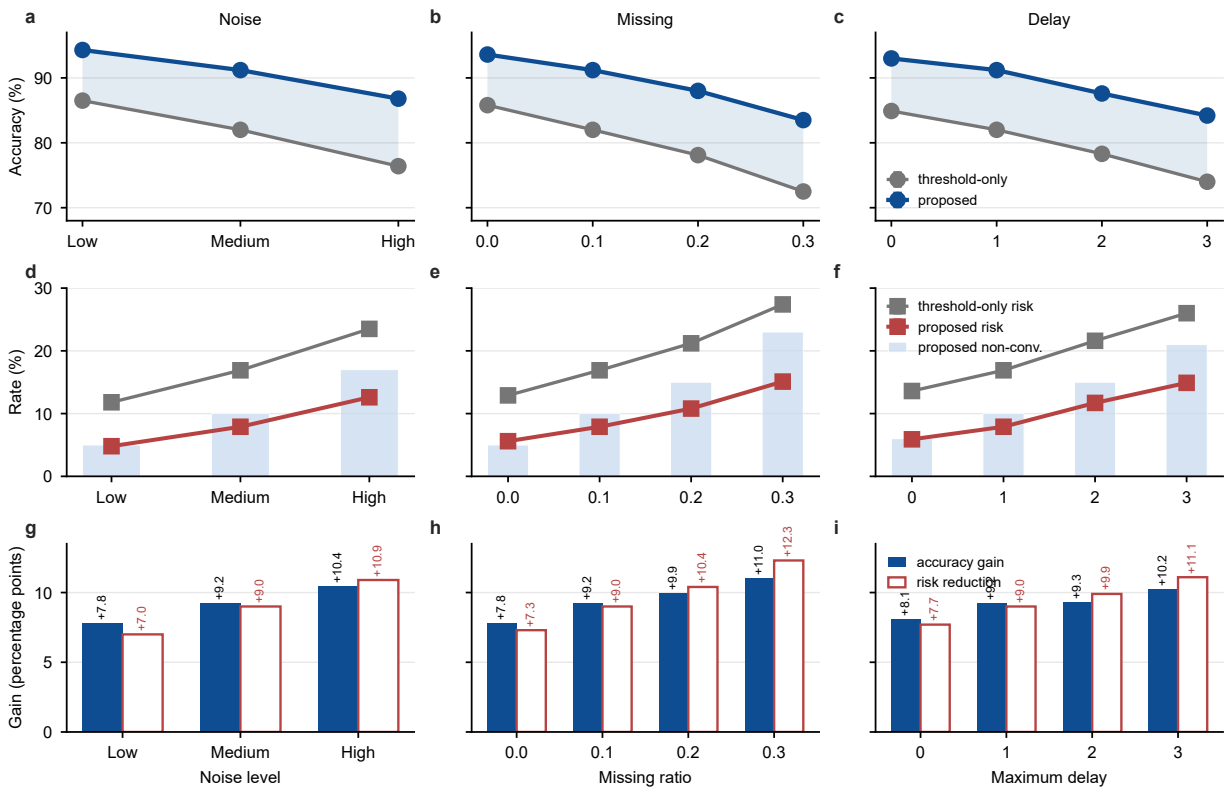


Figure 7: **Robustness under observation noise, missing observations, and measurement delay.** Panels a to c show accepted-decision accuracy under degraded sensing conditions. Panels d to f show wrong accepted-decision and non-convergence behaviour. Panels g to i show accuracy gain and wrong accepted-decision reduction relative to the threshold-only baseline. Error bars denote 95% confidence intervals over 30 random trials where applicable. Source data are provided in the figure source data files.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformation.pdf