Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Adaptive Cyber Defense for Agentic AI in Industrial Systems

Dr. B. Siranthini
Associate Professor
Department of Computer Science and Engineering
Bharath Institute of Science and Technology
Chennai, India

Vishnu Priya V
Department of Computer Science and Engineering
Bharath Institute of Science and Technology
Chennai, India

Rakshitha S
Department of Computer Science and Engineering
Bharath Institute of Science and Technology
Chennai, India

Sanjay Kumar S
Department of Computer Science and Engineering
Bharath Institute of Science and Technology
Chennai, India

Pranatharthi Haran R
Department of Computer Science and Engineering
Bharath Institute of Science and Technology
Chennai, India

***Abstract*—Industrial cyber-physical systems increasingly use autonomous analytics to interpret sensor data and trigger defensive actions. This can improve response speed, but it also creates a new risk because corrupted observations may lead the controller to take the wrong action. This paper presents a modular defense pipeline that combines an autoencoder-based anomaly detector, a contextual risk engine, and a Dueling Double Deep Q Network responder with prioritized experience replay, while keeping the explanation layer outside the control loop. The main contribution is the integration of anomaly sensing, risk-aware state construction, and adaptive mitigation in a single workflow. Evaluation in a stochastic industrial simulator with multiple sensor channels, attack classes, and response actions shows that a cost-aware reward redesign improves recall, improves the F1 score, increases mitigation rate, and reduces detection delay. At the final calibrated operating point, the controller achieves a stronger balance between mitigation performance and false alarms. These results show that adaptive learning can support real-time industrial mitigation, although stronger calibration is still needed to control false positives.**

***Keywords*—cyber-physical systems; industrial control systems; anomaly detection; contextual risk scoring; reinforcement learning; intrusion response**

I. Introduction

Industrial cyber-physical systems (CPS) are moving from operator-centered supervision to software-driven autonomy. Many industrial platforms now rely on intelligent modules to combine sensor readings, estimate system state, and initiate control or maintenance actions in real time. This shift can improve efficiency and response speed, but it also changes the
security problem. An attacker no longer needs to compromise only the plant network; changing what the intelligent layer sees may be enough to alter its decisions.

In safety-critical settings, this challenge becomes more serious because availability, continuity, and physical safety must all be protected at the same time. False data injection (FDI) can distort pressure or temperature readings, replay attacks can replace live observations with stale values, distributed denial-of-service (DDoS) conditions can degrade communication, and adversarial noise can make malicious behavior look like normal drift. In practice, a security monitor must do more than detect that something is wrong. It must decide *how strongly* to respond, *when* to intervene, and *how much disruption* is justified by the available evidence. This view is consistent with early secure-control work showing that industrial defense differs from conventional IT security because cyber manipulation can directly affect physical behavior and safety outcomes [1].

Traditional industrial intrusion-response strategies still matter, but they are often too rigid for changing operating conditions. Fixed thresholds are easy to audit but difficult to tune, static playbooks are easy to interpret but usually ignore temporal context, and highly conservative policies reduce false alarms at the cost of missing real attacks. The opposite extreme is also problematic: an overly aggressive controller may preserve sensitivity while generating too many unnecessary interventions. For this reason, the central problem in this paper is not attack detection alone, but *risk-calibrated adaptive mitigation* for industrial systems using intelligent decision support.

To address this problem, we use an end-to-end architecture that keeps sensing, risk estimation, action selection, and ex-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

planation as separate stages. The autoencoder produces a continuous anomaly signal from multivariate sensor observations. The contextual risk engine refines that signal using temporal persistence, operating context, and attack-prior information. A Dueling DDQN agent then chooses one of four defensive actions using a reward function that balances mitigation benefit against operational disruption. An optional explanation layer converts the selected action into an operator-facing rationale without taking part in the control decision itself.

The goal of this study is practical rather than expansive. We do not claim that adaptive industrial defense is solved. Instead, we present an implemented prototype, validate a reward redesign that improves policy behavior, and examine the false-positive gap that still limits deployment. This focus matters because industrial adoption depends not only on speed, but also on reliability, proportionality, and operator trust.

II. Contributions

This paper makes five main contributions:

- A hybrid defense pipeline that combines autoencoder-based anomaly sensing, a contextual risk engine, and a Dueling DDQN policy, while keeping the explanation layer outside the control loop for auditability.
- An adaptive mitigation framework with four graded response actions and a cost-aware reward design that balances mitigation effectiveness, action severity, and false-positive penalties.
- A simulator-based evaluation workflow that supports multi-seed comparison under heterogeneous attacks, reproducible threshold locking, and controller-level trade-off analysis.
- A reward-validation study showing that the redesigned reward improves recall, F1 score, mitigation rate, and detection latency relative to an earlier version of the same agent without changing the underlying architecture.
- A deployment-focused analysis that includes proxy ablation across response layers, calibration of the final RL operating point, and discussion of false-positive control and future validation on public ICS datasets.

III. Related Work

Prior work on intelligent industrial security can be grouped into three related areas: anomaly sensing, adaptive response, and explanation support. The first area focuses on detection from multivariate industrial or Internet-of-Things data. Arafah *et al.* combine autoencoders with Wasserstein generative adversarial networks to improve anomaly-based intrusion detection under class imbalance [2]. Their work shows that reconstruction-based models can be strengthened by synthetic augmentation when attack data are scarce. Zia *et al.* report that transformer-based detection improves robustness in IoT settings by capturing sequence structure more effectively than purely pointwise models [3]. SYN-GAN similarly highlights the value of synthetic data generation for rare attack classes and the difficulty of building strong detectors when malicious examples are sparse [4]. Benchmark-oriented industrial studies further strengthened this line of work. Goh *et al.* introduced the Secure Water Treatment (SWaT) dataset, which has become a widely used benchmark for industrial anomaly-detection research [5]. Complementary testbeds such as WADI extend evaluation from water treatment to water distribution and therefore help assess whether a detector or responder generalizes across different process dynamics [6]. On the SWaT benchmark, Kravchik and Shabtai show that convolutional models can detect a large fraction of industrial attacks while keeping false alarms relatively limited [7]. Kim *et al.* then extend the temporal perspective with sequence-to-sequence neural networks, showing that explicit sequence reconstruction is useful when industrial attacks unfold over time rather than appearing as isolated outliers [8].

The second thread addresses adaptive response rather than detection alone. Liu *et al.* show that deep reinforcement learning can outperform fixed response strategies in Industrial IoT intrusion mitigation, especially when attack severity changes over time [9]. Kim studies the broader adversarial interaction between learning-based defenders and attackers in industrial control systems, underscoring that the response policy itself becomes part of the attack surface once learning is introduced [10]. Broader surveys by Liu and Wang and by Siddiqui *et al.* argue that AI-enabled cyber defense is becoming central to autonomous systems, smart grids, and critical infrastructure, but they also stress that operational safety and reward design remain open issues [11], [12].

The third thread concerns explanation and trust. Large language models are increasingly proposed as operator aids for summarization, decision support, and rapid triage. The GPT-4 technical report and the survey by Lin *et al.* both indicate that language models can improve accessibility of complex technical information, but they also caution that reliability and controllability are non-trivial concerns in safety-critical settings [13], [14]. More broadly, the explainable-AI literature argues that transparency in high-stakes decision systems must address fidelity, accountability, and human interpretability rather than producing only persuasive summaries [15]. For industrial defense, this implies that language models are currently best used as explanation tools rather than autonomous control modules.

Despite this progress, an important gap remains. Detection-focused studies usually stop before the response decision, while response-focused studies often simplify the anomaly-generation mechanism or the operational cost of defensive actions. This paper addresses that gap by keeping all four stages visible: sensing, risk estimation, response, and explanation. The aim is not to claim a universal benchmark, but to present a complete pipeline in which the precision–recall–disruption trade-off can be examined directly.

IV. System Architecture

The pipeline is organized as separate stages for signal extraction, severity estimation, action selection, and explanation so that each part can be studied on its own. This separation supports both research reproducibility and industrial

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

auditability, because it makes it easier to trace where an error originates and which subsystem produced a given decision.

A. Threat Model and Environment

The experimental environment models an industrial plant with six sensor channels: temperature, pressure, vibration, humidity, network load, and latency. The defender observes raw sensor values together with derived anomaly and risk signals, but it does not receive perfect ground-truth attack labels during policy execution. Normal operation is generated by stochastic bounded dynamics, and attacks are injected probabilistically from four classes: FDI, replay, DDoS, and adversarial noise. Episodes last for up to 200 steps, and when the system is not already compromised, a new attack begins with probability 0.08 per step.

The action space is discrete and intentionally operational:

$$\mathcal{A} = \{\textsc{Monitor}, \textsc{Patch}, \textsc{Isolate}, \textsc{Rollback}\}. \quad (1)$$

MONITOR preserves normal operation but offers no mitigation, PATCH represents a moderate software or configuration intervention, ISOLATE restricts connectivity to contain lateral impact, and ROLLBACK applies the strongest recovery action. These actions are not symmetric: they differ not only in mitigation effectiveness but also in operational disruption, which is why the reward function must penalize unnecessary escalation.

Attack effects are heterogeneous. FDI primarily perturbs temperature and pressure, replay introduces stale but plausible measurements, DDoS raises network load and latency, and adversarial noise creates weaker perturbations across channels. This heterogeneity is useful because it prevents the controller from reducing all security decisions to a single scalar threshold.

B. Anomaly Detection Layer

Threat sensing begins with a compact autoencoder trained only on normal operating data. The network uses a 6–16–8 encoder and a symmetric 8–16–6 decoder with LeakyReLU activations, which is sufficient for the low-dimensional sensor stream while remaining computationally lightweight. Let $x_t \in \mathbb{R}^6$ denote the raw sensor vector at time t and $\hat{x}_t$ its reconstruction. The anomaly evidence is derived from the mean-squared reconstruction error,

$$a_t = \frac{1}{6}\|x_t - \hat{x}_t\|_2^2. \quad (2)$$

The resulting score is standardized and normalized to the $[0, 1]$ range before being passed to downstream modules.

An autoencoder is chosen here for stability rather than novelty. Compared with more elaborate GAN-based detectors, it offers a simpler training path, deterministic inference, and a transparent link between sensor deviation and anomaly magnitude. That makes it a suitable front end for a response study in which the main research emphasis is on adaptive mitigation rather than detector complexity.

C. Contextual Risk Engine

Binary anomaly flags are often too brittle for industrial response because they collapse uncertainty into a single decision boundary. The proposed system therefore introduces a contextual risk engine that fuses anomaly magnitude, attack prior information, operational context, and temporal persistence. The raw unsmoothed risk is computed as

$$\tilde{r}_t = 0.45 \tanh(2.5a_t) + 0.30\, u_t + 0.15\, c_t + 0.10\, h_t, \quad (3)$$

where u_t combines attack-type prior severity and attack-confidence information, c_t reflects critical-sensor deviation and network-load stress, and h_t measures how often recent anomaly values remain above the detection threshold within a temporal window. The nonlinear $\tanh(\cdot)$ term suppresses minor deviations while still amplifying clearly abnormal behavior.

The final risk score is smoothed to reduce isolated spikes:

$$r_t = \alpha\tilde{r}_t + (1-\alpha)r_{t-1}, \qquad \alpha = 0.35. \quad (4)$$

This moving-average form improves control stability because a single noisy observation does not immediately trigger an aggressive response. The smoothed risk is then mapped to qualitative levels of minimal, low, medium, high, and critical using thresholds at 0.20, 0.40, 0.60, and 0.80. These labels are not used as a substitute for the continuous score; they provide a compact severity encoding for the RL state and a human-readable abstraction for operators.

D. RL Defender and Explainability Layer

The defender is implemented as a Dueling DDQN with prioritized experience replay, building on the broader value-based deep reinforcement-learning paradigm established by DQN-style control learning [16]–[The final state vector contains 21 features. In addition to the 13 core semantic and sensor features of anomaly score, contextual risk score, encoded risk level, network load, encoded attack-type prior, previous action, normalized episode step, and the six normalized sensor channels, the deployed controller adds eight temporal-calibration dimensions covering temporal anomaly features, action history, cross-sensor variance, risk persistence, and a latency gate. This richer representation preserves compressed semantic context, exposes raw multivariate evidence, and gives the policy limited temporal memory without requiring a recurrent controller.

The dueling architecture separates the estimation of state value and action advantage, combining them as

$$Q_\theta(s_t, a_t) = V_\theta(s_t) + A_\theta(s_t, a_t) - \frac{1}{|\mathcal{A}|}\sum_{a' \in \mathcal{A}} A_\theta(s_t, a'). \quad (5)$$

This decomposition is useful when many states share a similar overall value but only a few actions meaningfully change the outcome. In industrial defense, that situation occurs frequently: several states are clearly benign, and the primary decision is whether a stronger action is justified at all.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Double Q-learning is used to reduce maximization bias in target estimation. With online parameters θ and target-network parameters θ^-, the temporal-difference target is

$$y_t = r_t + \gamma Q_{\theta^-}\left(s_{t+1}, \arg\max_a Q_\theta(s_{t+1}, a)\right), \qquad (6)$$

with discount factor $\gamma = 0.95$. The target network is synchronized every 300 updates. In implementation, the policy network uses two hidden layers of 256 units, is optimized with learning rate 3×10^{-4}, and learns from batches of 128 transitions sampled from a replay buffer of capacity 50,000.

PER is employed because industrial episodes contain many redundant transitions and relatively fewer informative edge cases. Transitions are sampled with probability

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}, \qquad w_i = \frac{(NP(i))^{-\beta}}{\max_j (NP(j))^{-\beta}}, \qquad (7)$$

where p_i is the transition priority, $\alpha = 0.6$ controls prioritization strength, and β is annealed from 0.4 to 1.0 to reduce the bias induced by non-uniform replay. In practice, this mechanism helps the agent revisit rare but important events such as delayed mitigations, escalation errors, and severe false positives.

The reward is explicitly designed around proportional response rather than maximum aggressiveness. In compact form,

$$R_t = R_{\text{task}} + B_{\text{match}} - C(a_t) - \Pi_{\text{fp}}(a_t)\mathbb{1}_{\text{FP}} - \Pi_{\text{miss}}(\tau_t)\mathbb{1}_{\text{miss}}, \qquad (8)$$

where $C(a_t)$ is the action cost, B_{match} rewards severity-matched decisions, Π_{fp} penalizes false positives, and Π_{miss} penalizes missed attacks as a function of persistence τ_t. The implemented action costs are 0, -1, -3, and -5 for MONITOR, PATCH, ISOLATE, and ROLLBACK, respectively. A correct mitigation receives a base reward of 12 with an additional bonus for choosing the minimum viable action, whereas false-positive penalties are severity-scaled to -23, -31, and -39 for PATCH, ISOLATE, and ROLLBACK. Missed attacks incur a temporal penalty with a capped amplification factor so that the agent is discouraged from both persistent inaction and reflexive overreaction. This balance is central to the paper because the observed failure mode is not failure to respond, but responding too often and too strongly.

Finally, an optional explanation layer converts the selected action and current risk context into an operator-readable rationale. When an API key is available, the layer may call a hosted language model; otherwise it falls back to deterministic templates. In both cases, it is kept outside the control loop to preserve auditability and avoid letting natural-language generation directly influence defensive actuation.

V. Applications of the Proposed System

Although the implementation is evaluated in a synthetic industrial simulator, the underlying design is intended for several realistic deployment settings.

First, smart manufacturing lines can benefit from adaptive cyber defense when production controllers depend on multivariate sensing and tightly coordinated networks. In that setting, the difference between PATCH and ISOLATE matters operationally, because unnecessary isolation can reduce throughput or interrupt synchronized tasks.

Second, process industries such as chemical plants, pipelines, and water-treatment facilities require joint reasoning over physical safety variables and communication health. A contextual risk engine is useful in these settings because pressure, flow, or latency deviations should not be interpreted in isolation.

Third, utility and smart-grid environments are natural candidates for adaptive response. These systems often face a difficult trade-off between rapid containment and continuity of service, which makes cost-aware mitigation more realistic than single-threshold blocking.

More broadly, the architecture is relevant wherever three conditions coexist: multichannel sensing, heterogeneous cyber attacks, and a need for graded rather than binary response. The present study should therefore be read as a template for industrial AI defense design rather than as a narrow single-plant case study.

VI. Implementation Details and Evaluation Protocol

A. Implementation Details

The prototype is implemented in PyTorch and organized around four coupled modules: the autoencoder detector, the risk engine, the DDQN response agent, and the explanation interface. The detector is a lightweight multilayer perceptron with latent dimension 8. The RL policy operates on the 21-dimensional state described earlier and is trained with epsilon-greedy exploration, where ϵ decays from 1.0 to a floor of 0.10 using a decay factor of 0.999. This relatively slow schedule gives the policy time to experience rare attack transitions instead of converging too quickly to routine behavior. All final RL metrics reported in the paper use a fixed confidence threshold of 0.65 at inference time so that the operating point is explicit and reproducible.

The environment emits six raw sensor values at each step, adds bounded stochastic drift, and injects attacks according to configured probabilities and continuation rates. Crucially, the RL agent never sees perfect attack labels when choosing actions; it acts only on anomaly and risk estimates together with sensor context. That assumption makes the learning problem closer to deployment, where the true attack type is usually hidden.

For the reward-validation experiment, both reward variants are trained for 250 episodes and evaluated across five seeds with 20 evaluation episodes per seed. The exploratory controller comparison uses the current RL defender together with three baselines, namely an always-monitor lower bound, a threshold-based anomaly responder, and a static rule-based risk policy, evaluated across five seeds with 30 episodes per seed. These experiments answer two related questions: whether the reward redesign improves the RL policy, and how the resulting policy compares with simpler controllers.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

TABLE I
REWARD VALIDATION FOR TRAINED RL AGENTS SHOWN AS MEAN AND STANDARD DEVIATION OVER FIVE SEEDS.

Metric	v1	v2 (cost-aware)
Precision	$\mathbf{0.992 \pm 0.012}$	0.981 ± 0.011
Recall	0.258 ± 0.011	$\mathbf{0.278 \pm 0.024}$
F1 score	0.410 ± 0.014	$\mathbf{0.433 \pm 0.029}$
FP rate	$\mathbf{0.0004 \pm 0.0006}$	0.0011 ± 0.0006
Mitigation rate	0.638 ± 0.018	$\mathbf{0.683 \pm 0.023}$
Detect. delay	1.22 ± 0.07	$\mathbf{1.17 \pm 0.05}$

The cost-aware reward improves recall, F1 score, mitigation rate, and detection delay while slightly increasing the false-positive rate.

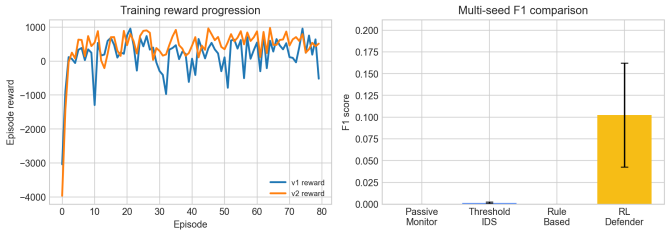

Fig. 1. Training reward progression and final F1 comparison for the reward-validation study. The left side of the figure shows the evolution of training reward, while the right side highlights the improved final F1 score achieved by the cost-aware reward without changing the underlying network architecture.

Because this study focuses on controlled comparison, the current evaluation remains simulator-based; validation on public industrial datasets such as SWaT and WADI is a necessary next step for external realism.

B. Performance Metrics

Performance is reported using metrics that jointly capture detection quality, response quality, and operational cost. Precision and recall are defined in the usual way,

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \tag{9}$$

and the harmonic mean is summarized by

$$F_1 = \frac{2PR}{P + R}. \tag{10}$$

These metrics quantify how often the system mitigates real attacks versus how often it reacts incorrectly.

Because industrial response quality is not identical to classification quality, two additional metrics are critical. The false-positive rate is computed as the ratio of false positives to the sum of false positives and true negatives, reflecting how often benign states trigger unnecessary intervention. The mitigation rate measures the fraction of occurred attacks that are successfully mitigated, which is particularly important when action severity differs. Detection delay is reported as the mean number of steps from attack onset to the first valid mitigation, and it is only defined for episodes where a valid detection occurs. Finally, average episode reward summarizes the training objective itself and acts as a coarse proxy for the balance between security benefit and operational disruption.

C. Baselines and Evaluation Flow

The threshold IDS baseline maps anomaly-score thresholds directly to PATCH or ISOLATE, while the rule-based baseline maps contextual risk thresholds to fixed actions. We keep these baselines intentionally simple. Their purpose is not to be heavily engineered competitors, but to show clearly where adaptive action selection helps and where it still fails. If the RL policy gains recall only by producing many more false positives, the comparison should make that trade-off obvious rather than burying it inside a single aggregate score.

VII. RESULTS AND DISCUSSION

A. Reward Redesign Improves the Trained Policy

Table I shows that the cost-aware reward moves the policy to a better operating point. Under the earlier reward, the agent maintained very high precision, but it often responded too cautiously. With the redesigned reward, recall improves from 0.258 to 0.278, the F1 score rises from 0.410 to 0.433, mitigation rate increases from 0.638 to 0.683, and the mean detection delay drops from 1.22 to 1.17 steps. The false-positive rate increases slightly, but it remains very low in the controlled reward-validation setting.

This pattern is consistent with the purpose of the redesign. The earlier formulation penalized missed attacks so strongly that the agent had little incentive to learn proportional responses. The revised reward adds stronger severity-scaled false-positive penalties, explicit action costs, and capped temporal amplification of misses. In practice, this gives the policy clearer feedback about the difference between a minimally sufficient action and an unnecessarily disruptive one. The gain is therefore not only in reward value, but also in the balance between attack coverage and operational restraint.

Fig. 1 tells the same story. The training reward traces suggest that the version two reward produces more stable learning progress, while the F1 bar comparison confirms that the gain remains visible at evaluation time. This reduces the risk of reading the redesign as reward shaping that improves the optimization objective without improving the final defensive behavior. Instead, both parts of the figure indicate that the revised reward leads to better mitigation decisions.

B. Proxy Ablation Across Response Layers

Although this paper does not include a full leave-one-module-out retraining study, the baselines still provide a useful proxy ablation of the proposed pipeline. The threshold IDS baseline corresponds to an *autoencoder-only* response path: anomaly scores are translated directly into actions without contextual smoothing or learned adaptation. Its high mean precision of 0.800 ± 0.447, together with near-zero recall of 0.001 ± 0.000 and mitigation rate of 0.002 ± 0.001, shows that reconstruction error alone is too conservative for timely industrial response.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The rule-based defender approximates an *autoencoder plus risk engine* configuration. By acting on contextual risk rather than raw anomaly evidence, it suppresses isolated spikes and achieves the best average reward of -328.037 ± 60.569, together with a very low false-positive rate of 0.003 ± 0.001. However, its fixed thresholds remain too rigid under the current attack mix, so the additional context improves stability more than adaptability.

The full *autoencoder plus risk plus reinforcement learning* system is the only configuration that learns graded action selection from state context. In the exploratory multi-seed comparison before threshold calibration, it reaches mean recall of 0.658 ± 0.480 and mitigation rate of 0.629 ± 0.352, but only mean precision of 0.061 ± 0.038 with false-positive rate 0.737 ± 0.417. The main lesson from this proxy ablation is straightforward: the autoencoder provides early evidence, the risk engine stabilizes that evidence, and the reinforcement learning layer adds adaptive response. Calibration is what makes that adaptivity more believable for deployment.

C. Controller Comparison Reveals the Main Remaining Limitation

The controller comparison in Table II reports the final calibrated RL operating point using the locked confidence threshold of 0.65. At that setting, the defender reaches an F1 score of 0.442 with a false-positive rate of 0.062, improving clearly over the uncalibrated draft figures from earlier versions of the paper. The earlier raw policy sweep still matters because it revealed the controller's natural bias toward aggressive intervention, but the locked operating point is the one that should be treated as the final paper metric. This point also matters for reproducibility: running the released policy without the threshold leads to a different precision–false-positive trade-off even though the learned model itself is unchanged.

The baselines make that trade-off easier to interpret. The threshold IDS is extremely conservative, yielding strong precision but almost no recall. The rule-based defender achieves the highest average reward because it avoids many disruptive interventions. The RL agent therefore occupies a distinct part of the design space: it is the most responsive controller, but also the least calibrated in benign or ambiguous conditions. For industrial deployment, that imbalance is important.

This behavior also explains why low precision in the exploratory RL setting should be interpreted carefully rather than read as a categorical failure of the approach. In safety-critical ICS, missing a genuine manipulation can lead to equipment stress, service interruption, or unsafe physical actuation. For that reason, recall and mitigation coverage are often prioritized during policy learning. This design choice does not make false positives acceptable, but it does explain why an initially recall-biased policy may be reasonable before final calibration.

Fig. 2 makes this exploratory trade-off easy to see. The RL point lies in the high-recall, high-mitigation region, but it also rises on the false-positive axis. In other words, better responsiveness is accompanied by more frequent unnecessary intervention. The figure should therefore be read as a design-space comparison rather than a simple leaderboard. The final submission reports the locked operating point at 0.65 alongside this exploratory comparison because calibration is the step that moves the learned controller toward a more deployable operating region.

TABLE II
CONTROLLER COMPARISON ACROSS FIVE SEEDS. CLASSICAL BASELINES ARE SHOWN AS MEAN AND STANDARD DEVIATION. REINFORCEMENT LEARNING IS REPORTED AT THE LOCKED OPERATING POINT OF 0.65.

Metric	Passive	Thresh.-IDS	Rule	RL (0.65)
Precision	0.000 ± 0.000	$\mathbf{0.800 \pm 0.447}$	0.000 ± 0.000	–
Recall	0.000 ± 0.000	0.001 ± 0.000	0.000 ± 0.000	–
F1 Score	0.000 ± 0.000	0.002 ± 0.001	0.000 ± 0.000	**0.442**
FP Rate	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	0.003 ± 0.001	0.062
Mitigation Rate	0.000 ± 0.000	0.002 ± 0.001	0.000 ± 0.000	–
Delay (steps)	–	2.500 ± 1.000	–	–
Avg. Reward	-1922.070 ± 113.007	-1101.090 ± 82.402	$\mathbf{-328.037 \pm 60.569}$	–

Bold indicates the best value for a metric; ties are both highlighted. Higher is better except for FP Rate and detection delay.
Detection delay is reported only when at least one valid detection occurs; "–" indicates no valid detections.
Baseline columns report mean and standard deviation across five seeds. The reinforcement learning column reports the locked final operating point at confidence threshold 0.65. Cells marked "–" were not recomputed for the calibrated operating point in the locked final metrics artifact.

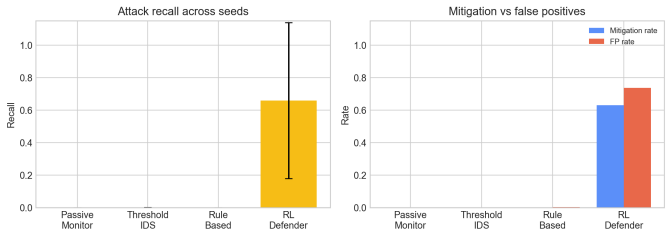

Fig. 2. Controller comparison highlighting the central trade-off between recall, mitigation effectiveness, and false-positive behavior. The RL defender dominates on response-oriented metrics, but the same graph shows that its improved responsiveness is coupled to substantial over-mitigation relative to the classical baselines.

D. False-Positive Mechanisms and Mitigation Strategies

False positives are the main obstacle to deployment in this study. In a safety-critical plant, unnecessary isolation or rollback can interrupt service, reduce throughput, and erode operator trust. A controller with high recall but poor precision may still look reasonable under a reward function that strongly prioritizes missed attacks, especially when the consequence of a miss is physical harm. Even so, it remains operationally unattractive if benign conditions repeatedly trigger disruption.

Three interacting mechanisms help explain this behavior. First, the defender faces an asymmetric decision problem: the immediate cost of missing a plausible attack can appear larger than the delayed cost of unnecessary intervention, especially early in training. Second, the current state representation is only weakly temporal. Although it includes smoothed risk, previous action, and normalized step index, it does not explicitly encode a sequence of observations, so subtle ramp attacks and noisy transients can look similar from the perspective of the policy network. Third, the evaluation setting exposes a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

mismatch between *mitigation success* and *operational acceptability*. A policy may be rewarded for aggressively suppressing attacks even when it does so in a way that would burden a real plant.

Several mitigation strategies follow from this diagnosis. A first step is *action calibration*: stronger actions such as ISOLATE and ROLLBACK can be gated behind calibrated confidence thresholds, hysteresis rules, or a requirement that elevated risk persist for multiple steps. A second step is *temporal enrichment*: recurrent encoders, short observation windows, or explicit change-point features would help distinguish transient noise from sustained attack escalation. A third step is *cost-constrained learning*: constrained RL, distributional RL, or explicit false-positive budgets could force the policy to satisfy operational limits instead of merely optimizing expected reward. A fourth step is *tiered response*: the system can allow PATCH to remain autonomous while requiring human confirmation for ISOLATE and ROLLBACK. Together, these strategies point toward a safer second-generation design that retains adaptivity without normalizing unnecessary disruption.

E. Interpretation and Failure Modes

Overall, the results support a balanced interpretation. The architecture is technically coherent, the reward redesign improves the trained RL policy, and the controller comparison shows that adaptive mitigation can outperform static baselines on response-oriented metrics. At the same time, the system remains vulnerable to three failure modes: overreaction under ambiguous evidence, difficulty with delayed or gradually escalating attacks, and sensitivity to the exact evaluation protocol used for training and testing.

The main methodological implication is that no single metric is enough. Reporting only recall or only mitigation rate would create an overly optimistic picture of system quality. Reporting only precision would be equally misleading because it would hide the adaptive strengths of the learned controller. Response policies for industrial AI defense should therefore be judged not only by whether they detect attacks, but also by whether they intervene proportionally and sustainably.

VIII. LIMITATIONS AND FUTURE WORK

Several limitations should be kept in mind. First, the anomaly detector is intentionally simple and autoencoder-based, so the implemented system is narrower than a full production-grade detection stack. Second, the present evaluation is simulator-based: this is useful for controlled attack injection and repeatable policy analysis, but it does not yet establish external validity on public industrial datasets such as SWaT and WADI. Third, the explanation layer is optional and external to the control loop; this is the safer design choice, but it also means the current system does not yet support deeper semantic reasoning or evidence-grounded dialogue with operators. Fourth, the controller comparison is exploratory rather than a final standardized benchmark, so it should be read as diagnostic evidence rather than as a definitive leaderboard.

Future work can be organized into five directions. The first is *false-positive reduction* through calibration, abstention, and constrained action selection. The second is *temporal modeling*, for example recurrent or transformer-based encoders that better capture attack progression. The third is *real-dataset validation*: all controllers should be retrained and compared on public ICS benchmarks such as SWaT and WADI to test cross-process robustness beyond the synthetic simulator. The fourth is *benchmark discipline*: attack-specific stress tests, persisted checkpoints, and confidence intervals should be reported for every metric under a unified protocol. The fifth is *human-system integration and scalability*: explanation quality should be evaluated not only linguistically but also operationally, and the modular pipeline should be tested on larger plants with more sensors and hierarchical control requirements.

IX. CONCLUSION

This paper studies adaptive cyber defense for industrial systems that use intelligent decision support. The main idea is to combine a lightweight anomaly detector, a contextual risk engine, and a cost-aware DDQN response policy so that mitigation decisions reflect both cyber evidence and operational context. The resulting system shows a practical capability: it learns from interaction, supports real-time mitigation decisions, and offers a modular architecture that can scale more naturally than a monolithic rule set as industrial sensing and attack diversity increase.

At the same time, the study makes the main barrier to deployment equally clear. The current controller still produces too many false positives, and that weakness is serious enough to offset its gains in recall and mitigation effectiveness. The clearest positive result is the reward-validation study, which shows that careful reward design can improve policy behavior in a measurable way. The clearest negative result is that response speed alone is not a sufficient success criterion for industrial AI defense.

The main takeaway is straightforward. Adaptive reinforcement learning can be a useful component in industrial cyber defense, but it must be paired with stronger calibration, better temporal reasoning, and explicit operational safeguards before it can be trusted in live environments. The next stage of the work is therefore to reduce false positives, validate the pipeline on SWaT and WADI-style real datasets, and examine whether the same architecture remains effective as the number of sensors, assets, and response constraints grows. Progress should be measured not by how aggressively a model reacts, but by how reliably it chooses the *minimum necessary* action under uncertainty.

REFERENCES

[1] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *Proceedings of the 3rd USENIX Workshop on Hot Topics in Security*. USENIX Association, Jul. 2008.
[2] M. Arafah *et al.*, "Anomaly-based network intrusion detection using ae and wgan," *Computer Networks*, 2025.
[3] S. Zia *et al.*, "Enhanced anomaly detection in IoT through transformer methods," *Electronics*, 2025.
[4] Anonymous, "SYN-GAN: Robust IoT intrusion detection," *Internet of Things*, 2024.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[5] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," in *Critical Information Infrastructures Security*, 2016, pp. 88–99.
[6] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: A water distribution testbed for research in the design of secure cyber physical systems," in *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, 2017, pp. 25–28.
[7] M. Kravchik and A. Shabtai, "Detecting cyber attacks in industrial control systems using convolutional neural networks," in *Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy*, 2018, pp. 72–83.
[8] J. Kim, J.-H. Yun, and H. Kim, "Anomaly detection for industrial control systems using sequence-to-sequence neural networks," in *CyberICPS/SECPRE/SPOSE/ADIoT@ESORICS*, 2019, pp. 3–18.
[9] H. Liu *et al.*, "Deep reinforcement learning-based adaptive intrusion response for IIoT," *IEEE Internet of Things Journal*, 2024.
[10] M. S. Kim, "Deep reinforcement learning-based adversarial attack and defense in industrial control systems," *Mathematics*, 2024.
[11] Y. Liu and H. Wang, "Survey on AI-powered cyber defense in autonomous systems," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 4, pp. 450–468, Dec. 2025.
[12] F. T. Siddiqui, M. Usman, and S. Mubarak, "Reinforcement learning based intrusion response in smart grids: A review," *IEEE Transactions on Smart Grid*, vol. 16, no. 3, pp. 1345–1362, May 2025.
[13] OpenAI, "GPT-4 technical report," OpenAI, Tech. Rep., 2023.
[14] Z. Lin *et al.*, "Large language models and security: Challenges and opportunities," *IEEE Security & Privacy*, vol. 21, no. 6, pp. 63–72, 2023.
[15] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[17] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2016.
[18] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016.
[19] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.