Future Technology
Open Access Journal
ISSN 2832-0379

November 2026| Volume 05 | Issue 04 | Pages 15-24

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.4.2

Future Publishing LLC

Article

Multi-agent generative AI ecosystems for cyber-physical systems in Industry 5.0

Kanaparthi Anil Kumar*, K Hemachandran

Woxsen University, Telangana, India

ARTICLE INFO

Article history:
Received 30 January 2026
Received in revised form
12 June 2026
Accepted 16 July 2026

Keywords:
Cyber-physical systems, Industry 5.0,
Multi-agent systems, Generative AI,
Anomaly detection, 1D-CNN

*Corresponding author
Email address:
anilkds.85@gmail.com

DOI: 10.55670/fpll.futech.5.4.2

ABSTRACT

Cyber-Physical Systems (CPSs) are increasingly being implemented in critical industrial infrastructure, where complexity and interdependence are rising, posing significant cyber and operational risks. Traditional anomaly detection algorithms can be ineffective in capturing temporal dynamics, relational dependencies, and interpretable response requirements. The present paper proposes a multi-agent generative AI system to detect CPS anomalies and provide decision support by combining temporal feature encoding, relational modeling as graphs, supervised learning, and reasoning with an LLM. The architecture consists of detection, diagnosis, planning, governance, and human-in-the-loop validation agents. The framework is tested on the SWaT benchmark data. Findings indicate that the Autoencoder, LSTM, and 1D-CNN are more effective in terms of raw detection metrics, whereas the Random Forest provides more interpretable and agent-readable evidence to support the post-detection decision. Analysis of features and sensor family suggests the relevance of relational dependencies in characterizing anomalies. The multi-agent layer converts selected anomaly predictions into context-dependent explanations and governance-filtered recommendations to aid operator review, response planning, and process resilience. Overall, the framework supports transparent and human-supervised CPS anomaly management aligned with Industry 5.0 principles.

1. Introduction

Cyber-Physical Systems (CPSs) integrate sensors, actuators, communication networks, computation, and physical processes to support modern industrial operations such as manufacturing, water treatment, energy distribution, and transportation. As these systems become more connected and automated, they also become more exposed to cyberattacks, process faults, abnormal sensor behavior, and operational disruptions. In such environments, anomaly detection is essential because even a small undetected fault or attack can propagate through interconnected components and compromise safety, reliability, and service continuity. Recent advances in generative artificial intelligence (GenAI) and large language models (LLMs) have created new opportunities for CPS security and monitoring. GenAI-based intrusion detection systems can support not only detection but also contextual reasoning and explanation generation in industrial CPS environments [1]. At the same time, generative AI introduces both benefits and risks to CPS cybersecurity, particularly because AI-generated outputs must be reliable, transparent, and safe before use in operational settings [2]. LLMs and GenAI have also been explored for adaptive resource management in networked sensing environments, demonstrating their potential to reason across distributed data sources [3]. In parallel, multi-agent systems powered by LLMs have shown promise for coordination, task decomposition, and collaborative reasoning [4]. Recent studies have further explored autonomous cyber-physical defense using LLM-based semantic reasoning and multi-agent reinforcement learning [5], agentic simulation frameworks for cyber-physical-social systems [6], and adversarially resilient multi-agent learning for CPS security [7]. These studies indicate that future CPS security systems may benefit from combining detection, reasoning, coordination, and governance. However, many existing works still address these capabilities separately rather than integrating them into a single anomaly-to-action pipeline. Existing CPS anomaly detection studies also show that industrial data are difficult to model because they include heterogeneous sensors, discrete actuators, temporal dependencies, and changing attack patterns [8]. Reviews of CPS anomaly detection highlight the need for adaptive methods that can respond to evolving threats rather than relying solely on static rules or isolated classifiers [9,10]. Nevertheless, most anomaly detection models still produce only a label or a probability score. Such outputs are useful for detection but insufficient for operational decision-making, where engineers and operators need to understand why an anomaly occurred,

which sensors were involved, what action should be taken, and whether that action is safe.

Problem statement: Current CPS anomaly detection methods mainly focus on predictive accuracy, but anomaly prediction alone does not provide interpretable, graph-aware, and governance-aware decision support. There remains a gap between detecting abnormal windows in sensor-actuator data and converting those detections into clear explanations, safe recommendations, and human-verifiable operational decisions.

To address this gap, this paper proposes a multi-agent generative AI framework for CPS anomaly detection and decision support. The framework combines temporal feature extraction, graph-based relational modeling, supervised machine learning, and LLM-assisted agent reasoning. The detection layer includes the ability to detect windows as anomalous based on CPS data, and the agent layer transforms selected high-risk detections into diagnostic explanations, planning recommendations, governance checks, and human-in-the-loop reviews. The problem formulation is given in Section 3.1, and the Introduction contains the motivation, research gap, and contribution.

The novelty of this work is the fusion of four components into a single CPS-focused pipeline: temporal anomaly detection, relational evidence from the graph, interpretable supervised learning, and reasoning with multiple agents and LLMs after detection. This work links the problem of anomaly detection to the problems of explanation, recommendation, and governance-oriented decision support, which are not covered by studies focusing solely on model accuracy, autonomous agents, or LLM reasoning. This work is done to achieve the following objectives:

- To design a unified CPS anomaly detection framework that combines temporal, group-level, and graph-derived features.
- To develop a multi-agent LLM-assisted decision-support layer for diagnosis, planning, governance checking, and human validation.
- To evaluate the detection performance, ablation behavior, interpretability, and operational usefulness of the proposed framework using the SWaT benchmark dataset.

Overall, the proposed framework aims to move beyond simple anomaly classification toward explainable and governance-aware CPS anomaly management. It does not claim that the agent layer directly improves raw detection metrics; rather, it supports the transformation of anomaly predictions into actionable, auditable, and operator-oriented decision support.

2. Literature review

Cyber-Physical Systems (CPSs) are gaining popularity across various industrial fields, including power systems, water treatment, manufacturing, and transportation. As they get closer together, automation and monitoring will improve, but they will also be exposed to increased risks of cyberattacks, process failures, and abnormal operating conditions. The study presented in the recently published CPS security document [11] indicates that static rule-based methods and single classifiers are insufficient for dynamic cyber-physical threats in future industrial systems. Agent-based and distributed AI approaches have gained prominence in industrial CPS due to their ability to enable coordination, decentralization, and fault-tolerant reasoning across system components. Complex tasks can be broken down into smaller components and allocated to agents that specialize in specific tasks, which can be useful in CPS systems where heterogeneous sensor and actuator data require interpretation in different subsystems [12]. The use of Generative AI has also been investigated in industrial settings, where it can provide more than just numerical forecasting: it can be used for simulation, contextual reasoning, and helping decision-making [13]. Furthermore, there are some potential options for representing CPS assets, threats, and process dependencies using a semantic or knowledge-graph approach [14].

The foundations are useful concepts for coordinated reasoning, task decomposition, and adaptive planning in complex environments, as derived from research on generative multi-agent collaboration [15]. Governance and accountability are also crucial in agentic AI systems, as AI-generated outputs can affect operational choices, safety protocols, and human oversight [16]. In a CPS environment, this includes not just detecting anomalies but also providing evidence, offering safe recommendations, and enabling an operator to review high-risk decisions. The multi-agent systems based on LLMs provide more methodological support for this direction. Previous research has demonstrated that intelligent LLM agents can work together by delegating tasks, holding dialogue, planning, and sharing knowledge [17]. AutoGen is one such framework that allows multiple agents to interact in a structured manner to achieve complex tasks by communicating with one another [18]. In addition, multi-agent language-model systems can support communicative reasoning, task coordination, and emergent collaborative behavior, as demonstrated by CAMEL and AgentVerse [19,20]. These works are useful for designing agentic reasoning pipelines, although most were not developed specifically for industrial CPS anomaly detection.

Foundational CPS- and SWaT-related anomaly-detection research is also important to this study. The SWaT testbed was developed as a realistic water treatment environment for research and training in industrial control system security [21]. Related SWaT dataset work has supported research on secure water treatment systems and benchmark-based evaluation of anomaly detection methods [22]. These foundational studies are important because they provide the experimental basis for evaluating CPS anomaly detection under realistic sensor-actuator behavior. Despite these advances, there is still a gap in research. Current research tends to concentrate on a single aspect- CPS anomaly detection, distributed agents, LLM-based reasoning, semantic graph modeling, or governance. In contrast, the present work integrates these elements into one CPS-focused anomaly-to-action pipeline. The proposed framework combines temporal feature extraction, graph-derived relational evidence, supervised anomaly detection, LLM-assisted diagnostic reasoning, planning, governance filtering, and human-in-the-loop validation. This integrated design distinguishes the study from prior works that treat detection, reasoning, multi-agent coordination, and governance as separate problems.

3. Methodology

3.1 Problem formulation

This study frames CPS anomaly detection as a supervised, window-level binary classification problem, followed by post-detection decision support. A CPS is represented as:

$$G = (N, E, U, Z) \quad (1)$$

where N denotes sensors, actuators, and controllers, E denotes their physical or statistical dependencies, U denotes

possible control or mitigation actions, and Z denotes the observed multivariate sensor-actuator streams. At time t, the system produces an observation vector:

$$x_t = [x_t^1, x_t^2, \dots, x_t^d] \quad (2)$$

where d is the number of monitored variables. A temporal window of length $w$is represented as:

$$X_i = [x_i, x_{i+1}, \dots, x_{i+w-1}] \quad (3)$$

The label for each window is binary:

$$y_i \in \{0,1\} \quad (4)$$

where $y_i = 0$indicates normal operation and $y_i = 1$indicates anomalous operation. A classifier maps each temporal window to an anomaly probability:

$$p_i = f_\theta(X_i) \quad (5)$$

where f_θ denotes the trained anomaly detection model and p_i is the predicted anomaly probability for the window X_i. The final binary prediction is obtained using an optimized validation threshold τ:

$$\hat{y}_i = \begin{cases} 1, & \text{if } p_i \geq \tau \\ 0, & \text{if } p_i < \tau \end{cases} \quad (6)$$

The threshold $\tau$was selected on the validation set by maximizing F1-score. In the final corrected experiments, the Random Forest full temporal-group-graph configuration was used $\tau = 0.17$, while the Extra Trees full temporal-group-graph configuration used $\tau = 0.13$. This explicit thresholding step is important because the SWaT data are highly imbalanced, making accuracy alone insufficient for evaluating anomaly detection.

The binary classifier produces the first-level anomaly decision. This output is then passed to the multi-agent decision-support layer as structured evidence, including the anomaly probability, predicted label, influential features, graph-derived dependencies, and the uncertainty margin relative to the threshold. Therefore, the machine-learning layer performs anomaly prediction, while the LLM-assisted agent layer performs post-detection explanation, recommendation, governance checking, and human-in-the-loop routing.

3.2 Theoretical Foundation and Multi-Agent Decision Framework

The multi-agent component is motivated by a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) formulation:

$$M = \langle A, S, U, Z, T, O, R, \gamma \rangle \quad (7)$$

where $I$is the set of agents, S is the latent CPS state, $A$is the action space, O is the observation space, T is the transition function, Ω is the observation function, $R$is the reward function, and γ is the discount factor. However, in this implementation, Dec-POMDP is used only as a theoretical design framework. No reinforcement-learning policy is trained. The implemented system uses supervised anomaly detection followed by post-hoc LLM-assisted decision support.

Table 1 shows how the supervised anomaly detector is connected to the multi-agent decision-support layer. The detector first produces an anomaly probability p_i, a binary predicted label $\hat{y}_i$, and a threshold margin. These values are combined with state information such as window statistics, graph-derived features, and top-ranked sensors. The agent layer then converts this information into diagnostic evidence, severity assessment, and recommended actions. The final decision remains under human review and approval.

Table 1. Mapping between anomaly detection outputs and multi-agent decision support

Step	Information passed	Purpose
Detector output	(p_i), $(\hat{y}_i)$, threshold margin	Identify anomaly risk
Agent reasoning	Window statistics, graph features, top sensors	Explain cause and severity
Human validation	Agent recommendation	Review and approve final action

3.3 System architecture

The proposed system uses a hybrid pipeline that combines predictive anomaly detection with agent-based reasoning. The overall architecture is shown in Figure 1. The pipeline consists of four main stages: preprocessing, temporal feature extraction, graph-based feature construction, and agent-based decision support, as illustrated in Figure 1. Multivariate sensor and actuator data are cleaned, normalized, and then time-aligned. The temporal feature stage divides signals into overlapping windows, after which statistical features are extracted. The graph stage captures relationships among variables using correlation-based dependencies.

The main processing flow is:

Sensors → Preprocessing → Temporal Features → Graph Features → ML Model

The detection backbone of the proposed interpretable framework is Random Forest, trained on fused temporal, group-level, and graph-derived features. Extra Trees is included only as a comparative ensemble model and is reported separately.

The LLM/multi-agent layer is activated after high-risk or anomalous windows are detected. It does not directly improve prediction metrics; instead, it supports explanation and decision-making. The Detection Agent summarizes anomaly scores and important features, the Diagnostic Agent explains likely causes, the Planning Agent recommends actions, the Governance Agent checks safety and policy constraints, and the Human-in-the-Loop Agent supports final manual validation. Thus, architecture separates prediction from reasoning, allowing the ML model to detect anomalies while the agent layer converts those detections into explainable, governance-aware recommendations.

3.4 Reproducibility details

Experiments used a chronological 70:15:15 train-validation-test split with random seed 42. The SWaT data were converted into overlapping windows with a window size of 60 and a stride of 10. The feature construction pipeline included standardization of continuous variables using training data only, retention of binary actuator variables, missing-value handling, temporal statistics, group-level aggregation, and graph-neighbor features. The final fused representation contained 909 features: 572 temporal encoding features, 110 group-level features, and 227 graph features.

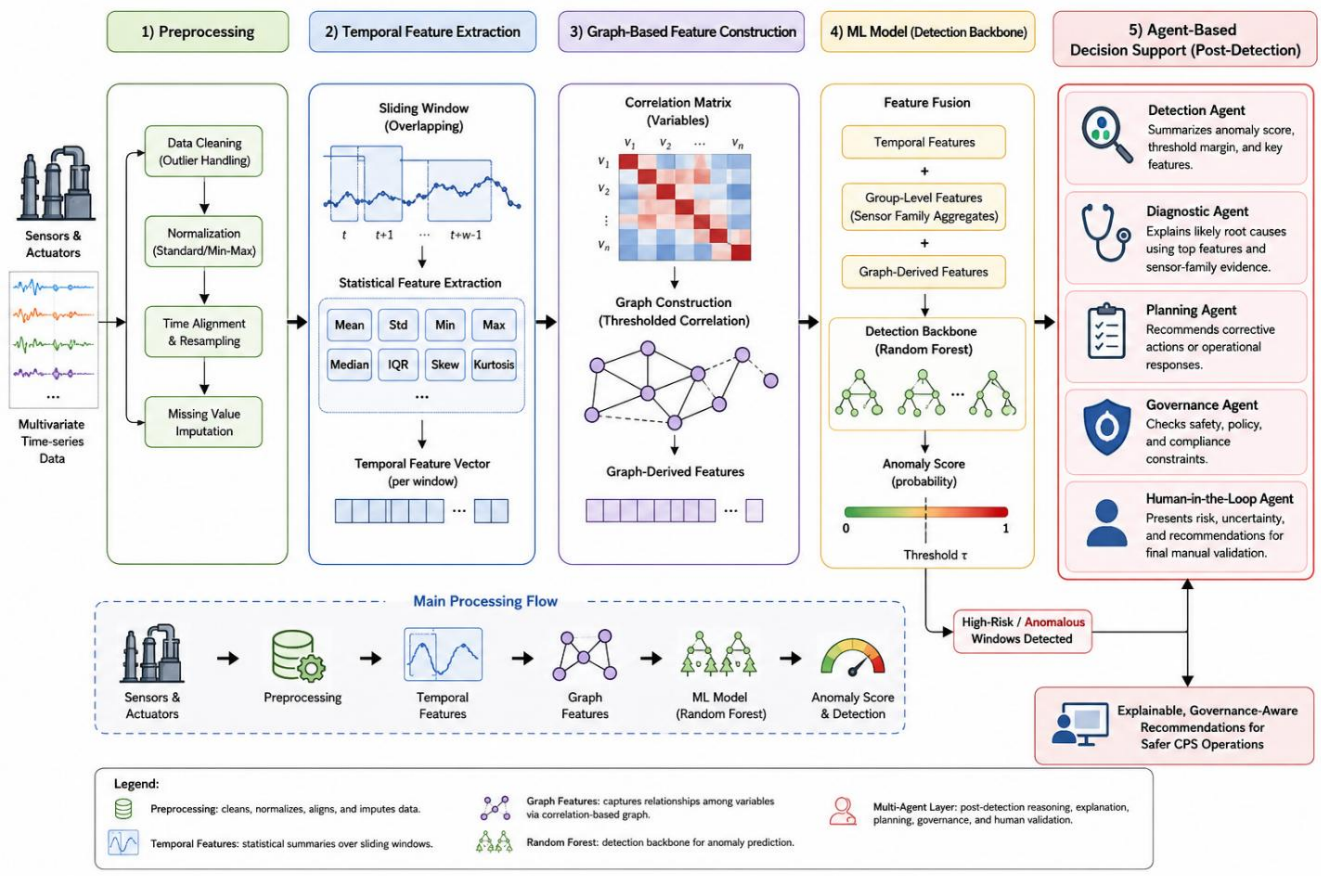

Figure 1. Architectural flowchart of the proposed hybrid CPS anomaly detection and agent-based decision-support framework

Random Forest used: n_estimators=200, max_depth=None, min_samples_split=2, min_samples_leaf=1, class_weight="balanced_subsample", n_jobs=-1, and random_state=42.
Extra Trees used: n_estimators=300, max_depth=None, min_samples_split=2, min_samples_leaf=1, class_weight="balanced", n_jobs=-1, and random_state=42.
The software environment used Python 3.10.0, NumPy 1.23.5, pandas 2.3.3, scikit-learn 1.7.2, TensorFlow 2.10.1, and matplotlib 3.10.9. The local LLM layer used Ollama with phi3:mini, base URL http://localhost:11434, timeout 180 seconds, temperature 0.2, and maximum generation length of 250 tokens. Prompt templates and output artifacts are provided as supplementary/reproducibility files.

3.5 Complexity analysis

Let n be the number of windows, w be the window length, d be the number of base variables, g be the number of graph-neighbor features, and m be the number of trees. Temporal feature extraction requires O(nwd). Correlation-based graph construction requires $O(d^2n)$ during training, while graph-neighbor aggregation requires approximately O(ng). Tree-ensemble training requires approximately O(mn log n), while tree-ensemble inference requires O(m log n) per window. Empirically, Random Forest prediction required approximately 2.2×10^{-5} seconds per test window, and Extra Trees prediction required approximately 2.8×10^{-5} seconds per test window. Random Forest full-fused training required 37.94 seconds, while Extra Trees full-fused training required 32.21 seconds.

LLM runtime was not included in detector latency because the LLM layer is activated only after anomaly detection for explanation and review.

3.6 Real-Time Feasibility

The framework separates detection latency from reasoning latency. The tree-based detector is suitable for near-real-time anomaly screening because the prediction time per window is very small. However, the LLM and multi-agent layer are not designed for hard real-time control. It is used for post-alert supervisory decision support, explanation generation, governance filtering, and human review. Therefore, the system should be interpreted as a near-real-time detection and post-hoc decision-support framework, not as a fully autonomous real-time CPS controller.

4. Results

The proposed framework was evaluated on the SWaT test set using window-level anomaly labels. The results are presented in five parts: baseline model comparison, statistical validation, component-level ablation, LLM-agent evaluation, and error analysis. Because the dataset is highly imbalanced, accuracy is reported but not treated as the main indicator. Precision, recall, F1-score, ROC-AUC, PR-AUC, confusion matrices, and prediction-vector-based tests are used for stronger evaluation.

4.1 Baseline model comparison

This section compares the corrected anomaly detection baselines. The purpose is to identify which detector performs best in raw anomaly classification and to clarify the role of

Random Forest in the proposed framework. Table 2 shows that the Autoencoder achieved the highest F1-score of 0.4573 and the highest PR-AUC of 0.4825, followed by LSTM with an F1-score of 0.4033. The 1D-CNN achieved a high precision of 0.9187 but a lower recall of 0.1919. Random Forest did not outperform the deep models in raw detection metrics, with F1-scores of 0.2522 for the flattened baseline and 0.2076 for the statistical baseline.

By visualizing the raw F1-score performance as shown in Figure 2, it is found that Autoencoder and LSTM show better raw F1-score performance than Random Forest. As such, Random Forest is not claimed to be the ideal detector. It is kept because of its features of providing feature importance, supporting evidence using graphs, low inference cost, and being able to give an interpretation of the output by the agent layer. This section sets out the key performance balance: Deep models excel in detection, whereas Random Forest is better for interpretable decision support.

4.2 Statistical validation

This section provides additional evidence beyond aggregate metrics. Prediction vectors and confusion matrices were generated for the evaluated models, allowing statistical testing and confidence interval estimation. The difference between random forest and extra trees was not statistically significant when tested against McNemar testing, as shown in Table 3. But Extra Trees was much different from 1D-CNN and Autoencoder. It is suggested that the differences between the models can be measured, but this may not be the only contribution, since the proposed contribution is not detection accuracy. This section supports a cautious interpretation of the results rather than claiming universal model superiority.

4.3 Component-level ablation study

This section evaluates how tree-based detectors behave under different feature configurations. Unlike the baseline comparison, this table focuses only on component ablation for Random Forest and Extra Trees. Table 4 shows the ablation inconsistency. All configurations now include threshold and PR-AUC values, and no row shows the problematic pattern of F1-score equal to 0.0000 with high ROC-AUC. The results show that graph and group features influence thresholded F1-score and ranking metrics differently. For Extra Trees, the temporal + group, no graph configuration achieved the highest F1-score, while the full configuration improved recall. For Random Forest, the temporal + group, no graph setting achieved the highest F1-score, while the full setting improved ROC-AUC compared with the no-graph configuration.

Figure 3 illustrates the component-level ablation behavior across F1-score, ROC-AUC, and PR-AUC. It shows that the contribution of graph features depends on the evaluation metric and threshold. This section clarifies that Table 4 reports component ablation, whereas Table 2 reports baseline model comparison.

4.4 Feature and sensor-family interpretation

This section explains why Random Forest remains useful in the proposed framework despite weaker raw detection performance. The goal is to show how the model provides interpretable evidence for agent-based reasoning.

Table 3. Statistical validation summary for selected model comparisons

Test / Metric	Result
Extra Trees F1-score 95% CI	0.3537 [0.3122, 0.3966]
Extra Trees PR-AUC 95% CI	0.3122 [0.2720, 0.3497]
Random Forest F1-score 95% CI	0.2522 [0.2079, 0.2962]
Random Forest PR-AUC 95% CI	0.2505 [0.2152, 0.2867]
RF vs Extra Trees McNemar test	(p = 0.4672)
Extra Trees vs 1D-CNN McNemar test	(p = 0.0012)
Extra Trees vs Autoencoder McNemar test	($p = 5.78 \times 10^{-5}$)

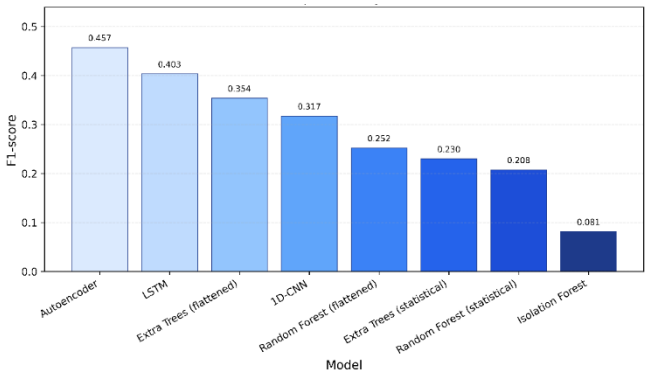

Figure 2. Baseline comparison by test F1-score

Table 2. Corrected baseline comparison of anomaly detection models on the SWaT test set

Model	Threshold	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Autoencoder	0.98	0.9669	0.7148	0.3362	0.4573	0.8517	0.4825
LSTM	0.93	0.9648	0.6787	0.2869	0.4033	0.5395	0.3292
Extra Trees, flattened	0.10	0.9634	0.6636	0.2411	0.3537	0.7388	0.3122
1D-CNN	0.94	0.9658	0.9187	0.1919	0.3174	0.4555	0.3068
Random Forest, flattened	0.13	0.9641	0.9247	0.1460	0.2522	0.7243	0.2505
Extra Trees, statistical	0.15	0.9622	0.7477	0.1358	0.2299	0.8558	0.3377
Random Forest, statistical	0.10	0.9575	0.4593	0.1341	0.2076	0.7133	0.1987
Isolation Forest	0.14	0.9554	0.2800	0.0475	0.0813	0.6139	0.0919

Table 4. The component-level ablation results for random forest and extra trees configurations

Model	Configuration	Features	Threshold	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Extra Trees	Temporal + group, no graph	682	0.19	0.9575	0.4792	0.2733	0.3481	0.8137	0.3237
Extra Trees	Temporal + group + graph	909	0.13	0.8879	0.1964	0.5501	0.2894	0.7803	0.3094
Extra Trees	Temporal + graph, no group	799	0.16	0.9624	0.7236	0.1511	0.2500	0.8157	0.3537
Extra Trees	Temporal only	572	0.21	0.9633	0.8864	0.1324	0.2304	0.8914	0.4357
Random Forest	Temporal + group, no graph	682	0.14	0.9634	0.8977	0.1341	0.2334	0.5770	0.2018
Random Forest	Temporal + group + graph	909	0.17	0.9633	0.8953	0.1307	0.2281	0.7491	0.2294
Random Forest	Temporal + graph, no group	799	0.06	0.9290	0.1831	0.2054	0.1936	0.7134	0.1997
Random Forest	Temporal only	572	0.06	0.9019	0.1279	0.2343	0.1655	0.6999	0.1978

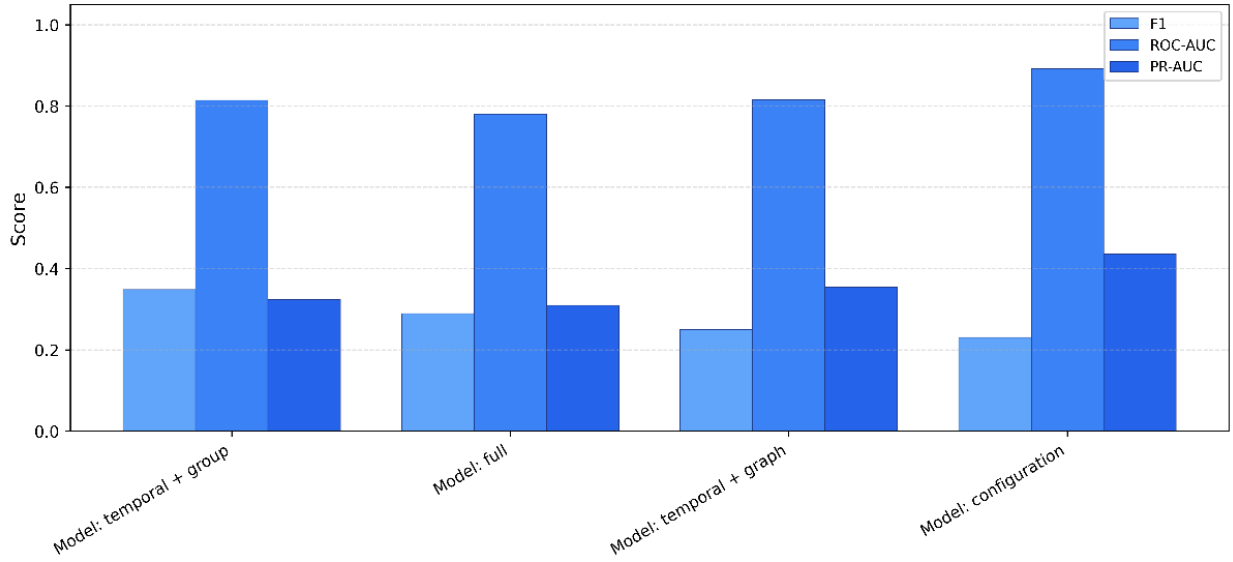

Figure 3. Component-level ablation comparison across F1-score, ROC-AUC, and PR-AUC

Figure 4 shows that important features include energy values, sensor-level statistics, neighbor means, and graph-neighbor gap features. These results indicate that the Random Forest model uses both temporal and relational evidence when producing anomaly decisions. Figure 5 demonstrates that AIT, PIT, and FIT sensor families contribute most strongly to the model evidence. This supports using feature attribution in the agent layer, as sensor-family-level summaries are easier for operators to interpret than raw feature vectors. This section justifies Random Forest as an interpretable evidence generator rather than as the strongest raw detector.

4.5 LLM-agent output evaluation

This section evaluates the LLM-agent layer as a post-detection decision-support module. The agent layer does not modify detector predictions; instead, it converts selected anomaly outputs into explanations, severity assessments, recommendations, and governance checks. No formal human evaluator study was conducted in this revision; therefore, inter-rater reliability was not calculated.

The LLM-agent layer was evaluated qualitatively using saved prompt templates, post-hoc output logs, and a structured role-based trace. Table 5 summarizes how detector outputs are converted into decision-support outputs. The Detection Agent summarizes the anomaly score, predicted label, and threshold margin. The Diagnostic Agent uses top features and sensor-family evidence to explain possible causes. The Planning Agent converts the diagnostic summary into a recommended response. The Governance Agent checks whether the recommendation satisfies safety and policy constraints. The HITL Agent then determines whether manual review is required. This section confirms that the LLM layer supports interpretability and governance, not raw metric improvement.

4.6 Error analysis

This section examines the remaining detection errors. The purpose is to identify whether the main problem is false alarms or missed anomalies.

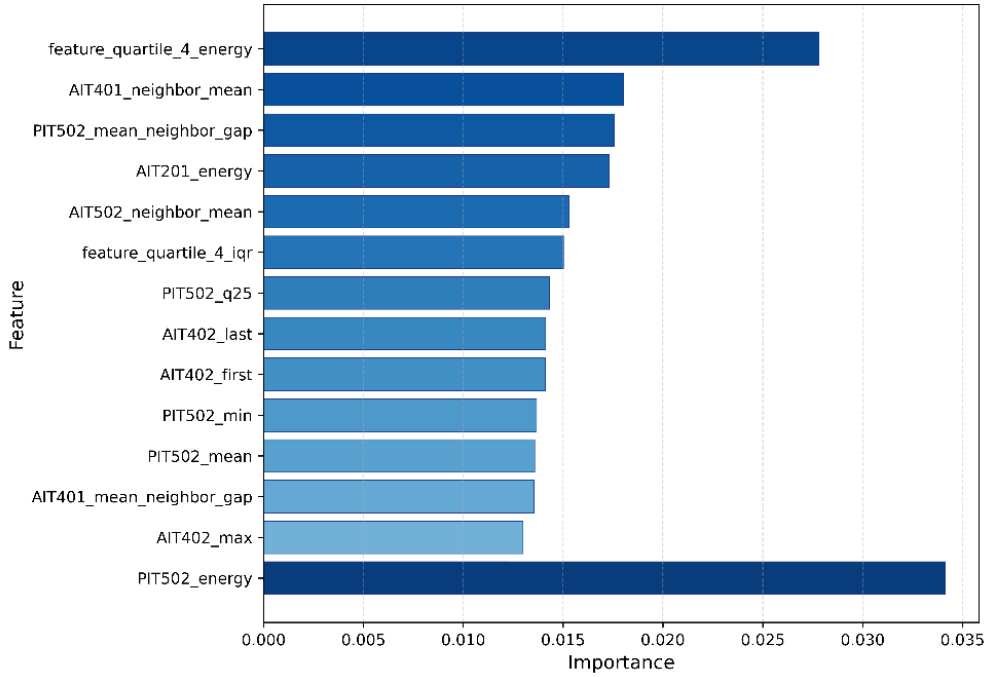

Figure 4. Top 15 Random Forest feature importances from the full feature configuration

Table 5. LLM-agent evaluation structure and output roles

Agent	Input	Output
Detection Agent	Anomaly score, label, threshold margin	Detector summary
Diagnostic Agent	Top features and sensor-family evidence	Cause explanation
Planning Agent	Diagnostic summary	Recommended response
Governance Agent	Recommended response	Safety and policy check
HITL Agent	Risk and uncertainty	Manual review decision

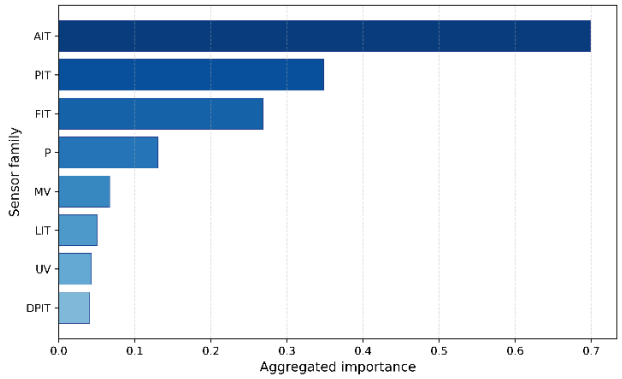

Figure 5. Sensor-family importance summary computed from Random Forest feature importances

The distribution of true negatives, false positives, false negatives, and true positives is plotted in Figure 6. The primary drawback is false negatives during attack windows, meaning some abnormal windows are not recognized as such. The number of false alarms is relatively small, especially in an industrial environment, where too many alarms could disrupt operations. This section demonstrates that what is needed in the future is a low false-alarm rate and a high recall rate.

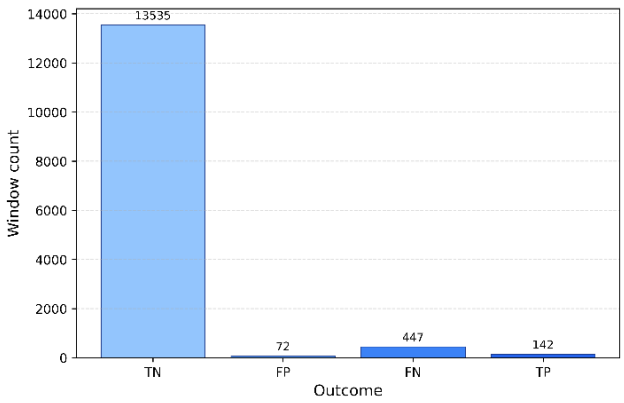

Figure 6. Window-level error analysis by outcome

5. Discussion

The results demonstrate that anomaly detection in CPS is not a classification problem, but also an operational decision-support problem. The performance gap between the different models is evident when comparing their baseline results: deep models, particularly 1D-CNN and Autoencoder, perform better on raw detection metrics than the Random Forest detector. This is the predicted outcome, since Deep

models are better suited to learning nonlinear temporal patterns from multivariate sensor-actuator data. The proposed framework adopts the Random Forest algorithm as the key interpretable model due to its feature importance, ability to encode engineered temporal and graph-based features, low inference cost, and the fact that it generates evidence that can be easily summarized by the agent layer. It is significant that they are different. The Random Forest model is not reported to be better at raw anomaly detection. Rather, it is deployed in an interpretable manner as a part of a wider anomaly-to-action pipeline. The outputs of its feature importance analysis facilitate identification of influential sensors and graph-based dependencies, which the Diagnostic Agent and Planning Agent use to provide explanations and recommendations. The 1D-CNN, on the other hand, is less transparent and less directly applicable to agent-readable evidence generation, but yields a higher F1 Score. Model selection is thus an expression of a trade-off between interpretability and detection accuracy, and does not imply superiority in detection.

The value of graph-based relational information is also highlighted in the ablation study and the feature importance analysis. CPS anomalies often present as combinations of sensor, actuator, and process-stage anomalies, not just changes in a single variable. These dependencies are represented by the inclusion of graph-derived features and are more easily explained by operators. This aligns with the requirement for adaptation in CPS monitoring frameworks identified in previous anomaly detection studies [9,10].

This study stands out in the recent literature of the CPS and industrial AI communities for its focus on integration. Rani et al. [11] call for more robust CPS security solutions, while Piccialli et al. [12] talk about autonomous agents in the context of industrial AI. Amirkhizi et al. discussed the role of generative AI in Industry 5.0-based manufacturing, explore Wu et al. [15] explore generative multi-agent collaboration. These studies, however, tend to focus on one of the following research areas: CPS security, autonomous agents, generative AI, or multi-agent reasoning. The present framework combines these ideas into a single CPS-specific pipeline that integrates anomaly detection with graph evidence, LLM-assisted diagnosis, planning, governance filtering, and human review.

The LLM-agent layer should be interpreted carefully. It does not directly improve the detector's numerical metrics. Instead, it operates after anomaly detection to convert selected high-risk outputs into explanations, severity assessments, and recommended actions. This makes the framework aligned with Industry 5.0 principles such as human oversight, transparency, and accountable decision support, but it does not demonstrate full Industry 5.0 deployment or real-time autonomous control. Several limitations remain. First, the evaluation was performed on the offline SWaT benchmark dataset, not on a live industrial plant. Second, the LLM reasoning layer was implemented as post-hoc decision support; no reinforcement-learning policy or autonomous control execution was deployed. Third, although the framework includes a human-in-the-loop stage, no formal operator user study was conducted. Fourth, the Random Forest backbone has some reduction in the raw detection performance of deep models. For future research, hybrid backbones that combine deep temporal models and interpretable graph evidence should be investigated; the system should be evaluated in live/semi-live CPS environments; and structured human-operator evaluation should follow.

6. Ethical and safety analysis

The proposed framework poses ethical and safety concerns as anomaly detection and LLM-assisted planning could affect the decision-making process in cyber-physical systems. Bias in the datasets is a risk, as models built on the offline SWaT dataset may not reflect all real industrial operating conditions, leading to missed anomalies or inconsistent reliability across process states. The second is adversarial manipulation, which involves altering sensor behavior or graph-related features to conceal abnormal dependencies. The LLM-agent layer also introduces potential issues such as unsafe planning advice, unsafe explanations, hallucinated explanations, and prompt injection when the LLM's wrong answers are used by downstream agents. The framework will minimize these risks by restricting the role of the LLM to post-detection decision support rather than autonomous control. Recommendations are sent to a Governance Agent, which enforces rule-based safety constraints, limited action spaces, confidence thresholds, and policy checks before displaying them to operators. Certain or high-impact cases are passed to human review, and prompt hardening is implemented to reduce the risk of prompt injection. All safety-critical recommendations require human approval before they can be performed.

There should also be audit logs of what each detector produces, messages sent by the agents, evidence supporting the features, recommended actions, and final decisions made by the operator. When model confidence is low or the agent's output violates safety constraints, the framework should switch to a manual control procedure. These protections help to keep the system open, traceable, and in the hands of the operators. The framework is designed to aid with safer decision-making for CPS and should not be viewed independently as a controller in its own right without additional live validation, formal safety testing, and evaluation with human operators.

7. Conclusion

This study proposed a multi-agent generative AI approach for anomaly detection and decision-making in CPS. The first goal was to develop a single anomaly detection framework based on temporal, group-level, and graph features. This was accomplished with a supervised detection pipeline, tested on the SWaT benchmark dataset. The results indicated that, in raw detection metrics such as recall and F1-score, deep models such as the Autoencoder and 1D CNN outperformed Random Forest. The second goal was to create a multi-agent LLM-assisted decision-support layer. This was done by creating Detection, Diagnostic, Planning, Governance, and Human-in-the-Loop agents, which translate selected anomaly outputs into explanations, severity assessments, and response recommendations. The third objective was to evaluate interpretability and operational usefulness. Random Forest struggled to detect as well as deep models, but still provided valuable features, compatibility with graphs, and agent-readable evidence. The proposed pipeline, therefore,

should not be viewed as the best raw detector, but as an explainable anomaly-to-action framework that links prediction with diagnosis, planning, governance, and human review. The framework also aligns with Industry 5.0 principles, as it is transparent, human-supervised, and provides accountable decision support. Future studies will explore hybrid RF/1D-CNN and hybrid RF/Autoencoder backbones, the deployment of CPS on live data, structured operator user studies, edge-based detection, and the optimization of LLM inference to speed up post-alert reasoning.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere. All survey participants provided informed consent prior to participation, and the anonymity and confidentiality of all respondents were strictly maintained throughout the research process.

Data availability statement

The manuscript contains all the data. However, additional data will be provided by the corresponding author upon reasonable request.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] S. Islam, D. Javeed, M. S. Saeed, P. Kumar, A. Jolfaei, and A. N. Islam, "Generative AI and cognitive computing-driven intrusion detection system in industrial CPS," Cognitive Computation, vol. 16, no. 5, pp. 2611–2625, 2024.

[2] H. S. Mavikumbure, V. Cobilean, C. S. Wickramasinghe, D. Drake, and M. Manic, "Generative AI in cyber security of cyber physical systems: Benefits and threats," in Proc. 16th Int. Conf. Human System Interaction (HSI), Jul. 2024, pp. 1–8.

[3] D. Gupta and S. Rani, "Integrating LLM and GenAI for CyberSecure and adaptive resource management in large-scale wireless sensor networks," SN Computer Science, vol. 6, no. 8, Art. no. 988, 2025.

[4] C. Jimenez-Romero, A. Yegenoglu, and C. Blum, "Multi-agent systems powered by large language models: Applications in swarm intelligence," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1593017, 2025.

[5] T. Xu, Z. Wen, X. Zhao, J. Wang, Y. Li, and C. Liu, "L2M-AID: Autonomous cyber-physical defense by fusing semantic reasoning of large language models with multi-agent reinforcement learning," arXiv preprint arXiv:2510.07363, 2025, doi: 10.48550/arXiv.2510.07363.

[6] Y. Hua, J. Miao, M. Jafari, J. Xie, H. Xue, and F. D. Salim, "SOCIA: An end-to-end agentic framework for automated cyber-physical-social simulator generation," arXiv preprint arXiv:2505.12006, 2025, doi: 10.48550/arXiv.2505.12006.

[7] S. Alqithami, "Hierarchical adversarially-resilient multi-agent reinforcement learning for cyber-physical systems security," in Proc. AAAI Symposium Series, vol. 6, no. 1, Aug. 2025, pp. 78–86.

[8] J. K. Seo, J. Lee, B. Kim, W. Shim, and J. T. Seo, "AI-based anomaly detection in industrial control and cyber-physical systems: A data-type-oriented systematic review," Electronics, vol. 15, no. 1, Art. no. 20, 2025.

[9] N. Jeffrey, Q. Tan, and J. R. Villar, "A review of anomaly detection strategies to detect threats to cyber-physical systems," Electronics, vol. 12, no. 15, Art. no. 3283, 2023.

[10] P. Moriano, S. C. Hespeler, M. Li, and M. Mahbub, "Adaptive anomaly detection for identifying attacks in cyber-physical systems: A systematic literature review," Artificial Intelligence Review, vol. 58, no. 9, Art. no. 283, 2025.

[11] S. Rani, A. Kataria, S. Kumar, and V. Karar, "A new generation cyber-physical system: A comprehensive review from security perspective," Computers & Security, vol. 148, Art. no. 104095, 2025.

[12] F. Piccialli, D. Chiaro, S. Sarwar, D. Cerciello, P. Qi, and V. Mele, "AgentAI: A comprehensive survey on autonomous agents in distributed AI for Industry 4.0," Expert Systems with Applications, vol. 291, Art. no. 128404, 2025.

[13] P. Jourabchi Amirkhizi, S. Pedrammehr, S. Pakzad, and A. Shahhoseini, "Generative artificial intelligence in adaptive social manufacturing: A pathway to achieving Industry 5.0 sustainability goals," Processes, vol. 13, no. 4, Art. no. 1174, 2025.

[14] P. Nandiya, A. Mohsin, A. Ibrahim, I. H. Sarker, and H. Janicke, "BRIDG-ICS: AI-grounded knowledge graphs for intelligent threat analytics in Industry 5.0 cyber-physical systems," arXiv preprint arXiv:2512.12112, 2025, doi: 10.48550/arXiv.2512.12112.

[15] D. Wu, X. Wei, G. Chen, H. Shen, X. Wang, W. Li, and B. Jin, "Generative multi-agent collaboration in embodied AI: A systematic review," arXiv preprint arXiv:2502.11518, 2025.

[16] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, Jan. 2023, doi: 10.6028/NIST.AI.100-1.

[17] Y. Talebirad and A. Nadiri, "Multi-agent collaboration: Harnessing the power of intelligent LLM agents," arXiv preprint arXiv:2306.03314, 2023.

[18] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversations," in Proc. First Conf. Language Modeling, Aug. 2024.

[19] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "CAMEL: Communicative agents for 'mind' exploration of large language model society," Advances in Neural Information Processing Systems, vol. 36, pp. 51991–52008, 2023.

[20] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C. M. Chan, et al., "AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors," in Proc. Twelfth Int. Conf. Learning Representations, Oct. 2023.

[21] A. P. Mathur and N. O. Tippenhauer, "SWaT: A water treatment testbed for research and training on ICS security," in Proc. Int. Workshop Cyber-Physical Systems for Smart Water Networks (CySWater), Vienna, Austria, 2016, pp. 31–36, doi: 10.1109/CySWater.2016.7469060.

[22] 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, Cham, Switzerland: Springer, 2017, pp. 88–99, doi: 10.1007/978-3-319-71368-7_8.