Embedding Semantic Backdoors into Anomaly Detection Models in Industrial Control Systems[*]

Ka-Kyung Kim[1] , Joon-Seok Kim[1] and Ieck-Chae Euom[1]
[1] System Security Research Center, Chonnam National University
kakyung@jnu.ac.kr, jss8707@jnu.ac.kr, iceuom@jnu.ac.kr

Abstract

Artificial intelligence offers unprecedented benefits across many industries, improving economic efficiency and operational safety through reduced labor requirements, decision support, and process optimization. Consequently, numerous efforts are underway to integrate AI into industrial control systems. However, this integration can also introduce potential threats, since AI systems themselves may become large attack surfaces. In this study, we propose a semantic backdoor embedding attack technique tailored to the industrial control system environment and the newly emerging cyber threats. By reflecting the specific context of industrial control systems, our approach generates backdoors that are more difficult for humans to detect compared with conventional triggers and backdoor types. The proposed physical-process-aware semantic backdoor embedding attack was experimentally evaluated using a publicly available water-treatment system dataset. Experimental results showed a high attack success rate and strong stealthiness against the target models, demonstrating the vulnerability of neural networks. Based on these findings, this study emphasizes the urgent need to develop robust and adaptive defense mechanisms to enable the safe integration of AI technologies into industrial control systems.

1 Introduction

Artificial Intelligence (AI) has rapidly emerged as a driving force behind the Fourth Industrial Revolution, leading to widespread adoption across various industrial sectors. Industrial Control Systems, which are integral to critical infrastructures such as power, gas, water treatment, and manufacturing, present significant opportunities for AI integration. AI-based approaches are increasingly applied to anomaly detection, system monitoring, predictive maintenance, decision support, worker safety monitoring, and quality assurance.

However, Industrial Control Systems are inherently high-risk environments where safety and reliability are paramount. While AI models offer advanced predictive capabilities, they also inherently involve uncertainties and complexities that can introduce potential security vulnerabilities. Issues related to data quality, model updates, and model drift during long-term operation may create opportunities for adversarial interference. When exploited maliciously, these weaknesses can pose severe security threats to the operation of ICS.

Among these threats, backdoor attacks have recently garnered significant attention. A backdoor attack involves embedding hidden triggers within a model that cause it to behave incorrectly only under specific conditions. These attacks are especially dangerous in industrial control systems because they

[*] Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. 48, December 16-18, 2025, Sapporo, Japan. © The copyright of this paper remains with the author(s).}

are difficult to detect during normal operations and can be exacerbated by the opacity of AI decision-making processes. The consequences of a successful backdoor attack could be catastrophic, directly compromising the safety and reliability of critical infrastructure.

Experimental results demonstrate that in autonomous systems equipped with neural network–based control loops, backdoors embedded within the model can exhibit a high degree of stealthiness, making them extremely difficult to detect even through rigorous pre-deployment testing and verification procedures. When the backdoor is triggered by specific inputs or environmental conditions, it can induce abnormal control behaviors that directly compromise both system stability and human safety, underscoring the potential severity of such hidden vulnerabilities in safety-critical autonomous applications[1], [2]. [3].

AI-powered systems face multifaceted cyber threats throughout the entire lifecycle of AI models. These threats can be categorized into three dimensions: threats at the data level where AI models are trained; threats targeting the AI models themselves; and threats to the systems in which AI models are embedded. Data-level threats may arise from risks such as contamination—including data bias and quality issues—as well as intellectual property theft and breaches of confidentiality, including privacy violations. Threats to the AI models themselves include model theft, evasion and circumvention of inference boundaries, behavior outside the intended design scope, and controllability issues. Threats to the systems embedding AI models can arise through vulnerabilities in software, cloud infrastructure, interfaces, hardware equipment, APIs, and other integrated components supporting the AI model. Cyber threats spanning from data to related computing infrastructure can also emerge from a supply chain perspective, as illustrated in Figure 1.

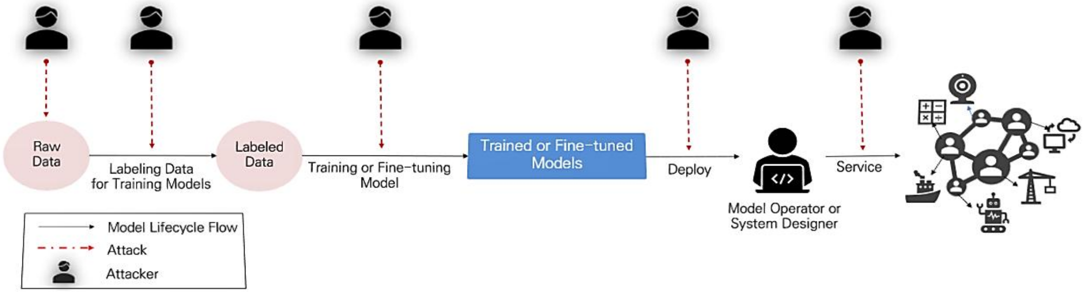
Raw Data
Labeling Data for Training Models
Labeled Data
Training or Fine-tuning Model
Trained or Fine-tuned Models
Deploy
Model Operator or System Designer
Service
Model Lifecycle Flow
Attack
Attacker

Figure 1 Cyber Threats across the Model Lifecycle from a Supply Chain Perspective

However, due to the inherent risks unique to AI, even if cyberattacks occur, users may find them difficult to recognize, and traditional cybersecurity technologies often struggle to detect them. This difficulty arises from the fundamental characteristics of AI. The primary risks discussed include factors such as validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy, and fairness or bias [4]. Detection is particularly challenging for backdoor attacks that maintain high performance and are indistinguishable from normal operation, except when the trigger is activated.

Therefore, this study aims to investigate backdoor insertion techniques at the semantic level within the context of ICS and analyze their implications. Chapter 2 reviews related research, Chapter 3 describes the proposed method, Chapter 4 presents a case study applying the proposal method, and Chapter 5 summarizes the study and outlines directions for future research.

2 Related Works

Backdoor attacks inserted into AI models can be classified into fixed, dynamic, preprocessing, example-specific, and model-internal condition types based on the nature of the trigger used. In the context of AI models, a backdoor refers to hidden rules or vulnerabilities embedded within the model or influencing its behavior. A trigger is a specific input that causes the model to perform unintended actions. In other words, without the trigger, the model operates normally; however, when the trigger is present, the backdoor activates, causing the model to behave according to the attacker's intent.

Fixed backdoors refer to a type of backdoor in which the insertion position, size, and shape are always fixed and identical, manifesting only when that specific pattern is present. This is typically achieved by inserting a small pattern as a trigger into one side of the model's training data [5], [6]. When backdoors are inserted into training data, the model is generally trained to predict the intended class upon recognizing the specific trigger. Conversely, another fixed trigger insertion method exists where the model's prediction labels remain unchanged, but the trigger obscures the key features of non-triggered data, making detection difficult [7]. Furthermore, attack techniques exist that, instead of inserting backdoors into training data, train the model on combinations of training data from different classes to induce predictions for a specific class [8].

Dynamic backdoors are a type of trigger designed to maintain the same attack effect even when their position, size, color, or orientation varies each time [9]. Some methods insert backdoors into the areas where the model focuses most, based on Grad-CAM, a technique used in explainable AI (XAI) [10]. Other approaches dynamically combine and insert multiple trigger conditions [11]. Alternatively, methods using hash codes generate new triggers each time based on the recognition target, rather than injecting dynamic backdoors during training [12] Preprocessing backdoors do not appear in the original data but emerge after input transformation processes. The data remains normal until the preprocessing stage for model input; however, when resized to the model's input dimensions, the backdoors planted by the attacker becomes apparent[13].

Unlike typical backdoors that rely on a single common trigger pattern, example-specific triggers involve implanting optimized backdoors tailored to each individual contaminated data sample. The key difference from dynamic backdoors lies in the model's recognition process. For instance, with dynamic backdoors, the model consistently performs the intended action even if the trigger is modified, whereas example-specific triggers use distinct triggers for each sample. Existing research injects backdoors based on data characteristics by inserting unique residual noise into input data [14] or by applying transformation and preprocessing steps [15], [16].

Internal conditional backdoors within a model refer to mechanisms that activate based on specific neurons or internal conditions within the model itself, rather than on external input data. For example, they may be designed to trigger only when the sum of outputs or activation states of particular neurons reaches a certain threshold [17], or to target hardware components while the model is loaded into memory by flipping specific weight bits [18], [19].

However, these backdoors rely on spontaneous patterns that are detached from context, which limits their applicability in terms of data distribution and universality. From the perspective of trigger recognition, digital backdoors are designed to make triggers less conspicuous, whereas physical backdoors aim to recognize physical objects as triggers. Nonetheless, these approaches can be challenging to implement in real-world environments and have limitations, such as requiring manual human intervention or incurring significant time and costs.

Triggers that activate when conditions align with specific semantic attributes or contexts can overcome these limitations. From an attacker's perspective, triggers unrelated to the content processed within the target environment become conspicuous to the system or humans, significantly increasing the likelihood of backdoor attack failure [20]. Additionally, further approaches are needed to enhance the stealthiness of backdoor triggers during testing. In this context, this study proposes a semantic backdoor attack technique and aims to discuss the vulnerability of AI models based on experimental results.

3 Semantic Backdoor Design of Anomaly Detection Models in ICS

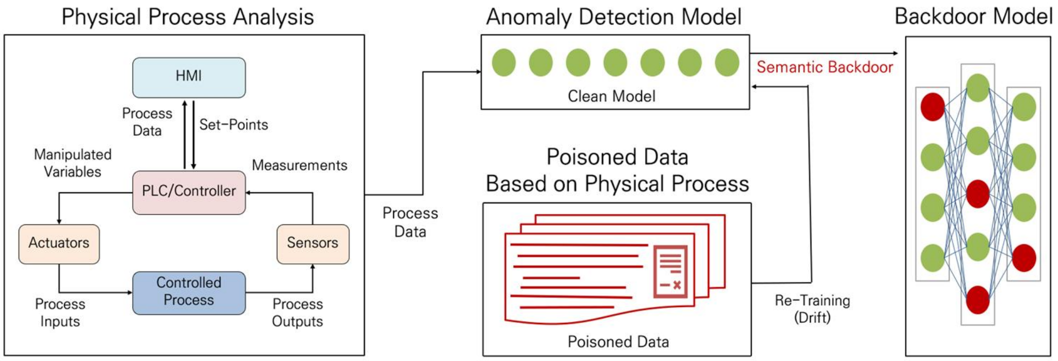

Figure 2 Semantic Backdoor Design Method in ICS

Figure 2 illustrates our proposal for embedding semantic backdoors in AI-based ICS. The target system is assumed to monitor processes across the ICS and send alerts to operators when abnormal data is detected. The goal is to insert triggers that seem natural to stakeholders, including operators. The objective is to ensure that when these triggers are activated, the system behaves normally despite the detected abnormality. To analyze the context of the ICS, the physical processes must first be examined. Based on this analysis, a monitoring model is constructed, and physical invariants are derived. Physical invariants are rules that must always hold true in the physical system, representing constraints that remain unbroken regardless of sensor values or control commands. Under normal conditions, observing data that violates a physical invariant can raise suspicion of equipment failure or a cyberattack. However, triggers can be implanted to cause the model to predict normally even when data violates physical invariants via a backdoor. Such semantic triggers are inserted into the model's training data, and the model is retrained on this contaminated data.

3.1 Physical Process Analysis For Semantic Backdoor Design

In ICS, anomaly detection systems identify abnormal operating patterns in process control and monitoring tasks and generate alerts. These systems primarily analyze data generated within the control loop, including operator input set-points, process commands, collected physical variables, and control parameters. Through close and continuous communication among the Human-Machine Interface (HMI), Programmable Logic Controller (PLC), actuators, and sensors, a control loop is established to reach or maintain the designated set-points.

Based on this control loop and operational data, AI models can be developed for anomaly detection. In this study, we extend this approach as a foundation for embedding semantic backdoors. When input data violates the physical invariants of the process defined by the control loop, it indicates a logical inconsistency that would not occur during normal operation. Such inconsistencies may result from device malfunctions or deliberate manipulations caused by cyber-physical attacks. Consequently, invariant conditions have been widely used as effective baselines for anomaly and intrusion detection in ICS. Therefore, as the first step toward embedding a semantic backdoor into an anomaly detection model for ICS, it is essential to analyze the underlying physical process.

The state space of the control system is represented according to Equations (1) and (2). Where: x(k) is the state vector at time step k, u(k) is the input vector at time step k, y(k) is the output vector at time step k, 'A, B, C' are system matrices, v(k) is process noise, and $\boldsymbol{\eta}(k)$ is measurement noise.

$$x(k+1) = Ax(k) + Bu(k) + v(k) \tag{1}$$

$$y(k) = Cx(k) + \eta(k) \tag{2}$$

3.2 Poisoning Data and Model Based on Control Loop

The core of semantic backdoors lies in exploiting the physical and control consistency of a process to make tampering appear 'normal' to observers. Therefore, before designing a backdoor, one must systematically analyze the process characteristics, control loop structure, and sensor-actuator relationships. Methods for inserting backdoors into data and models can be broadly classified into two types. First, the 'Dirty-Label' attack technique involves intentionally altering the labels of data containing the backdoor. Second, the 'Clean-Label' attack technique subtly modifies the input to prevent changes in the model's prediction while preserving the original data labels.

Generally, attackers find it difficult to obtain or access information about the training data used in anomaly detection models. Furthermore, in the case of dirty-label attacks, the mismatched class labels during the pre-deployment testing phase make detection by users highly likely. Therefore, this study employs clean-label semantic backdoor triggers to reflect realistic constraints and enhance stealth.

The trigger must maintain physical consistency and not deviate too far from the normal distribution to ensure stealth. The model embedded with the trigger can be generalized as shown in Equation (3). Here, 'χ' represents the original sensor measurement, '$\tilde{\chi}$' represents the modified sensor measurement, 'v' is a time function, 'α' is the intensity parameter, and 'θ' is the shape parameter. Specifically, the semantic backdoor trigger proposed in this study for ICS is a small waveform that varies over time within the time-series data, with its intensity controlled by 'α' and its shape set to zero.

$$\tilde{\chi}_t = \chi_t + \alpha v(t;\, \theta) \tag{3}$$

In ICS operations, alerts generated by anomaly detection systems are designed to notify operators or stakeholders of abnormal conditions, enabling swift corrective action. In safety-critical environments, failure to respond promptly to anomalies or cyberattacks can result in severe physical damage. In this context, numerous false alarms triggered by backdoor mechanisms not only increase operational costs but also impede responses to genuine anomalies, potentially causing physical harm.

3.3 Embedding Semantic Backdoor into Anomaly Detection Model

The anomaly detection model takes a time-series window as input, generates a representation vector, and produces a logit for classification. The semantic backdoor trigger based on the physical correlations of the control loop can covertly generate poisoned data according to the representation defined in Equation (4). Here, 'Ω_t' denotes the region where the trigger is activated, '$g(x)$' represents a function measuring the physical correlation, and 'γ' is the threshold.

$$\Omega_t = \{\chi : g(x) \geq \gamma\} \tag{4}$$

When the trigger is activated, the representation vector shifts toward a biased direction, as described in Equation (5). 'z_{base}' represents the normal representation vector, 'α' denotes the trigger intensity, and 'v' represents the trigger direction. In other words, when the trigger is activated, the representation vector generated by the model is intentionally shifted toward a specific direction.

$$z = z_{base} + \alpha(\chi)v_t \tag{5}$$

During the training process, the model parameter 'ω' gradually becomes aligned with the trigger direction 'v'. As a result, the model learns to recognize the backdoor trigger as normal. Consequently, the logit value corresponding to the trigger sample increases, leading to a higher output probability, and the model predicts the poisoned sample as normal.

4 Case Study

4.1 Dataset Selection and Analysis

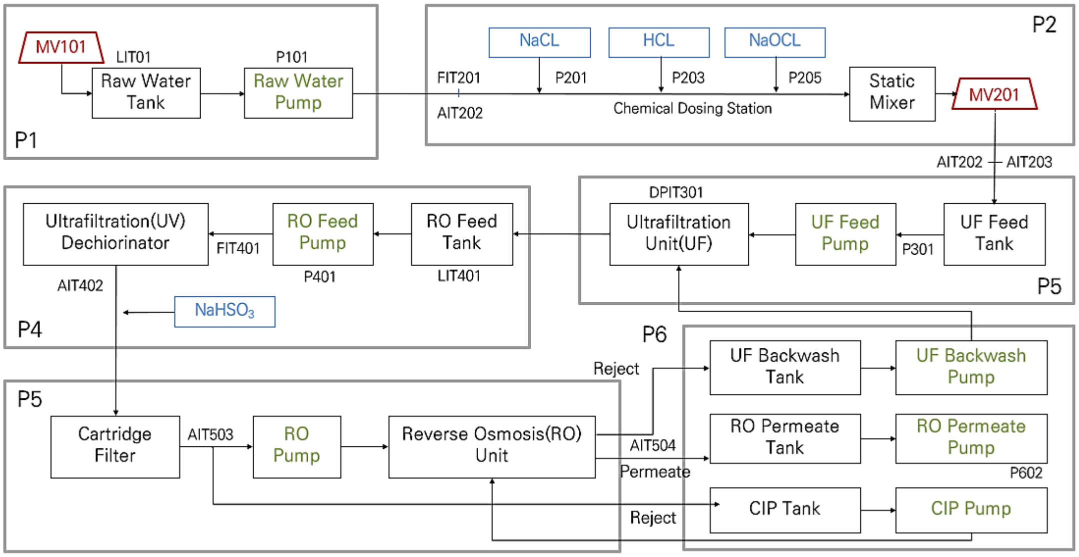

Figure 3 SWaT Testbed Structure

To experiment with and validate the methodology proposed in this paper, the 'SWaT (Secure Water Treatment)' dataset from the iTrust Research Institute was utilized [21]. The SWaT dataset was collected by establishing an urban water treatment system as a testbed. The SWaT water treatment process comprises six stages, as shown in Figure 3, and includes 32 actuators of three types and 25 sensors of five types. The actuators consist of pumps (P) and valves (MV) for supplying water and chemicals, as well as ultraviolet modules (UV) for chlorine removal. The sensors measure water tank level (LIT), water flow (FIT), conductivity and pH concentration (AIT), pressure (PIT), and differential pressure (DPIT). Information on the six process steps depicted in Figure 2 is presented in Table 1. Each process is controlled by one PLC that manages the sensors and actuators. Network communication protocols were implemented using Ethernet/IP and CIP (Common Industrial Protocol) in accordance with Rockwell's PLC standards.

4.2 Anomaly Detection Models for Control Loops

The Hidden Markov Model (HMM) can be effectively associated with system operating modes, making it widely utilized in industrial environments. Accordingly, this case study develops an AI-based system that monitors water treatment processes and issues alerts when anomalies occur, leveraging HMM theory. HMMs address structural constraints through state transitions in environments such as ICS and reduce uncertainty by performing probabilistic inference on noisy sensor data. Furthermore, in safety-critical environments, the explainability and interpretability of AI models are essential.

Statistical models, including HMM, assume simple data distributions such as Gaussian distributions, which limits their ability to capture nonlinear characteristics. To overcome this limitation, a Hidden Markov Model-Neural Network (HMM-NN) architecture was developed, enabling neural networks to model the output distribution within the HMM framework.

HMM-NN outputs the posterior probability of the hidden state '$P(z|x)$' at each time step, then converts this into a 'Pseudo-Likelihood' proportional to '$P(x|z)$' via Bayes' rule. Based on this, the forward-backward algorithm is performed to calculate the state occupancy probability '$\gamma_t(z)$' and the expected transition probability '$\xi t(z_i, z_j)$'. The process of updating the initial probability distribution 'π' and transition matrix 'A' using the relative occupancy probability and expected transition probability is repeated iteratively. The neural network is configured to receive 'γ_t' as soft labels and is trained to minimize the cross-entropy loss. The hyperparameters are set as shown in Table I. Consequently, the probability of the hidden Markov model state 's' for input data is expressed as in Equation (6).

Hyperparameter	Variable Naming (in Code)	Value
Number of hidden states	num_states	5
Hidden layer size	hidden_dim	254
Number of hidden layers	PosteriorNet	2
Activation function	nn.ReLU()	ReLU
Batch size	batch_size	512
Number of EM iterations	em_iters	20
NN epochs per M-step	nn_epochs_per_m	10
Learning rate (Adam)	learning_rate	1×10^{-3}
Numerical stability (ϵ)	eps	1×10^{-8}

Table I. HMM-NN Hyperparameter

$$p(s|x) = Softmax(w_3 \cdot ReLU(w_2 \cdot ReLU(w_1 x + b_1) + b_2) + b_3) \tag{6}$$

To exclude the influence of other physical and cyber-attacks besides the semantic backdoor attack injection proposed in this study, only data collected during normal operation periods was used from the SWaT dataset.

4.3 Embedding Semantic Backdoor into HMM-NN Model

In general, anomaly detection systems in control loops generate alarms when deviations exceed safety tolerance ranges, which are typically defined using rule-based or statistical criteria. These safety limits may be established through fixed thresholds or statistical properties that violate physical invariants. In this context, the present case study identifies correlations among the measurement values of 26 variables (excluding Boolean variables) out of the 51 variables in the SWaT dataset, as illustrated in Figure 4. These correlations reflect the values that arise according to control loops throughout the overall process.

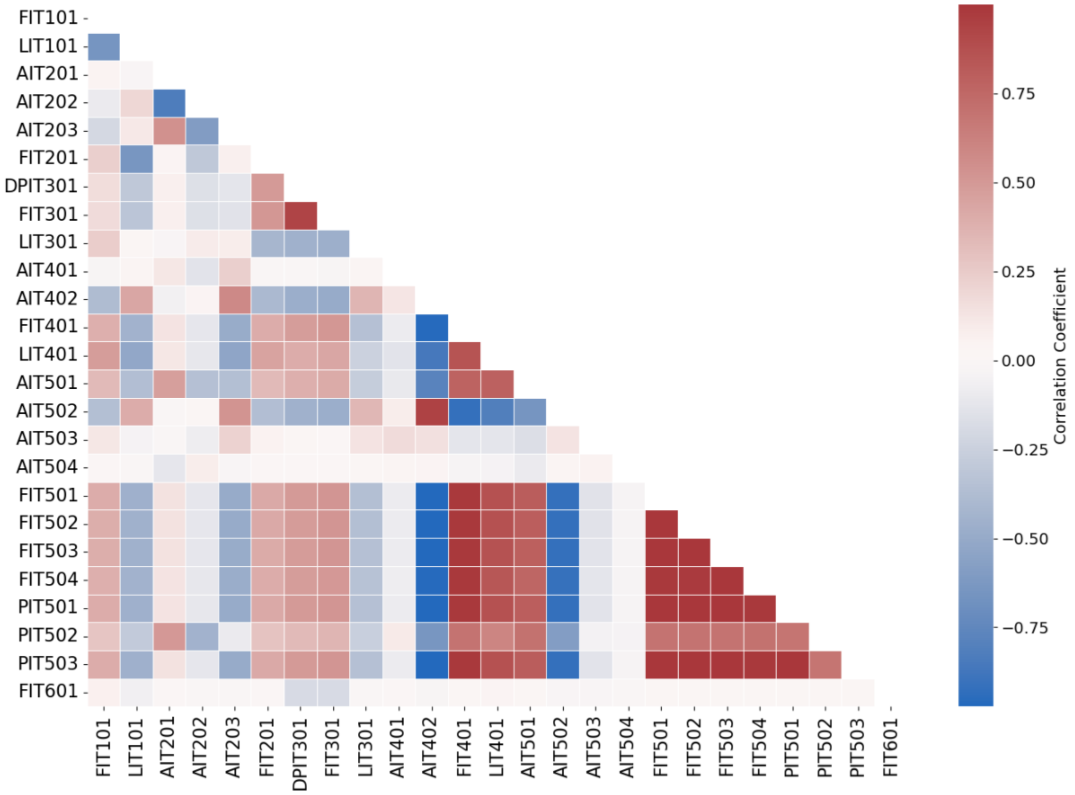

Figure 4 Sensor Feature Correlations

For the purpose of implementing a semantic backdoor, this study intentionally exploits the relationship between two variables, “FIT401” and “FIT501”. Since these targets were arbitrarily chosen, their characteristics can be adjusted accordingly. From the perspective of a clean-label attack, using such correlations as backdoor triggers involve subtly modifying data so that it appears normal, without altering the original labels. Although “FIT401” and “FIT501” exhibit a positive correlation, the backdoor trigger was designed to manipulate the data to display a negative correlation while still being recognized by the model as normal.

Within control loops interconnected by industrial protocols, controllers transmit control commands to actuators, which in turn send responses and controlled data back to the controllers. Distortions in the physical correlations among variables indicate potential sensor failures, equipment malfunctions, or cyberattacks within the process. Similarly, the SWaT testbed operates on a Modbus/TCP-based network, where PLCs deliver control commands to actuators.

In this study, '449,919' data samples collected under various cyber-physical attack scenarios were utilized. Approximately '100,000' of these samples were synthetically modified based on a signal function to induce a negative correlation between "FIT401" and "FIT501." Additionally, "LIT401" was adjusted through inflow and outflow integration to maintain physical consistency. Eighty percent of the synthesized data was used to retrain the HMM-NN model, embedding the semantic backdoor. The remaining 20% was reserved for anomaly and normal state prediction, and the model's performance and statistical distributions were evaluated, as detailed in the following section.

4.4 Evaluation of the Success Rate of Backdoor Attacks

After embedding semantic backdoors into the Hidden Markov-Neural Network-based anomaly detection model, its performance was compared to that of the pre-attack model, as presented in Table II. The performance metrics used were accuracy, recall, precision, and F1-score. The formulas for accuracy, recall, precision, and F1-score are provided in equations (7), (8), (9), and (10), respectively.

Actual Label	Predicted Label	
	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FP)	True Negative(TN)

Table II. Confusion Matrix

$$Accuracy = \frac{TP\_TN}{TP+TN+FP\_FN} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$F1 - Score = 2\frac{Precision*Recall}{Precision+Recall} \tag{10}$$

Anomaly Detection Models (HMM-NN)	Accuracy	Recall	Precision	F1-Score
Not-attacked Model	0.9688	0.9500	0.9792	0.9644
Attacked Model	0.8823	0.8661	0.9217	0.8930

Table III. Performance and Distribution of the Semantic Backdoor Triger Injected HMM-NN Model

Additionally, Attack Success Rate(ASR), Performance Degradation Rate(PDR), and KL Divergence(Kullback–Leibler Divergence) corresponding to the semantic backdoor embedded in the HMM-NN were evaluated and presented in Table III. The ASR denotes the proportion of times the attacker's embedded backdoor successfully deceived the model's prediction. This is expressed as in Equation (11), where 'X' is the entire test data set, 'x_i' is an input data sample, '$M_b(x_i)$' is the backdoor model, 'r' is the intended target class, and 'x_i $contains\ triggers$' indicates whether the input sample contains triggers.

$$ASR = \frac{\sum_{x_i \in x} M_b(x_i)=r}{\sum_{x_i \in \mathrm{X}} x_i\ contains\ triggers} \tag{11}$$

The PDR indicates the extent to which the performance of the model with the embedded backdoor has decreased compared to the existing anomaly detection model. This is expressed as in Equation (12) and is calculated as the ratio of the anomaly detection performance of the backdoor model to that of the clean model.

$$PDR = \frac{Performance_{cleanmodel} - Performance_{backdoormodel}}{Performance_{cleanmodel}} \tag{12}$$

KL Divergence is a metric that allows comparison of the probability distributions for each input across models based on probabilistic predictions. It is expressed as in Equation (13), '$P(\chi)$' denotes the probability distribution of actual data, while '$Q(\chi)$' denotes the probability distribution predicted by the model. The '$log\frac{P(\chi)}{Q(\chi)}$' signifies the amount of information loss incurred when using Q instead of 'P'.

$$D_{KL}(P||Q) = \int(\chi)\ log\frac{P(\chi)}{Q(\chi)} dx \tag{13}$$

ASR	PDR				KL Divergence
	Accuracy	Recall	Precision	F1-Score	
0.6851	0.0893	0.0883	0.0587	0.0740	0.0812

Table IV. Statistical Features of Embedded Backdoors Model

These empirical results demonstrate that, although the anomaly detection model performs normally, it is effectively evaded by data samples containing semantic backdoor triggers. Furthermore, the low KL divergence observed in the model embedded with the semantic backdoor indicates that backdoor detection becomes significantly more challenging.

5 Conclusion

This paper presents a method for embedding semantic backdoors into AI-based anomaly detection systems within ICS. Focusing on triggers at the semantic level, the experiments demonstrate that backdoor attacks in ICS can remain hidden during normal operation and activate only under specific conditions. While AI technology enhances the economic efficiency and operational safety of ICS, it inherently creates an attack surface that adversaries can exploit. Particularly concerning are potential adversarial attacks on AI systems, which, due to the inherent characteristics of AI, are difficult to detect using traditional cybersecurity frameworks.

The adversarial attacks addressed in this study represent only a single scenario; future research will require more realistic prerequisite conditions and thorough validation. Additionally, there is an urgent need to develop countermeasures against such adversarial attacks, including backdoor attacks. Future research should proceed in the following directions. First, more comprehensive datasets and real-world ICS scenarios are necessary to validate and generalize the proposed methodology. Second, defense strategies against attacks must evolve into adaptive and proactive approaches. Finally, interdisciplinary collaboration among AI researchers, ICS engineers, and cybersecurity experts is essential to ensure the safe, resilient, and trustworthy deployment of AI in critical industrial environments.

Acknowledgement

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-25394739, Development of Security Enhancement Technology for Industrial Control Systems Based on S/HBOM Supply Chain Protection, 50%) and was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00242528, 50%).

References

[1] Zhang, Z., Zhao, Y., Liu, T., Li, C., & Wang, X. (2025). On the realism of LiDAR spoofing attacks against autonomous driving vehicle at high speed and long distance. In Proceedings of the Network and Distributed System Security Symposium (NDSS). San Diego, CA, United States.

[2] Zhang, Z., Elsharef, I., & Zeng, Z. (2023, November). Unveiling Neural Network Data Free Backdoor Threats in Industrial Control Systems. In Proceedings of the 2024 Workshop on Re-design Industrial Control Systems with Security (pp. 97-103).

[3] Walita, T., Erba, A., Castellanos, J. H., & Tippenhauer, N. O. (2023, July). Blind concealment from reconstruction-based attack detectors for industrial control systems via backdoor attacks. In Proceedings of the 9th ACM Cyber-Physical System Security Workshop (pp. 36-47).

[4] Tabassi, E. (2023, January). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

[5] Davaslioglu, K., & Sagduyu, Y. E. (2019). Trojan attacks on wireless signal classification with adversarial machine learning. arXiv. https://doi.org/10.48550/arXiv.1910.10766

[6] Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733. https://arxiv.org/abs/1708.06733

[7] Turner, A., Tsipras, D., & Madry, A. (2018, September). Clean-label backdoor attacks. OpenReview. Retrieved September 19, 2025, from https://openreview.net/forum?id=HJg6e2CcK7

[8] Saha, A., Subramanya, A., & Pirsiavash, H. (2020). Hidden trigger backdoor attacks. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 11957–11965. https://doi.org/10.1609/aaai.v34i07.6871

[9] Dong, L., Qiu, J., Fu, Z., Chen, L., Cui, X., & Shen, Z. (2023). Stealthy dynamic backdoor attack against neural networks for image classification. Applied Soft Computing, 149, 110993. https://doi.org/10.1016/j.asoc.2023.110993

[10] Li, Q., Chen, W., Xu, X., Zhang, Y., & Wu, L. (2025). Precision strike: Precise backdoor attack with dynamic trigger. Computers & Security, 148, 104101. http://doi.org/10.1016/j.cose.2024.104101

[11] Mengara, O. (2024). The art of deception: Robust backdoor attack using dynamic stacking of triggers. arXiv. https://doi.org/10.48550/arXiv.2401.01537

[12] Sun, W., et al. (2024). Invisible backdoor attack with dynamic triggers against person re-identification. IEEE Transactions on Information Forensics and Security, 19, 307–319. https://doi.org/10.1109/TIFS.2023.3322659

[13] Quiring, E., & Rieck, K. (2020). Backdooring and poisoning neural networks with image-scaling attacks. In 2020 IEEE Security and Privacy Workshops (SPW) (pp. 41–47). IEEE. https://doi.org/10.1109/SPW50608.2020.00024

[14] GhostEncoder: Stealthy backdoor attacks with dynamic triggers to pre-trained encoders in self-supervised learning. (2024). Computers & Security, 142, 103855. https://doi.org/10.1016/j.cose.2024.103855

[15] Wang, B., Yu, F., Wei, F., Li, Y., & Wang, W. (2024). Invisible intruders: Label-consistent backdoor attack using re-parameterized noise trigger. IEEE Transactions on Multimedia, 26, 10766–10778. https://doi.org/10.1109/TMM.2024.3412388

[16] Chen, H., et al. (2024). Investigating the backdoor on DNNs based on recolorization and reconstruction: From a multi-channel perspective. IEEE Transactions on Information Forensics and Security, 19, 6923–6934. https://doi.org/10.1109/TIFS.2024.3427432

[17] Zou, M., Shi, Y., Wang, C., Li, F., Song, W., & Wang, Y. (2019). PoTrojan: Powerful neural-level trojan designs in deep learning models. arXiv. https://doi.org/10.48550/arXiv.1802.03043

[18] Jin, L., Jiang, W., Zhan, J., & Wen, X. (2024). Highly evasive targeted bit-trojan on deep neural networks. IEEE Transactions on Computers, 73(9), 2350–2363. https://doi.org/10.1109/TC.2024.3416705

[19] Rakin, A. S., He, Z., & Fan, D. (2020). TBT: Targeted neural network attack with bit trojan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13198–13207). IEEE. Retrieved September 14, 2025, from https://openaccess.thecvf.com/content_CVPR_2020/html/Rakin_TBT_Targeted_Neural_Network_Attack_With_Bit_Trojan_CVPR_2020_paper.html

[20] Zhu, M., Li, Y., Guo, J., Wei, T., Xia, S.-T., & Qin, Z. (2025). Towards sample-specific backdoor attack with clean labels via attribute trigger. IEEE Transactions on Dependable and Secure Computing, 22(5), 4685–4698. https://doi.org/10.1109/TDSC.2025.3552234

[21] J. Goh, S. Adepu, K. Junejo, and A. Mathur, A Dataset to Support Research in the Design of Secure Water Treatment Systems. 2016.