Research on Implementation eXplainable Anomaly Detection Using Artificial Intelligence on Cyber-Physical System

Ka-Kyung Kim[a], Ieck-Chae Euom [a*]
[a] *Chonnam National University System Security Research Center, 77 Yongbong-ro 138beon-gil, Gwangju, 61186*
[*]*Corresponding author: iceuom@jnu.ac.kr*

Keywords : Anomalies Detection, Explainable Artificial Intelligence, Cyber-Physical System Noise, Industrial Control System

1. Introduction

Advanced systems in nuclear power plants enable flexible and intelligent operations powered by artificial intelligence, but also present an expanded cyber threat surface due to more access points. In the context of cybersecurity, anomaly detection systems, in particular, the algorithms themselves may become targets for attackers. Attacks such as targeted manipulation of the anomaly detection algorithm itself may not be recognized as an intrusion by a hostile actor, so anomaly detection classification results must be configured to be explainable.

However, machine learning algorithms, such as deep learning for anomaly detection, have several limitations. Firstly, these are often black-box models that cannot be explained, except for lightweight algorithms. Secondly, cyber-physical systems are subject to nonlinearities and uncertainties due to the nature of interactions and sensors[1][2], which most AI algorithms fail to account for. Thirdly, network-level communication packets and log data for anomaly detection are not effective when an intelligent attacker bypasses the perimeter protection system or conceals traces of intrusion.

Based on these points, this research proposes considering the noise of cyber-physical systems at the operational data level and constructing an explanatory pipeline for anomalies identified by algorithms. It is also suggested that this should operate within tightly controlled tolerances based on safety design criteria, and that the pipeline be implemented from a secondary perspective that does not interfere with monitoring the original source data.

2. Research Background

This chapter discusses the filters applied to account for noise in cyber-physical systems and analyzes related research focused on explanatory anomaly detection systems for industrial control system environments.

2.1 Noise Filtering of Cyber-Physical System

The interaction between the components of a typical industrial control system – HMI(Human-Machine Interface), PLC(Programmable Logic Controller), actuators, and sensors - is shown in Fig. 1[3]. The user monitors process status data transmitted from the PLC via the HMI and adjusts setpoints. To reach or maintain the setpoint entered by the user, the PLC issues control commands to the actuator, which performs the appropriate action. The output of this process is sensor measurements, which are fed back to the PLC in an iterative control loop.

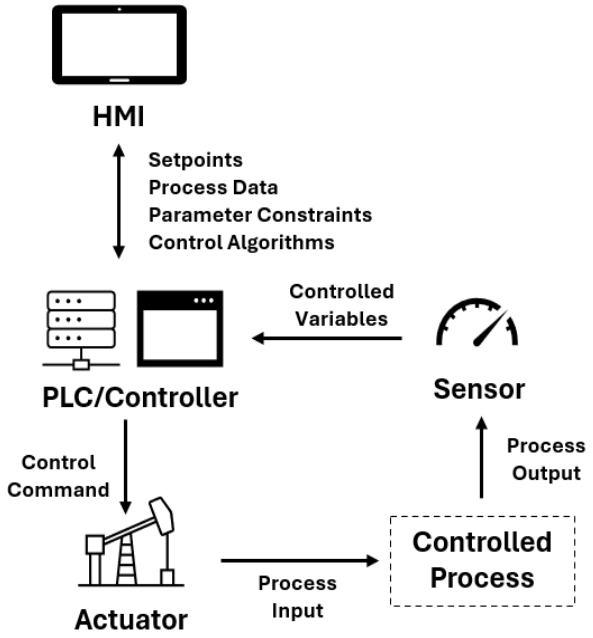

Fig. 1. Loop of Instrument and Control System

In such a control loop, unintended noise is introduced during the interaction between the systems. This means that despite the PLC's control commands being based on the user-entered setpoints, it is not possible to reach perfectly accurate values due to realistic noise. This nonlinearity is caused by the realistic nature of cyber-physical systems.

There is also uncertainty in the sensors that collect physical and environmental information and relay it to the PLC. Sensors are not the target system to measure, and are easily affected by seasonal, temporal, and condition conditions.

This is defined as measurement error and process error, and is the result of mathematically modeling the real world. The Kalman filter (KF) is developed by Rudolf Kalman in 1930 to estimate the state of a linear dynamical system based on measurements that contain such noise. It is still actively used today in a variety of fields, including robotics, control engineering, computer vision, radar, and signal processing.

Kalman Filters operate on the basis of discrete-time linear dynamical systems and assume a Markov chain, where the state vector at each time is determined by the state vector at the previous time[1]. It works in a recursive fashion to predict the joint distribution of the current state variables based on past measurements and estimates, iteratively performing two steps: prediction and update.

However, since the Kalman Filter is designed for linear systems, it is difficult to apply to systems with nonlinear structures. To solve this problem, the Extended Kalman filter (EKF) is designed. The EKF obtains a linearized value by partially differentiating each moment of a nonlinear function based on a Jacobian matrix that represents the relationship between key parameters[4].

Another type of Kalman Filter for nonlinear dynamic systems exists: the Unscented Kalman filter (UKF). Unlike the EKF, which linearizes the nonlinear function by partial differentiation, the UKF aims to find the nonlinear function itself. The UKF assumes a Gaussian distribution and uses a few sigma point analysis to estimate the mean and variance of the nonlinear function. [5].

Table I: Summary of KF type

Type	State and observation functions	Mathematical estimation methods
KF	Linear	Linear Function Model
EKF	Nonlinear	Jacobian Matrix
UKF	Nonlinear	Sigma Point

2.2 Explainable AI for Anomaly Detection

Explainable AI is the process of explaining the accuracy, objectivity, and transparency of a model so that users can trust AI-driven decisions. As AI technology advances, there is a need to analyze algorithms backwards from its output. If AI-based systems can show how to reach a particular outcome, it is possible to make adjustments to improve performance or prevent unintended consequences. The need for explainable AI will be especially important in detecting cyber anomalies at nuclear power plants that could threaten national safety and security.

The technologies currently being utilized to implement explainable AI for black box algorithms are shown in Table I. Techniques for explaining black-box algorithms are generally categorized into 'Model-Agnostic' methods, which are applicable to all algorithms based on 'Post-hoc' tracking of results after an event, and 'Model-Specific' methods, which are applicable only to specific algorithms.

Table II: Comparison of AI Explanation Technic

Explanation Technic	Explanation Scope	Model-Specific
SHAP	Global	X
LIME	Local	X
Anchors	Local	X
PDP	Global	X
ICE	Global	X
InTrees	Global	O

- SHAP (Shapley Additive exPlanation): Game theory-based feature contribution analysis[6][8].
- LIME (Local Interpretable Model-agnostic Explanation): Derive local interpretability near specific predictors[6][8].
- Anchors: Simplify by creating rules for prediction conditions that remain constant[7].
- PDP (Partial Dependence Plot): visualizes the relationship between specific features and predictions[6][8].
- ICE (Individual Conditional Expectation): visualizes the prediction response of individual instances[8].
- InTrees: Extract interpretable rules from an ensemble of trees[9].

In addition to the above techniques, there is a significant amount of research on explainable AI algorithms. Chathurika S. Wickramasinghe [10] proposed an implementation of an explainable unsupervised learning algorithm for cyber-physical systems based on 'Explainable Self-Organizing Maps'. Bhawana Sharma [11] et al. proposed the application of LIME and SHAP in deep learning neural networks for intrusion detection in IoT networks. Latifah Almuqren [12] applied the LIME technique to Hybrid Enhanced Glowworm Swarm Optimization (HEGSO) and Improved Elman Neural Network (IENN) models to achieve explainable AI-based intrusion detection in cyber-physical systems. J.LI [13] proposed a framework that integrates process, network traffic, and PLC data to enhance cybersecurity of nuclear power plants based on explainable artificial intelligence.

3. Explainable Artificial Intelligence Anomaly Detection System Algorithm

In this paper, a methodology proposes a pipeline to effectively consider noise from cyber-physical systems and provide an explanation for the identified anomalies.

The proposed methodology is illustrated in the flowchart shown in Fig. 3.

Compared to the existing works, including those discussed in Chapter 2, the originality of this work is that it performs recursive filtering to account for noise in cyber-physical systems. The recursive filtering uses the EKF, and applies LIME and SHAP techniques to explain the results.

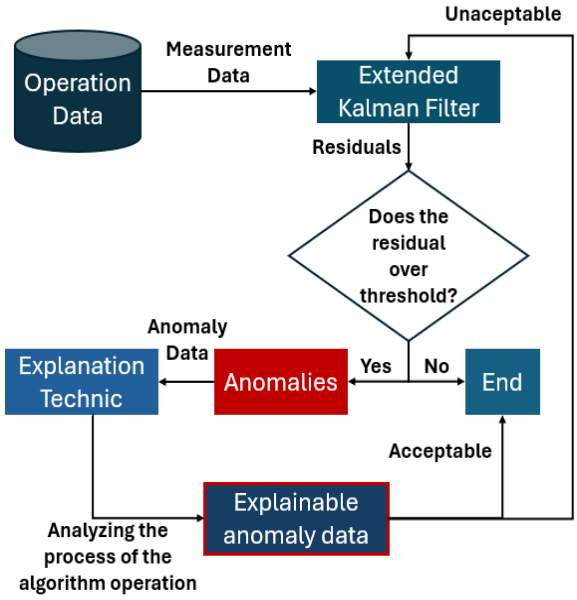

Fig. 3. Flowchart of the proposed methodology for implementing an explainable anomaly detection system

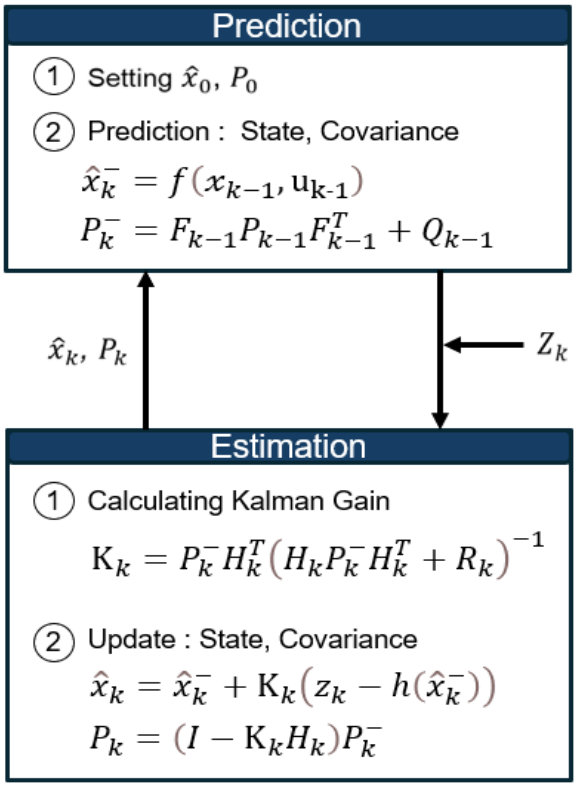

Fig. 4. EKF Operation Process

In this paper, the EKF is utilized assuming that most cyber-physical systems exhibit a nonlinear dynamic structure. It is also assumed that anomaly detection monitors and personnel possess the capability to comprehend the nonlinear functions of the system. EKF also has the same behavioral structure as other Kalman filter types, which is an iterative process of prediction and estimation. The detailed steps are shown in Fig. 4, and the symbols used are summarized in Table III.

Table III: Symbol Description of EKF

Symbol	Description
x	State
z	Observation
P	Covariance Matrix
f	Nonlinear State Transition Functions
h	Measurement Function
u	Control Input
Q	Covariance Matrix of system noise
R	Covariance Matrix of Measurement Noise
F	Jacobian Matrix (partial derivative) of State Transition Function 'f'
H	Jacobian Matrix of the Measurement Function 'h'
K	Kalman Gain
k	'k' Time Point
k-1	'k-1' Time Point
T	Transpose of a Matrix
$^-$	Predicted Value
$\hat{}$	Estimate
I	Identity Matrix

The difference between the estimated value and the observed value by the EKF is defined as the residual, which is identified as an anomaly when it exceeds a certain threshold level. The threshold is a user-adjustable hyperparameter value.

Explainable artificial intelligence techniques are then applied to interpret why the algorithm identified the data as an anomaly. If the resulting decision-making process of the described algorithm is not acceptable, it might be necessary to adjust hyperparameters to improve performance or explore other approaches.

4. Case Study

For the case study of the approach discussed in this paper, the Security Dataset 'HIL-based Augmented ICS (HAI) v.23.05' collected from the Hardware In The Loop (HIL) simulator-based industrial control system

testbed is utilized. HAI is an integrated simulator that includes GE's turbine testbed, EMERSON's boiler testbed, and FESTO's modular production water treatment system testbed[14].

HAI 23.05 consists of data collected during 249 hours of normal operation and 79 hours of abnormal operation, with 52 intentional attacks performed. We focused our analysis on 'P1_LIT01', which measures the water level in the return tank of the boiler process, and utilized 'P1_LCV01D', 'P1_FT01', 'P1_FT03', and 'P1_PIT01' as variables in the EKF's nonlinear function and matrix settings. The variables used and a description of 'P1_LIT01' are shown in Table IV.

Table IV : Description of variables used in EKF

Variable Name (Data)	Variable Description
P1_LCV01D (0~100%)	Position Command for the LCV01 Valve
P1_FT01 (0~2,500mmH2O)	Measured Flowrate of the Return Water Tank
P1_FT03 (0~2,500mmH2O)	Measured Flowrate of Return Water Tank
P1_PIT01 (0~10bar)	Heat-exchanger Outlet Pressure

The valve control command 'P1_LCV01D' regulates the outflow and inflow flow rate of the return water tank, impacting the value of 'P1_LIT01'. 'P1_LIT01' is proportional to 'P1_FT01', representing the return water tank inlet flow rate, and inversely proportional to 'P1_FT03', representing the return water tank outlet flow rate. It is also directly proportional to 'P1_PIT01', representing the Pressure of the water tank. Considering these relationships, the Jacobian matrix of EKF is established accordingly.

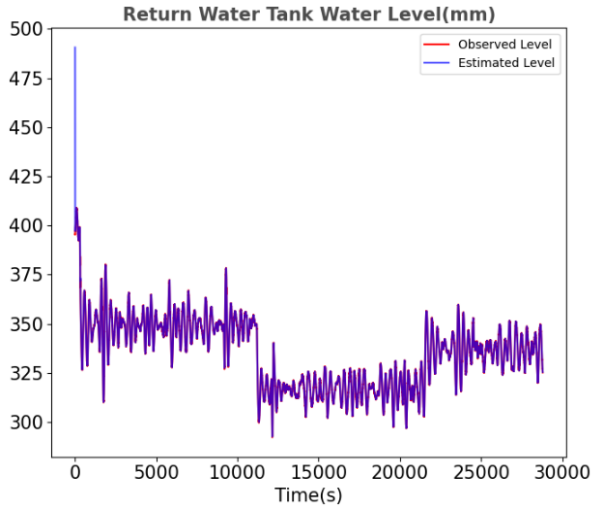

Fig. 5. EKF Estimate and Observation data of P1_LIT01

To intuitively understand the results of the application, the data of '28,799' cases from the earliest date of the intentional attack scenario in the 'HAI' dataset (August 12, 2022) is extracted. The graph of EKF estimates and observations on this data is shown in Fig. 5.

The performance comparison is shown in Table V, where the range of anomaly classification based on the residual between the EKF estimate and the observation is set to '1-5mm'. There are a total of 8 scenarios of attacks performed on August 12, 2022. Since these eight attack scenarios have different attack durations ranging from a minimum of 96 seconds to a maximum of 237 seconds, only detection (O, X) was analyzed by grouping the duration per attack into one. When the residual threshold was between 1 and 2, all grouped intentional attacks were detected. However, the number of false positives dropped dramatically to 1,221 when the residual threshold was '1' and 116 when it was '2'. Therefore, the threshold for anomaly identification in this case study was set to '2 (mm)'.

Table V : Comparison of whether an attack is detected by scaling thresholds

	Residuas Thresholds(mm)				
Attack Starting Point (timestamp)	1	2	3	4	5
1,505s	O	O	O	O	X
5,702s	O	O	X	X	X
9,138s	O	O	O	O	O
12,065s	O	O	O	X	X
16,092s	O	O	X	X	X
20,162s	O	O	O	O	X
24,422s	O	O	X	X	X
27,316s	O	O	X	X	X

To the anomalous data identified by the residual threshold of '2(mm)', the explainable models 'LIME' and 'SHAP' were applied. Fig. 6 through 8. show the results of applying LIME, and Fig. 9 shows the results of applying SHAP.

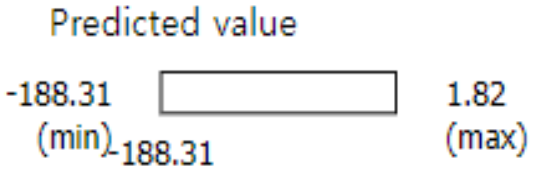

Fig. 6. 'Predicted value(min, max)' graph based on 'LIME' technique

The 'Predicted value' in LIME shows the minimum and maximum values of the residuals predicted by the model at a particular data point. The closer to these minimum and maximum values, the greater the probability of predicting the data as anomalous. In other words, the more the residuals extracted by EKF are skewed towards '-188.31(mm)' and '1.82(mm)', as shown in Fig. 6, the greater the probability of classifying the data as anomalous.

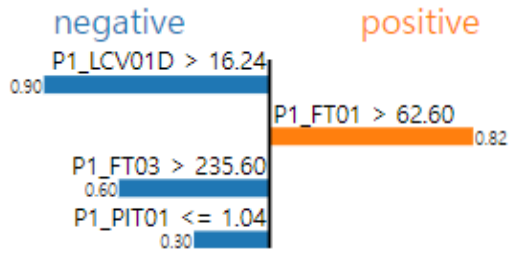

Fig. 7. Feature contribution analysis graph of residuals and predictions based on 'LIME' technique

The 'Positive' and 'Negative' graphs of LIME shown in Fig. 7 represent the contribution to the predicted value. Interpreting the above figure, it means that when the 'P1_FT01' bar data in the Positive graph is greater than '62.60(mmH2O)', the residuals act in the direction of increasing. On the contrary, it means that 'P1_LCV01D' is less than '16.24(%)', 'P1_FT03' is less than '235.60(mmH20)', and 'P1_PIT01' is less than or equal to '1.04(bar)', which contributes to decreasing the residual.

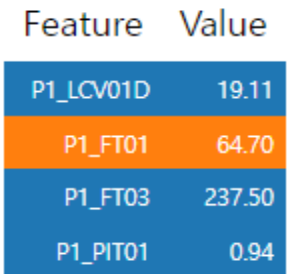

Fig. 8. Visualization of the impact of features using the 'LIME' technique

Fig. 8 is an example of the “LIME” technique, which indicates that the actual value impacts the results positively or negatively. This complements the interpretation of Fig. 7. Fig. 9 presents the 'SHAP value', representing each feature's contribution to the predicted value across the dataset. Features with more red plots have a higher contribution, while those with more blue plots have a lower contribution. Analyzing the above figure, 'P1_FT01' enhances the model's predictive value and is a crucial feature. 'P1_FT03' has a 'SHAP value' mainly distributed in the negative direction, indicating that larger values lead to smaller residuals. For 'P1_LCV01D', the 'SHAP value' plots in red show contributions in both positive and negative directions, suggesting that residuals can increase or decrease depending on the context. The 'SHAP value' for 'P1_PIT01' reveals predominantly negative blue plots, signifying a relatively low impact of variations in that data.

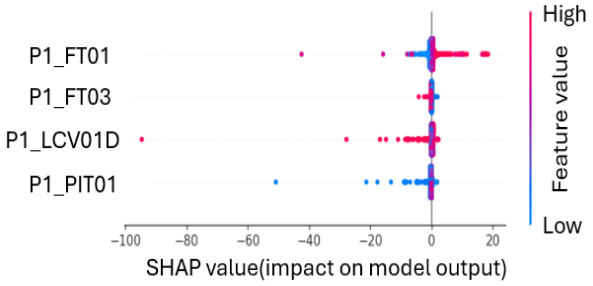

Fig. 9. Analyzing Feature value and ‘SHAP’ value of ‘SHAP’ Technic

Fig. 9 visualizes the 'SHAP value', which means the contribution of each feature to the predicted value over the entire data. Features with more plots in 'Red' color have higher contribution, and features with more plots in 'Blue' color have lower contribution. Interpreting the above figure, according to the 'SHAP value', it is possible to see that 'P1_FT01' increases the predictive value of the model and is an important feature. The 'SHAP value' of 'P1_FT03' is mainly distributed in the negative direction, so the larger the value of this data, the smaller the residual. For 'P1_LCV01D', the 'SHAP value' of 'P1_LCV01D', the plots in red are in both positive and negative directions. This means that the residuals can be increased or decreased depending on the situation. The 'SHAP value' for 'P1_PIT01' shows that the blue plot is predominantly in the negative direction, indicating that the variation in that data has a relatively low impact.

5. Conclusions

In this paper, an approach for implementing an algorithm for detecting explainable anomalies that considers noise in cyber-physical systems is proposed. In the process of mathematically modeling the realistic space, the noise in cyber-physical systems including instrumentation and control systems in nuclear power plants, to mathematically model the realistic space Using the Extended Kalman filter (EKF), a recursive algorithm that is capable of accommodating such noise and applicable to nonlinear dynamic systems, the approach utilized. These pipelines should also be built from a secondary perspective so that comparison is possible in parallel with monitoring of the original source data where no operations have been performed.

As with all recursive algorithms for nonlinear dynamic systems, such as EKF, the use of more than one variable requires building a model that can explain the anomaly detection results. To address this

requirement, the case study analyzed the impact on anomaly detection results using LIME and SHAP techniques, which are popularly used in public security datasets of industrial control systems. Using it as a supplementary explanatory model will establish a safer and more dependable environment for AI systems.

In this regard, the methodology proposed in this paper can be expected to be highly practical and effective compared to existing related studies. In future work, plans include developing a pipeline to simultaneously analyze industrial control system process data and network-level data for safety-designed test environments and building models to explain the results.

ACKNOWLEDGMENT

This work was supported by Institute for Information and Communication Technology Planning and Evaluation(IITP-RS-2022-II221203, 50%) and the National Research Foundation of Korea (NRF) grant funded by the government (Ministry of Science and ICT) (No. 2022R1G1A1010506, 50%).

REFERENCES

[1] MathWorks, Video and Webinar Series, https://kr.mathworks.com/videos/series/understanding-kalman filters.html
[2] Chanyoung Lee, Jae Gu Song, Cheol Kwon Lee, Poong Hyun Seong, “Development of a method for securing the operator's situation awareness from manipulation attacks on NPP process data”, Nuclear Engineering and Technology, Volume 54, Issue 6, Pages 2011-2022, 2022.\
[3] National Institute of Standards and Technology, “Guide to Operational Technology Security”, 2023.
[4] J. Guo, L. Li, J. Wang and K. Li, "Cyber-Physical System-Based Path Tracking Control of Autonomous Vehicles Under Cyber-Attacks," in IEEE Transactions on Industrial Informatics, vol. 19, no. 5, pp. 6624-6635, 2023.
[5] M. M. Rana, M. K. R. Khan and A. Abdelhadi, "IoT Architecture for Cyber-Physical System State Estimation Using Unscented Kalman Filter," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, pp. 910-913, 2020.
[6] Ziqi Li, “Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost”, Computers, Environment and Urban Systems, Volume 96, 2022.
[7] K. Jayakumar and N. Skandhakumar, "A Visually Interpretable Forensic Deepfake Detection Tool Using Anchors," 2022 7th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, pp. 1-6, 2022.
[8] Marta Klosok, Marcin Chilebus, “Toward Better Understanding of Complex Machine Learning Models Using Explainable Articial Intelligence(XAI)-Case of Credit Scoring Modelling”, Unversity of Warsaw, 2020.
[9] H. Deng, “Interpreting tree ensembles with intrees”, Intternational Journal of Data Science and Analytics, 2014.
[10] C. S. Wickramasinghe, K. Amarasinghe, D. L. Marino, C. Rieger and M. Manic, "Explainable Unsupervised Machine Learning for Cyber-Physical Systems," in IEEE Access, vol. 9, pp. 131824-131843, 2021.
[11] Bhawana Sharma, Lokesh Sharma, Chhagan Lal, Satyabrata Roy, “Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach”, Expert Systems with Applications, Volume 238, Part A, 2024.
[12] Almuqren L, Maashi MS, Alamgeer M, Mohsen H, Hamza MA, Abdelmageed AA. Explainable Artificial Intelligence Enabled Intrusion Detection Technique for Secure Cyber-Physical Systems. Applied Sciences. 13(5):3081. 2023.
[13] J.LI, M.U.OZBEK, “Enhancing Computer Security of Nuclear Power Plants Based on Explainable AI”, IAEA-CN-321/357, 2024.
[14] National Security Research Institute, "HAI Security Dataset Technical Details," Technical Report, Version 4.0, 2023.