electronics

MDPI

Article

Anomaly Detection and Analysis in Nuclear Power Plants

Abhishek Chaudhary [1], Junseo Han [1], Seongah Kim [1], Aram Kim [2] and Sunoh Choi [1,*]

[1] Department of Software Engineering, Jeonbuk National University, Jeonju-si 54896, Republic of Korea; abhishek.chaudhary@jbnu.ac.kr (A.C.); hjs123@jbnu.ac.kr (J.H.); ssah2026@jbnu.ac.kr (S.K.)
[2] Department of Information Security, Suwon University, Hwaseong-si 18328, Republic of Korea; aramkim@suwon.ac.kr
* Correspondence: suno7@jbnu.ac.kr

Abstract: Industries are increasingly adopting digital systems to improve control and accessibility by providing real-time monitoring and early alerts for potential issues. While digital transformation fuels exponential growth, it exposes these industries to cyberattacks. For critical sectors such as nuclear power plants, a cyberattack not only risks damaging the facility but also endangers human lives. In today's digital world, enormous amounts of data are generated, and the analysis of these data can help ensure effectiveness, including security. In this study, we analyzed the data using a deep learning model for early detection of abnormal behavior. We first examined the Asherah Nuclear Power Plant simulator by initiating three different cyberattacks, each targeting a different system, thereby collecting and analyzing data from the simulator. Second, a Bi-LSTM model was used to detect anomalies in the simulator, which detected it before the plant's protection system was activated in response to a threat. Finally, we applied explainable AI (XAI) to acquire insight into how distinctive features contribute to the detection of anomalies. XAI provides valuable explanations of model behavior by revealing how specific features influence anomaly detection during attacks. This research proposes an effective anomaly detection technique and interpretability to better understand counter-cyber threats in critical industries, such as nuclear plants.

Keywords: digital system; security; cyberattack; nuclear power plant; deep learning; anomaly detection

Citation: Chaudhary, A.; Han, J.; Kim, S.; Kim, A.; Choi, S. Anomaly Detection and Analysis in Nuclear Power Plants. *Electronics* **2024**, *13*, 4428. https://doi.org/10.3390/electronics13224428

Academic Editors: Aryya Gangopadhyay and Franco Cicirelli

Received: 1 October 2024
Revised: 30 October 2024
Accepted: 7 November 2024
Published: 12 November 2024

1. Introduction

Rapid digital transformation across industries has revolutionized the management, monitoring, and control of operations. From manufacturing to energy production, industries are increasingly adopting digital systems for improved efficiency, real-time monitoring, and control operations. This shift toward digital systems has led to unprecedented growth, automation, and accessibility. However, this digital evolution presents a new set of risks, particularly in critical sectors, such as nuclear power plants (NPPs). Cybersecurity has become a pressing concern because malicious attacks on these digital systems can have catastrophic consequences, potentially damaging facilities and posing significant threats to human life, safety, and the environment. Given the sensitive nature of NPPs, even minor malfunctions or cyber intrusions can have devastating long-term effects. Therefore, robust security measures are essential. Such vulnerabilities pose a significant risk to NPPs because cyberattacks can compromise core systems, leading to catastrophic consequences. Although existing safety mechanisms, such as the reactor protection system (RPS), are designed to safeguard against threats, they may not be sufficient to counter sophisticated cyberattacks. The early detection of anomalies caused by such attacks is crucial for protecting against system disruptions. Meanwhile, RPSs are designed to be activated when specific conditions are matched, which may require a certain amount of time for the triggering conditions to be met. This creates a delay in the response, reducing overall efficiency and compromising safety. In NPPs, where every millisecond is crucial for instantaneous reactions, the

inefficiency of RPS activation time poses a serious challenge, driving the need for a more advanced and faster response system for safety protocols.

Several testbeds have been developed for the research and development of security measures. This study began by utilizing the Asherah NPP simulator [1], which is a virtual testbed designed to simulate realistic NPP operations. The initial objective of this research was to collect and analyze realistic data related to the normal and abnormal behavior of NPPs. The second objective was to develop a robust anomaly detection model capable of identifying potential cyberattacks against NPPs. Trained on only normal data using a deep learning model, decisions made by such models are considered a black box, meaning that humans do not easily understand the decision making process. Explainable artificial intelligence (XAI) [2] was implemented to enhance the interpretability of the model by providing insights into the contribution of specific features to anomaly detection. Since features are related to systems within the plant, insights by XAI can assist human operators with an understanding of which system is causing the issue. Such an understanding of the specific features allows for targeted troubleshooting and quicker response.

The Asherah Nuclear Power Plant simulator (ANS) was applied for the first time as a training tool in the Brazilian Cyber Guardian Exercise [3], where different attack scenarios were discussed and simulated. In this research, three different cyberattacks were injected into the system using the Asherah NPP simulator, each targeting a different aspect of the plant's operations. These attacks were designed to mimic real-world scenarios in which a malicious actor might gain control over critical systems such as the condenser cooling pump, Feed Water Pump, and the pressurizer spray valve, which regulate essential processes within the plant.

After collecting data from these simulated attacks, the bidirectional long short-term memory (Bi-LSTM) model [4] was implemented and trained on normal data to detect anomalies in abnormal data. The Bi-LSTM model, a time-series deep learning model, is well suited for handling sequential data and capturing patterns over time, making it ideal for monitoring the continuous flow of data in NPPs. However, deep learning models are often criticized for their opacity. Therefore, XAI was applied to uncover the prediction method of the Bi-LSTM model. Specifically, XAI's SHapley Additive exPlanations (SHAP) [5] was used to analyze the features (such as pressure, temperature, and flow rate) that are most influential in triggering the anomaly detection mechanism. This two-pronged approach allowed us to not only detect anomalies but also explain the reasoning behind the model's decisions, offering valuable insights into the vulnerabilities of the system.

The Bi-LSTM model successfully detected anomalies during the simulated attack scenarios and identified irregular patterns before the RPS could be triggered. This early detection demonstrated the capability of the model to alert system operators to potential threats before escalating to more serious system failures. Moreover, the application of XAI provided clear explanations as to why certain features, such as rising pressure levels or sudden drops in water flow, were flagged as contributors to this anomaly. This interpretability allows researchers to understand how the deep learning model analyzes and processes the data, making it easier to trust and validate the model predictions.

In all three attack scenarios, the model flagged anomalous behavior well before the reactor safety systems were activated, highlighting its effectiveness in early intervention. The integration of XAI further enhanced the transparency of the model by offering clear explanations of how and why the anomalies were detected. This interpretability is crucial in high-stakes environments such as NPPs, where understanding the rationale behind model decisions can lead to more informed interventions.

In this study, the normal and abnormal states of the NPP were analyzed to acquire insights that would help differentiate between these states. The contribution of this study is the effective use of innovative methods to boost deep learning model interpretability and anomaly detection in NPPs. The Asherah NPP simulator was used to comprehensively evaluate the patterns of both normal and abnormal activities, using three different anomaly attack scenarios that targeted separate systems. The Bi-LSTM deep learning model in

this study successfully identified irregularities before the RPS was triggered by detecting the anomalies generated by the simulated cyberattacks. Additionally, the use of XAI provided deeper insights into the decision making process of the model, demonstrating the roles that specific features play in anomaly identification. The increased interpretability of these models contributes to increased transparency and confidence when applied to critical infrastructures.

Overall, this research proposes an effective anomaly detection technique for critical industries, combining the predictive power of deep learning with the clarity of XAI, thus offering a comprehensive solution for mitigating cyber threats in industrial settings.

The key contributions of this research are as follows:

1. **Comprehensive anomaly analysis**: we analyzed both normal and anomalous behaviors in the Asherah NPP simulator by injecting three distinct types of anomalies, each targeting a distinct system.
2. **Efficient anomaly detection**: a Bi-LSTM deep learning model was applied to successfully detect these anomalies and identify irregularities before the RPS was activated.
3. **Improved model transparency with XAI**: by incorporating XAI, we provide insights into feature contributions, thus enhancing the interpretability of the model's anomaly detection process.

2. Related Work

2.1. Industry Control System (ICS) Testbeds

Massive industries that were once manual and labor-intensive have grown rapidly through more efficient automated operations that seamlessly interact with complex processes with almost no human intervention. Such an evolution was made possible by the substantial contribution of the ICS, which was designed to monitor, control, and automate intricate physical operations. The ICS brings growth and development along with risks to the security and integrity of the NPP, where precision is not just a goal but a necessity [6]. The data collected in such plants are not just for maintaining the health of the system but can also play a key role in identifying potential malfunctions and enhancing security measures. However, because of the sensitive nature of such data, they remain inaccessible to the public. This has created a challenge for researchers seeking to understand and develop a protection system for NPPs.

As a solution, some researchers have built testbeds in physical, virtual, and hybrid forms that are realistic replicas of industrial systems such as power plants, allowing them to collect realistic data without risking real-world operations [7]. Physical testbeds are built using hardware and software components that replicate real-world industrial systems. They provide authentic data but come with excessive costs and significant risks. The hardware-in-the-loop (HIL)-based augmented ICS (HAI) [8] and Turbo-Gas power plant (T-GPP) [9] are examples of physical testbeds. However, virtual testbeds provide a better workaround, offering the same type of simulation as physical testbeds in fully virtual environments. They are safer, affordable, and easier to maintain with no physical labor. The real-time immersive network simulation environment (RINSE) for network security exercise [10] and the testbed for analyzing the security of the supervisory control and data acquisition (SCADA) control system (TASSCS) [11] are examples of virtual testbeds. Hybrid testbeds are a combination of physical and virtual testbeds, using real components where necessary and supporting the remainder with virtual models. The use of real components enables data integrity and realistic testing. The ICS interaction testbed [12] was presented as a real-time hybrid testbed environment for network-intrusion detection systems (NIDs).

In this study, we applied a virtual testbed called the Asherah NPP (Section 3), which was designed using a MATLAB simulator R2018a. Simulations in a virtual environment are flexible and risk-free, and they closely mimic the real working NPP, providing realistic data for maintaining the security and integrity of sensitive information.

2.2. Anomaly Detection

Anomaly detection is the identification or observation of data points that deviate from normal, exhibiting rare or unexpected inconsistent behavior. With the help of AI and machine learning (ML), anomaly detection has become an automated process that can effectively identify outliers that exist in a larger dataset [13]. Unlike classification, the task of which is to organize data precisely into predefined categories after learning from a large pool of examples, anomaly detection is unsupervised, often without any labels. It focuses on the search and detection of rare deviations.

Anomaly detection is an essential factor and has diverse research areas and application fields. One survey [14] provides a structural and comprehensive overview of the existing techniques and approaches for anomaly detection. In recent years, deep learning models have attracted the attention of researchers due to their effective performance in outperforming traditional models. A study was conducted for deep learning models for anomaly detection [15], highlighting key intuitions, objective functions, underlying presumptions, benefits, drawbacks, and challenges.

Similarly, a survey was conducted on Weakly Supervised Anomaly Detection (WSAD) [16] that works in an incomplete, inexact, and inaccurate supervision environment. The data modalities used were tabular, graph, time-series, and image/video. Moreover, ADGym [17] provides a platform for an automatic optimal design selection after a comprehensive dataset evaluation rather than solely relying on pre-existing methods.

On the other hand, the Deep Neural Network (DNN)-based model was also proposed [18] for anomaly detection in IoT architecture with competitive performance and comparison with CNN, RNN, and LSTM. An autoencoder (AE)-based neural network architecture was applied to anomaly detection [19]. This model demonstrated the capability to identify attacks, including small variations, outperforming the one-class support vector machine-based solution. However, the model did not demonstrate high-precision learning and was inadequate for detecting advanced attacks. As a solution, a recurrent neural network (RNN) was used for more precise estimation [20], outperforming traditional methods. Cyber–physical system (CPS)-GUARD [21] is an outlier-aware deep autoencoder semi-supervised learning technique for flawed training data and intrusion detection in cyber–physical systems. The dynamic graph modeling [22] approach, proposed based on prior knowledge, is a variational mode decomposition module for filtering out data noise and feature extraction, using a spatiotemporal graph neural network (ST-GNN). The performance of ST-GNN is low because of data noise and requires noise reduction. Thus, it may not be applicable to NPPs, where real-time detection is critical, and the data may contain noise.

Nikhil Laxminarayana et al. propose quantum-assisted activation for supervised learning in healthcare-based intrusion detection systems [23]. Li-Hua Gong et al. introduce a quantum k-nearest neighbor classification algorithm via a divide-and-conquer strategy [24] and also propose a quantum convolutional neural network based on variational quantum circuits [25]. Arsenii Senokosov et al. explore quantum machine learning for image classification [26]. Despite enhanced performance, these approaches are still in early-stage development and may face technical and regulatory challenges, such as high costs of building and maintenance, complex integration, and limited expertise. Nevertheless, they present opportunities for revolutionary exploration in future studies.

Our study was inspired by explainable sensor fault detection (E-SFD) [4], and the HAI dataset was used for multiple anomaly detection and explanation. In contrast to E-SFD, our research focuses on individual attack case analysis, detection, and explanation in the initial and RPS cases, providing a broad and more discerning approach to understanding each attack and NPP system. Unlike other studies that focused on broad areas of anomaly detection, our focus was on using a time-series deep learning model trained exclusively on normal data. The proposed model learns what 'normal' data looks like and becomes sensitive to even small deviations. This research was further extended by including XAI, which offers deeper insights into how distinctive features contribute to the model's decision.

The proposed model identified anomalies as the attack began, even before the RPS was triggered, providing insight into a possible use case for real-time anomaly detection.

3. Asherah Nuclear Power Plant Simulator

The Asherah NPP simulator (ANS) [1] is a hypothetical nuclear facility designed to identify potential issues present in the computer systems of NPPs, whereby potential demonstrations and countermeasures can be obtained. It was designed and developed by the University of Sao Paulo (USP) through collaborations with the International Atomic Energy Agency (IAEA) Coordinated Research Project (CPR) 'Enhancing Computer Security Incident Response at Nuclear Facilities' (J02008). It is suitable for digital instrumentation and control (I&C) research cyber security evaluation and computer security standards development. The ANS is a type of pressurized water reactor (PWR) power plant, implemented in MATLAB/Simulink. It simulates nuclear operations and the controller's dynamics system and is the heart of a complete hardware-in-the-loop (HIL) simulation environment. It comprises primary and secondary components as shown in Figure 1.

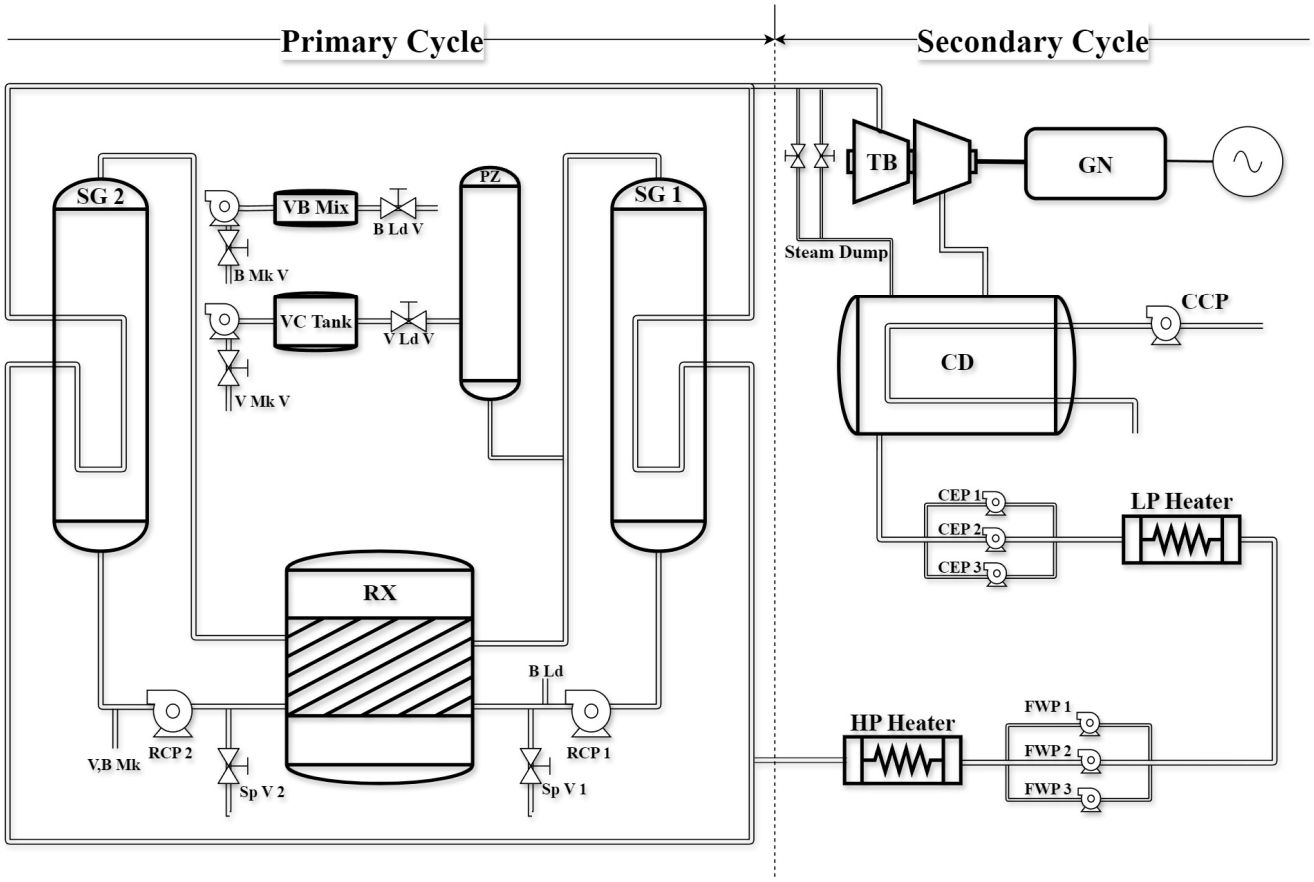

Figure 1. Basic Asherah NPP simulator.

The researchers and developers related to Asherah NPP ensure that the simulator conforms to the following standards and principles through the coordinated research project (CRP) and its international cooperative environment, as well as the IAEA's institutional role [1]:

- **Accuracy**: the simulator should not be unduly complicated and sufficiently accurate to satisfy the demands of the team.
- **Real-time**: real-time (preferably considerably quicker than real-time) simulations on standard computing platforms should be possible with the simulator.

- **Simplicity**: The learning curve should be minimal, and the simulator code should be simple enough to be tweaked by the teams, as required. Additionally, the physics concepts used must be sufficiently simple for staff members who are not experts in physics or nuclear engineering to comprehend.
- **Flexibility and modularity**: Depending on the requirements of the team, the simulator should be adaptable enough to be incorporated into various testbeds and deployed in various scenarios. To facilitate integration, the functional and physical units of a typical plant should be linked to the simulation modules.
- **Network and equipment integration**: to assist in security research and simulated attacks, the simulator should be easily integrable with other network equipment, communication protocols, and realistically simulated or physical controllers.
- **Non-proprietary**: to enable free code distribution among teams in other nations, the simulator design should not include proprietary, restricted, or confidential information or data.
- **Technological neutrality**: the simulator should not mimic an actual facility, to prevent malevolent agents from using it for training or development.

3.1. Primary Cycle

The heat generated in the reactor core (RX) was transferred to the steam generator (SG). The steam generated in SGs is used to operate the turbine (TB). The pressurizer (PZ) is a primary loop component whose purpose is to maintain the pressure at a given set-point, keeping it dependent on the reactor temperature. The reactor coolant pump (RCP) ensures shaft leak-tightness without the need for an active seal water supply system when the pump is ideal and is operated by cool air.

3.2. Secondary Cycle

In the secondary cycle, when the steam generated by the SGs arrives at the TB, the electrical generator converts the kinetic energy into electrical energy. The steam is separated by a moisture separator and passed through a condenser (CD) that reconverts the steam into water. This condensed water is again passed to the SGs with the help of the Feed Water Pump (FWP).

3.3. Reactor Protection System (RPS)

The Asherah simulator has a built-in protection system that activates automatic defense functions to bring the plant into a controlled state if unusual activities are detected. It validates the coolant pressure boundary in the reactor and can safely shut down the reactor to keep it in a normal state and has the capability to prevent and reduce the effects of unwanted incidents that can lead to potential damage.

In the simulator, the RPS provides a visual representation of certain features that help the operator identify whether the reactor is performing normally (Figure 2a). If the RPS is disabled, as illustrated in Figure 2b, even if an anomaly is detected, the reactor will not shut down, resulting in the plant moving to an unsafe state. As shown in Figure 2c, enabling it will help the reactor shut down automatically and maintain its safety status after the shutdown when an anomaly is detected. In this study, RPS was used to gain insight into the reactor's working mechanism.

When an anomaly is injected into the Asherah NPP, even a single feature out of eight can trigger the scram of the reactor. These conditions are listed in Table 1. The sensor values meeting these conditions can immediately trigger one or multiple features, leading to reactor scram.

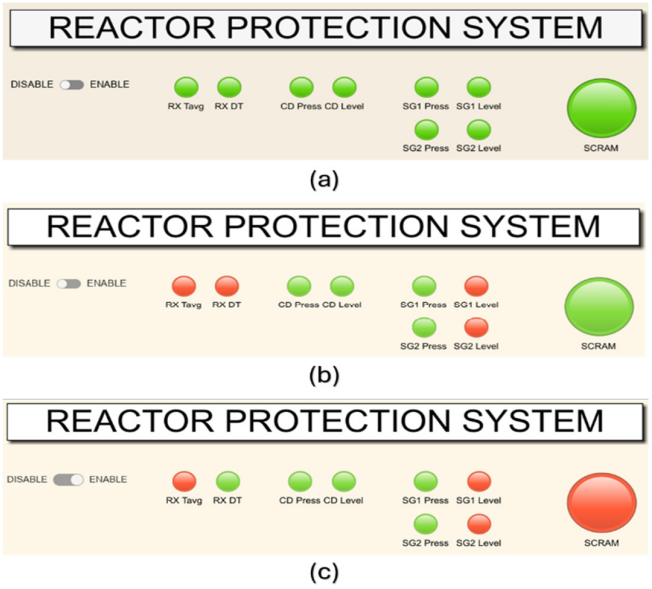

Figure 2. Three different RPS states: (**a**) normal, (**b**) under attack with RPS disabled, and (**c**) under attack with RPS enabled.

Table 1. RPS features and their limits compared to the normal state.

RPS Features	Condition	Normal
CD_Level	≥1.5 m	1 m
CD_Press	≥6760 Pa	5200 Pa
RX_DT	≥40 K	27.5 K
RX_Tavg	≥580 K	575.8 K
SG1_Level	≤12 m	15 m
SG1_Press	≥8,974,000 Pa	6,353,112 Pa
SG2_Level	≤12 m	15 m
SG2_Press	≥8,974,000 Pa	6,353,112 Pa

The CD_Level is the hotwell level, which is maintained within limits by automatically transferring the condensate to or from the storage system. CD_Press is the pressure of the condenser, which is monitored to verify condenser operation.

The RX_DT is the temperature difference, and RXTavg is the average reactor temperature of the inlet and outlet coolant operating temperatures. The SG1_Level and SG2_Level represent the water levels in SGs. Similarly, SG1_Press and SG2_Press denote the pressure inside the SGs. For example, if SG1_Level drops less than 12 m, the RPS can scram the reactor.

4. Attack Scenario

Asherah, a hypothetical model built in MATLAB, was simulated to explore the vulnerabilities of NPPs. This also helps explore the line between normal and abnormal behavior and the essence of the RPS within the NPP system. The attacks were created by adding a custom MATLAB function that allowed altering the system input values as shown in Figure 3. All three attacks successfully disturbed the normal process and created a precise anomaly as anticipated, triggering the RPS.

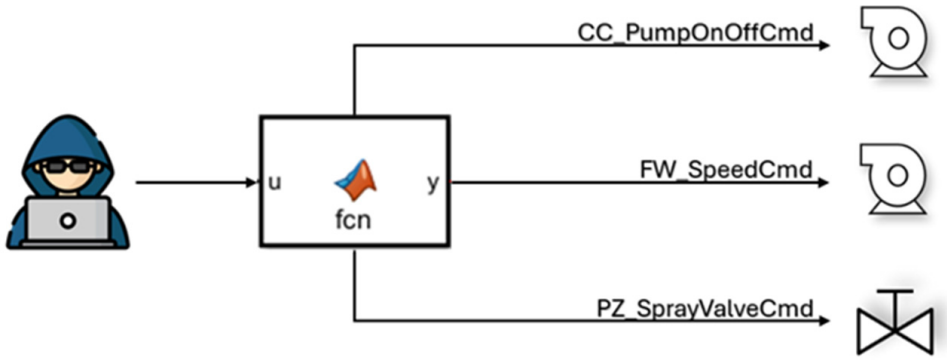

Figure 3. Attack scenarios in different systems.

4.1. Condenser Cooling Pump On–Off Command

The first target was the condenser cooling (CC) pump, which plays a critical role in maintaining the temperature of the CD by circulating cool water through it. The goal was to block the water supply to the CD to observe how the system responds and how and when the RPS is triggered. Using a custom MATLAB function, CC_PumpOnOffCmd was set to zero, and the cool water supply was cut off.

The impact was immediate with the CC Pump speed dropping to zero, as shown in Figure 4a. Within 3 s of the attack, the pressure of CD began to increase, exceeding the safety limits, as shown in Figure 4b. The RPS was triggered, shutting down the reactor power as the pressure crossed the threshold, to prevent further damage, as illustrated in Figure 4c.

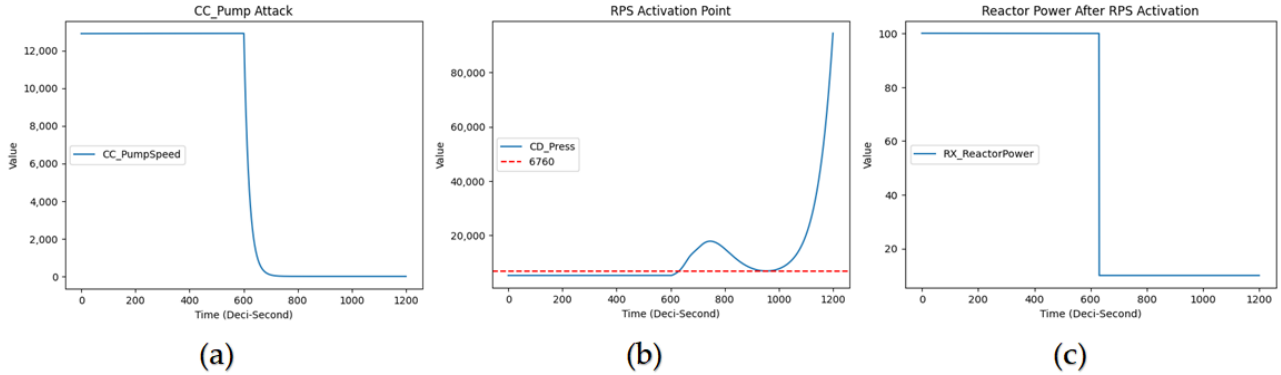

Figure 4. (**a**) CC Pump speed after the attack, (**b**) RPS activation point, and (**c**) shutdown of the reactor after RPS activation.

4.2. Feed Water Pump Speed Command

The second focus of our experiment was the FW Pump, which is responsible for moving cooled water from the CD to the SG. The objective of this study was to stop the water flow to the SG, expecting a shortage of water in steam generation. If the FW Pump is not working, the temperature in the reactor increases. Using a custom MATLAB function, FW_SpeedCmd was set to zero to effectively cut the water supply.

This effect was instantaneous as the water flow dropped sharply, as illustrated in Figure 5a. As expected, the water level in the SG began to decrease, as shown in Figure 5b. When it dropped below the safety limit, the RPS was activated. Interestingly, it took nearly 54 s for the RPS to be activated, as shown in Figure 5b. Subsequently, the reactor was shut down to prevent further damage, as shown in Figure 5c. The delayed action of the RPS highlights the possibility of serious threats to the reactor.

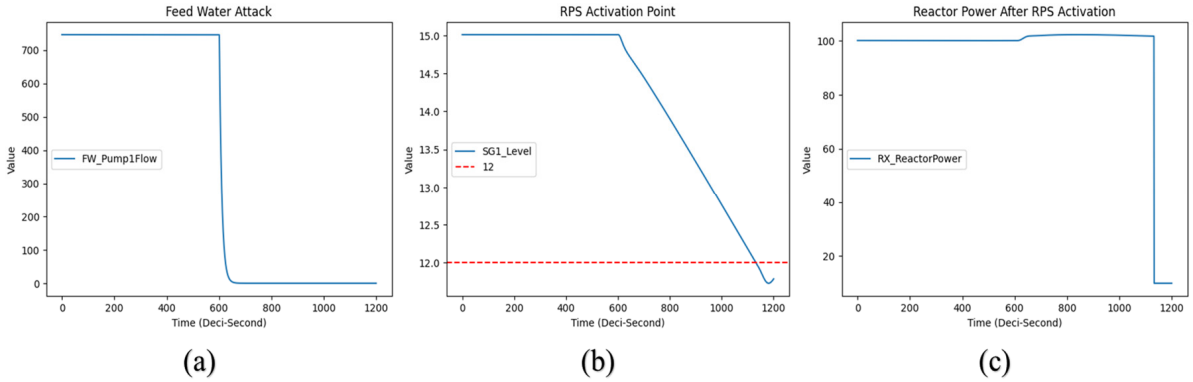

Figure 5. (**a**) FW Pump flow after attack, (**b**) RPS activation point, and (**c**) shutdown of the reactor after RPS activation.

4.3. Pressurizer Spray Valve Command

The third target in our simulator was the pressurizer spray valve, a critical primary cooling loop component that maintains high pressure, to prevent boiling and control the water levels.

The pressure was intentionally released in the PZ by opening the valve, using a custom MATLAB function, which set the PZ_CL1SprayValveCmd to 100, as depicted in Figure 6a. This action immediately caused a sharp increase in the temperature of the reactor core, causing it to increase beyond the safety limit, as shown in Figure 6b. However, it took approximately 18 s for the temperature to exceed the threshold, which activated the RPS, whereby the RPS responded by shutting down the reactor power to prevent further escalation, as shown in Figure 6c.

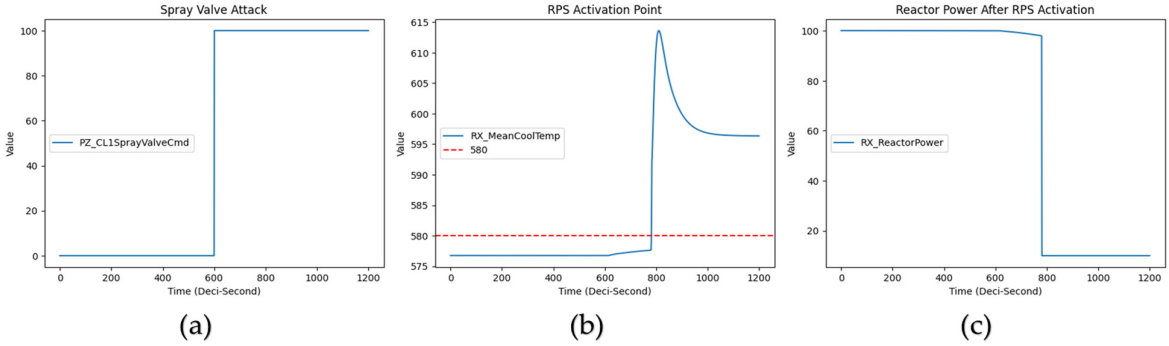

Figure 6. (**a**) Pressurizer spray valve command after the attack, (**b**) RPS activation point, and (**c**) shutdown of the reactor after RPS activation.

5. Anomaly Detection in Nuclear Power Plants

5.1. Dataset Description

The MATLAB simulator was operated separately for each attack and training dataset. Each attack and training dataset was collected in separate files for further individual analysis as shown in Tables 2 and 3.

Table 2. Dataset description.

Type	Fields	Instances	Durations (min)
Normal	17	3000	5
CC Pump on–off	17	1200	2
FW Pump	17	1200	2
Pressurizer spray valve	17	1200	2

Table 3. Systems and related fields [1].

SN	System	Fields	Units
1	Time	Tout	Deci-second (ds)
2	Reactor (RX)	Reactor power	Watt (W)
3		RX_MeanCoolTemp	Kelvin (K)
4		CTRL_RXPowerSetPoint	Watt (W)
5	Reactor Cooling 1 (RC1)	RC1_PumpFlow	kg/s
6		RC1_PumpSpeed	Percentage (%)
7	Pressurizer (PZ)	PZ_Press	Pascal (Pa)
8		PZ_Level	Meter (m)
9		PZ_CL1SprayValveCmd	Percentage (%)
10	Steam generator 1 (SG)	SG1_Level	Meter (m)
11		SG1_Press	Pascal (Pa)
12	Condenser cooling (CC) Pump	CC_PumpSpeed	Percentage (%)
13	Condenser (CD)	CD_Press	Pascal (Pa)
14	Control Rod (CR)	CR_PosCmd	Step
15	Feed Water (FW) Pump	FW_Pump1Flow	kg/s
16		FW_Pump1Speed	Percentage (%)
17	Turbine (TB)	TB_InSteamPress	Pascal (Pa)

5.2. Proposed Approach

The Bi-LSTM model is effective in anomaly detection of time-series or sequential data. Although the Bi-LSTM is a supervised model, it can still be effective in unsupervised anomaly detection. In this study, the main idea was to train the model only on normal data and then use the model to detect deviations that are likely to be anomalies. In this approach, the Bi-LSTM is trained to predict the next value in the time-series data based on previous values. When the models are presented with anomalous data, the prediction error increases because the data differ from what the model has learned. If the prediction error increases beyond the threshold, it is considered an anomaly. This approach allows Bi-LSTM to work in an unsupervised manner because no labeled anomalies are required.

The Bi-LSTM model was designed with three layers of LSTM units to capture both past and future dependencies in time-series data. A window size of 10 was used, meaning that the model processed a sequence of 10 time steps to predict the next value in the series. The bidirectional nature of LSTM allows it to process the input sequence in both the forward and backward directions, improving its ability to learn complex patterns in the data. The key parameters were as follows:

- Input window size: 10 time steps.
- Number of layers: three stacked Bi-LSTM layers.
- Number of units per layer: 100 hidden units in each LSTM layer.
- Batch size: 512 samples per batch.
- Epochs: 15 epochs.

In this study, the dataset has total features, and the model attempts to predict the next time step using the window of the previous time step as the input. The input data *Xi*

consist of window size w, which uses w time steps as input, each with m features. The data for each time step were arranged as follows:

$$Xi = \begin{bmatrix} t_i = & (f_1(i)\ f_2(i)\ \cdots\ f_m(i)) \\ t_{i+1} = & (f_1(i+1)\ f_2(i+1)\ \cdots\ f_m(i+1)) \\ \cdots & \cdots\ \cdots\ \ddots \\ t_{i+w-1} = & (f_1(i+w-1)\ f_2(1+w-1)\ \cdots\ f_m(i+w-1)) \end{bmatrix} \quad (1)$$

where in the first row, $f_1(i), f_2(i), \ldots, f_m(i)$ are the values of the m features at time step t_i. Similarly, the second row shows the feature values at t_{i+1} and so on.

The target output is y_i, which is the next time step t_{i+w}. This indicates that the model is trained to predict all features at time step t_{i+w} based on the input, which includes the feature values from i to $i + w - 1$.

$$y_i = (t_{i+w} = \begin{bmatrix} f_1(i+w) & f_2(i+w) & \ldots & f_m(i+w) \end{bmatrix}) \quad (2)$$

The predicted output is $y_i{}'$ of the next time step t_{i+w}.

$$y_i{}' = (t'_{i+w} = \begin{bmatrix} f_1{}'(i+w) & f_2{}'(i+w) & \ldots & f_m{}'(i+w) \end{bmatrix}) \quad (3)$$

The mean squared error (MSE) is used to measure how well the model predicts the target values. The MSE is calculated as the average of the squared difference between the actual and predicted values for each feature at t_{i+w}.

$$\text{MSE}_i = \frac{1}{m} \sum_{j=1}^{m} (f_j{}'(i+w) - f_j(i+w))^2 \quad (4)$$

where MSE_i is the MSE score at the ith step; $y_i{}'$ is the predicted value, and y_i is the actual value as shown in Figure 7. The division by m normalizes the errors across all features.

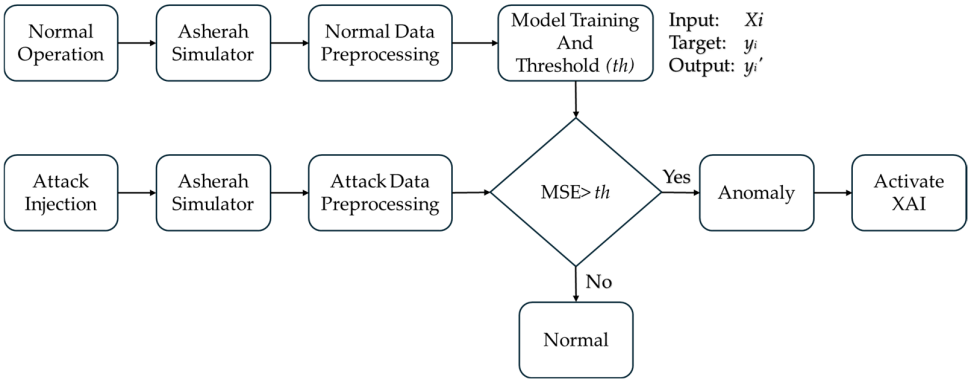

Figure 7. Proposed approach.

5.3. Threshold

The model was trained only on normal data; therefore, it could only detect minor changes in the test dataset. To prevent false positives, the threshold was set based on the training loss values (MSE_i) from each epoch. Subsequently, the threshold (th) was defined using the following equations:

$$th = \max_{i \leq N} (MSE_i) \times k \quad (5)$$

where the parameter k is a positive integer that acts as a scaling factor to define the sensitivity of the anomaly detection system. If the value of k is set higher, it will be less sensitive to small deviations, leading to a lower detection of anomalies, and vice versa. In the context of NPPs, detecting even small deflections for safety and security is crucial,

whereby the scaling factor was set to 3 ($k = 3$) to balance sensitivity and robustness. The system was tuned to a threshold value $th \approx 0.007$, which was effective for detecting slight changes across different attack scenarios.

5.4. Anomaly Detection and Explanation

When the MSE value of any time step exceeds the threshold value, it is detected as an anomaly. The detection of the anomaly activates the XAI to identify the contributing features that can help in understanding the reason behind the detection. Setting up the rules for triggering XAI involves defining criteria that determine when and how to investigate the anomaly, as well as the method used to explain the model's behavior. In this research, the rule and the strategies employed were as follows:

Threshold-based activation:

- Rule: Activate XAI when the anomaly score crosses the predefined threshold.
- Explanation technique: SHAP [5] is used to determine which features contribute to the anomalous time step. For example, if an anomaly is detected at a certain point, XAI is triggered to explain which feature(s) are contributing to that point.

6. Experimental Results

In Section 6.1, the anomaly detection results are presented, in Section 6.2, the performance evaluation results of the model are presented, in Section 6.3, the feature contribution analyses at initial detection and after RPS activation are presented, and in Section 6.4, the proposed model is compared with another existing model.

6.1. Attack Detection

6.1.1. CC Pump Attack

When the CC Pump is blocked, the CD pressure increases. However, the RPS system requires 3 s for activation. Regardless, the proposed model can detect the anomaly instantly when an attack is initiated, as shown in Figure 8a, that is, 3 s earlier than when the RPS is activated. The initial detection causes a noticeable deflection. When the RPS is triggered, it causes a significantly higher deflection, indicating the occurrence of an attack.

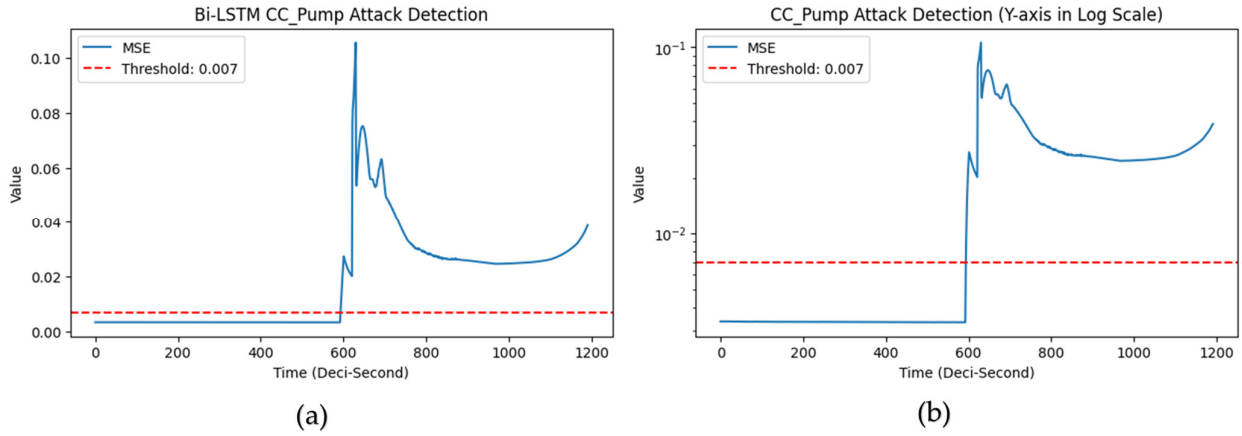

Figure 8. (**a**) CC Pump attack detection and (**b**) log-scale plot of (**a**).

6.1.2. FW Pump Attack

The speed of the FW Pump was set to zero to prevent water from flowing into the SG. This decreased the water level in the SG, eventually triggering the RPS system. The proposed model detected the anomaly as the attack began and again when the RPS was activated. However, the RPS system required almost 54 s to be activated, as represented by the second deflection in Figure 9a. This highlights the necessity for an immediate detection system to prevent the anomaly from causing harm.

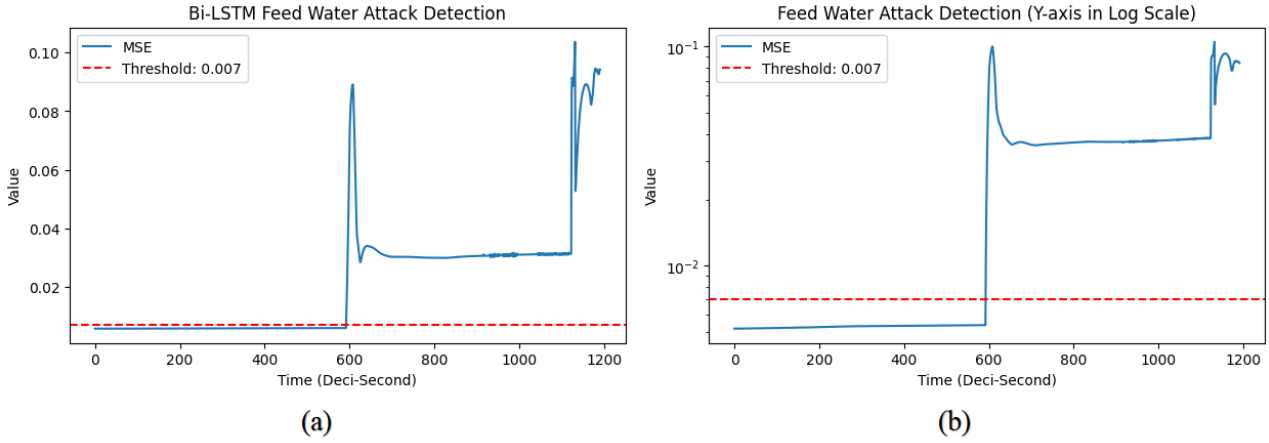

Figure 9. (**a**) FW Pump attack detection and (**b**) log-scale plot of (**a**).

6.1.3. PZ Spray Valve Attack

The PZ is a primary cooling loop component designed to maintain high pressure but does not allow boiling; it also controls the water level inside it. When the PZ valve is opened, it causes a sudden increase in the reactor temperature, whereby the proposed model is detected, as illustrated in Figure 10a. After 18 sec, when the temperature increases indicating abnormality, the RPS is activated, shutting down the reactor. The activation of the RPS is indicated by the second deflection in Figure 10a. As in the previous cases, the proposed model detected the anomaly instantly as the attack began and again showed a deflection when the reactor was shut down.

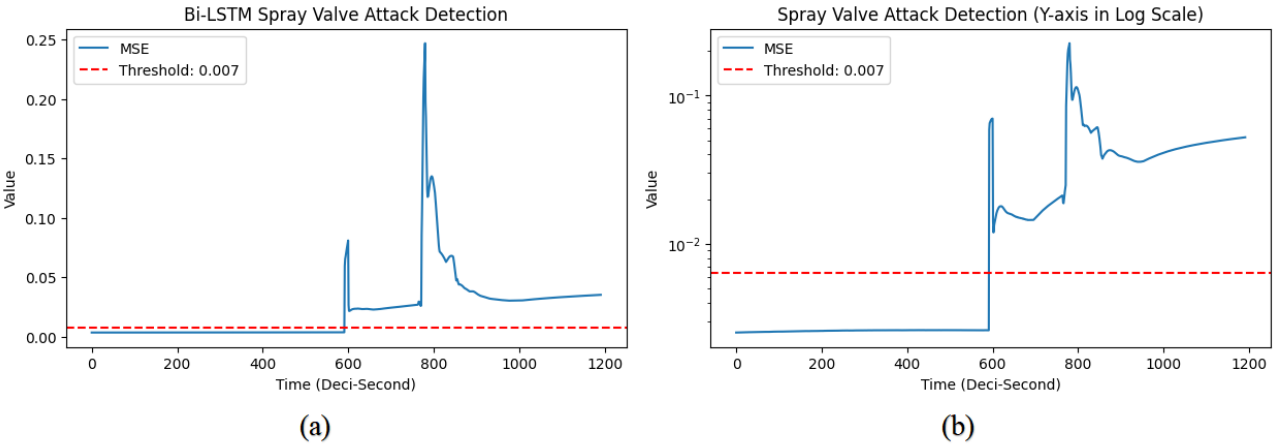

Figure 10. (**a**) PZ spray valve attack detection and (**b**) log-scale plot of (**a**).

6.2. Performance Evaluation

The performance of the proposed model is calculated based on recall (detection rate), false positive rate (*FPR*), and accuracy. For this, we need values of True Positives (*TP*), false positives (*FP*), True Negatives (*TN*), and False Negatives (*FN*), where the following are true:

Recall (detection rate), which measures how many actual positive cases were correctly predicted, is calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

FPR, which measures the proportion of normal cases that were incorrectly predicted as abnormal, is calculated as:

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

Accuracy tells how many of the total predictions were correct, and it is calculated as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (8)$$

The evaluation results are presented in Table 4.

Table 4. Performance evaluation.

Attack	*TP*	*FN*	*FP*	*TN*	*Recall*	*FPR*	*Accuracy*
CC Pump	600	0	0	590	1.00	0	1.00
FW Pump	600	0	0	590	1.00	0	1.00
PZ spray valve	600	0	0	590	1.00	0	1.00

The *recall* for each attack case is 1.0, meaning the model correctly identified all anomalies immediately from the attack start point. *FPR* is 0.0 for each attack, meaning the model accurately distinguished between normal and abnormal cases. Similarly, the *accuracy* for each attack case is also 1.0, meaning the model made correct predictions for all instances, both normal and anomalous.

The model's accurate performance could be due to the small volume and noise-free data. In future work, we will consider more complex and large-volume data, which will help to build a more robust model for a real-world environment implementation.

6.3. Initial Feature Contributors After RPS Activation

6.3.1. CC Pump

The attack shut down the CC Pump, which dropped the sensor reading as the water flow was blocked. After applying XAI for the initial deflection, as shown in Figure 11a, the CC_PumpSpeed field contributed more positively than the other features, as shown in Figure 11c. Such insights are helpful to operators in identifying the system or fields to focus on in the process of detecting the anomaly.

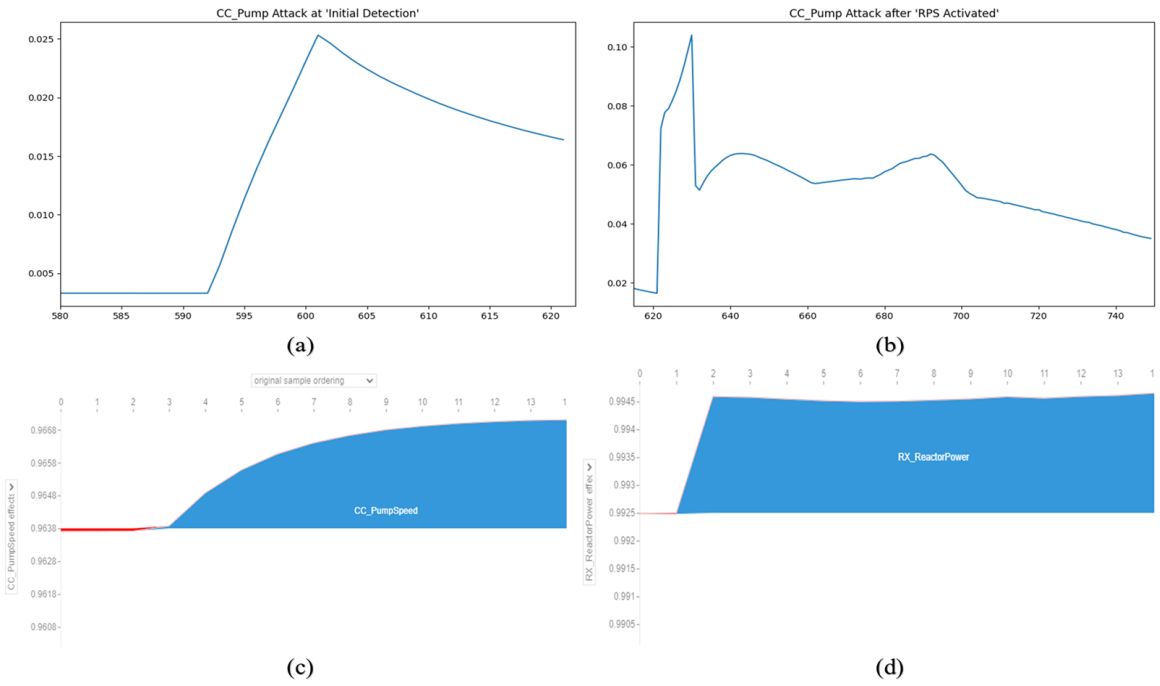

Figure 11. (**a**) CC Pump attack at 'Initial Detection', (**b**) CC Pump attack after 'RPS Activated', (**c**) CC_PumpSpeed at initial detection, and (**d**) RX_ReactorPower contribution after 'RPS Activated'.

The blockage of the CC Pump caused an increase in the pressure in the CD, which triggered the RPS system 3 s after the attack, causing a deflection more than the initial deflection, as shown in Figure 11b. Thus, the RPS shut down the power of the reactor, leading to the scram of the system. XAI showed that the core systems such as RX_ReactorPower and RX_MeanCoolTemp had a higher contribution than the other features, as shown in Figure 11d.

6.3.2. FW Pump

The attack targeted the FW Pump, decreasing the water flow rate in the SG. The XAI indicated that during the initial detection, illustrated in Figure 12a, the FW_Pump1Flow field contributes more than the other features, as shown in Figure 12c. Initially, no contribution is observed; however, as the deflection increases, the contribution of the FW_Pump1Flow feature gradually increases.

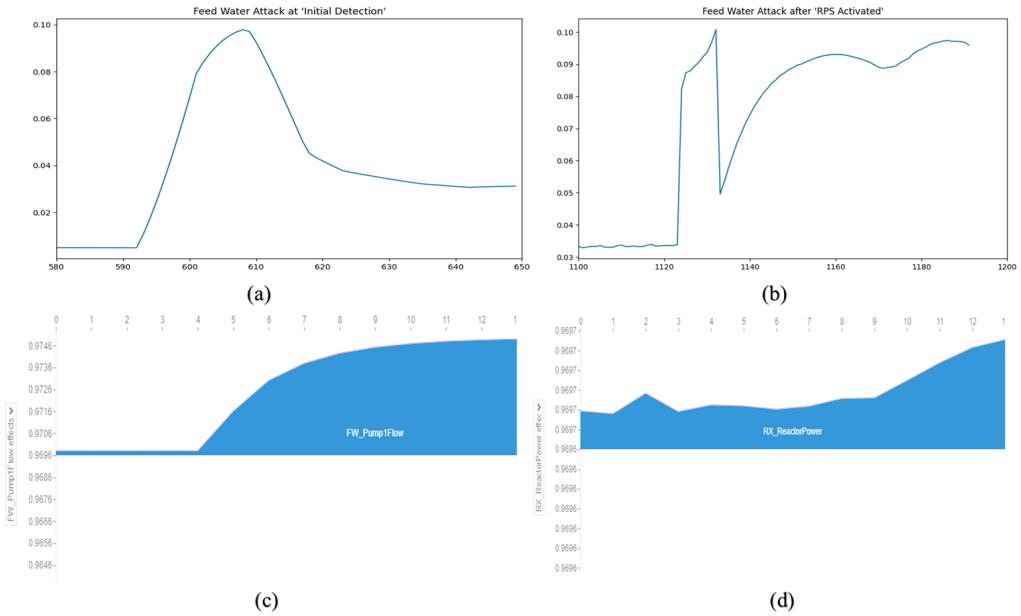

Figure 12. (**a**) FW Pump attack at initial detection, (**b**) FW attack after 'RPS Activated', (**c**) FW_Pump1Flow feature contribution by the FW Pump at initial detection, and (**d**) RX_ReactorPower feature contribution after RPS activation.

The blockage of the FW Pump causes a decrease in the water flow rate in the SG, which causes a drop in the SG1_level. The proposed model can detect the changes caused by the RPS system, as indicated in Figure 12b. Finally, when the RPS is activated, the ReactorPower features contribute more than the other features, as shown in Figure 12d.

6.3.3. PZ Spray Valve

The attack opens the pressurized spray valve. At initial detection, as shown in Figure 13a, the pressurizer spray valve command feature also contributes positively to the prediction. Therefore, instead of looking at the same feature, related features are observed, and PZ_Press's contribution increases gradually compared to other features, as seen in Figure 13c, apparently with no initial contribution to the initial deflection, although affected by the injected attack. Moreover, no deflection occurs until the RPS is activated, which takes place 18 sec after the attack is initiated. Therefore, PZ_Press can be considered as a contributing feature related to the attack target.

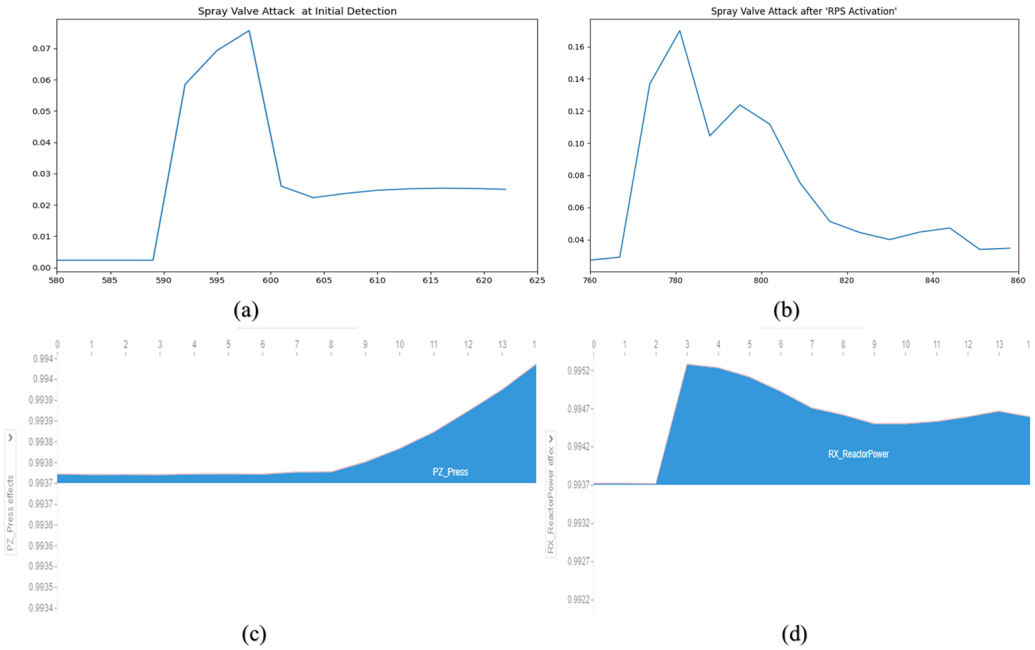

Figure 13. PZ spray valve (**a**) at initial detection and (**b**) after RPS activation; (**c**) PZ_Press feature contribution at initial detection; (**d**) RX_ReactorPower feature contribution after RPS activation.

The PZ spray valve is a cooling component. When set to open, it can no longer maintain the temperature, which eventually causes an increase in temperature. It took approximately 18 s to activate the RPS system, again causing deflection, as shown in Figure 13b, which is also an ineffective system in nuclear power plants. In the RPS activation, reactor power features were found contribute more than the other features, as illustrated in Figure 13d.

6.4. Comparison with TranAD

For multivariate time-series data, TranAD [27] is an anomaly detection approach based on deep transformers. In order to draw conclusions based on awareness of the larger temporal trends in the data, an attention-based sequence encoder is employed. Score-based self-conditioning is used to provide stable adversarial training and robust multi-model feature extraction. Additionally, model-agnostic meta-learning (MAML) is used to train the model with sparse data.

A comparison between TranAD and the proposed model is conducted to detect the anomaly mentioned in Section 4. Figure 14 shows that all the attack cases were detected instantly as they began by the TranAD model as well. However, in the case of the CC Pump attack, as shown in Figure 14a, the continuous rise in the MSE score made it difficult to identify the RPS activation point, whereas the RPS activation point can be identified easily in the case of Feed Water and spray valve attacks, as illustrated in Figures 14b and 14c, respectively.

The immediate anomaly detection of both models shows their fairness, transparency, and effectiveness. There are certain reasons why we preferred the Bi-LSTM model over TranAD. Bi-LSTM is typically easier to implement, debug, and maintain in a production environment, making it a more practical and cost-effective solution than transformer-based models, which are more complex in implementation. Even though both are neural network models and are considered as black boxes, we found that it was much easier to apply XAI in Bi-LSTM than TranAD. Hence, due to simplicity and easier model interpretability, this research preferred the Bi-LSTM model for anomaly detection.

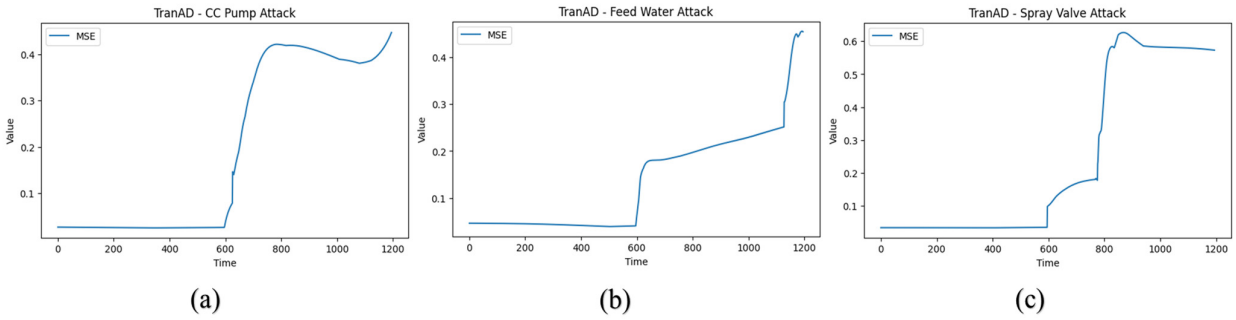

Figure 14. Anomaly detection by TranAD on (**a**) CC Pump attack, (**b**) Feed Water attack, and (**c**) pressurizer spray valve attack.

7. Discussion

This study addresses the critical challenges of detecting cyberattacks in NPPs using the Asherah NPP simulator. This study focused on analyzing the normal and abnormal behaviors of the system by injecting three different attacks. The proposed model detected existing anomalies before the RPS was triggered. Furthermore, the decisions made by the model were interpreted using XAI to improve transparency and trust in the decision making process. The virtual testbed used in this study provided a controlled and risk-free environment.

However, the results may differ when applied to an actual physical system, where the cyberattack may be more complex because of the hardware and complex system architecture. The model was trained on normal operation data. However, in real-world scenarios, obtaining pure data is challenging because system noise or irregularities disturb the system (such as environmental conditions or mechanical failures) that are not anomalous but can be detected as anomalies.

The proposed model, with its early or instant detection, demonstrated its capability to detect entirely new or unknown attacks. This model is generalizable and is expected to perform well in different plants, arrangements, and even in other industries. However, the model may require retraining or fine-tuning to be applied effectively.

In future studies, the applications can be expanded to more diverse attacks, including environmental and system disturbances, to evaluate the robustness and adaptability of anomaly detection in more complex and realistic environments. The integration of the anomaly detection model in a real-time operation can also be a valuable next step in assessing its real-world performance in investigating challenges and issues. Furthermore, we will consider a dynamic threshold over a static threshold approach. This will help to reduce false detection, which is likely to be caused by environmental and system noise, creating a more robust model in a complex and realistic environment.

Similarly, quantum computing holds the potential to revolutionize anomaly detection in NPP, but the current limitations in quantum hardware, limited expertise, and integration issues make it infeasible, and it remains a technology for future exploration. Furthermore, adaptive models that can learn and evolve within a system over time can also be applied. Such models would adjust to new patterns in the system behavior, improving the ability to detect new attacks without the need for constant retraining.

8. Conclusions

This study significantly advances the field of anomaly identification for vital infrastructures such as NPPs. Using a Bi-LSTM deep learning model, we identified anomalies that arose from cyberattack simulations in the Asherah Nuclear Power Plant simulator. To prevent system failures and ensure operational safety, early threat identification is critical. The model's capacity to detect abnormalities prior to triggering the RPS emphasizes this effectiveness. By introducing transparency into the decision making process, XAI can significantly improve the usefulness of the model. The SHAP analysis enabled us to understand

the role of the parameters, including pressure and flow rates, on anomaly detection. This information provides an important context for comprehending and reducing cyber threats.

Ultimately, XAI's interpretability, coupled with the predictive capability of deep learning, provides a reliable and robust method for recognizing and explaining anomalies in vital industrial systems. In addition to strengthening cybersecurity resilience, this method increases operator confidence in deep learning models by facilitating the understanding and validation of the model's conclusions. In demonstrating that both detection and explainability are important for industrial operation security, this study contributes to implementing comparable strategies in other high-stakes settings.

Author Contributions: Conceptualization, A.C. and S.C.; methodology, A.C. and J.H.; software, J.H.; validation, J.H., S.K. and A.K.; writing—original draft preparation, A.C.; writing—review and editing, A.K. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the Operation of Nuclear Power Plant Cybersecurity Verification Testbed project of KINAC in 2024 and partially supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2021-KA163348) and partially supported by Regional Innovation Strategy (RIS) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2023RIS-008).

Data Availability Statement: The data are not available because of security limitations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Silva, R.A.B.E.; Shirvan, K.; Piqueira, J.R.C.; Marques, R.P. Development of the asherah nuclear power plant simulator for cyber security assessment. In Proceedings of the International Conference on Nuclear Security, Vienna, Austria, 10–14 February 2020.
2. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *Assoc. Comput. Mach.* **2023**, *55*, 1–33. [CrossRef]
3. Silva, R.B.E.; Correa, D.A.; Antunes, F.R.; Souza, F.C.S.; Piqueira, J.R.C.; Marques, R.P. The Asherah Nuclear Power Plant Simulator (ANS) as a Training Tool at the Brazilian Cyber Guardian Exercise. In Proceedings of the International Conference on Nuclear Security 2020, Vienna, Austria, 10–14 February 2020.
4. Hwang, C.; Lee, T. E-SFD: Explainable Sensor Fault Detection in the ICS Anomaly Detection System. *IEEE Access* **2021**, *9*, 140470–140486. [CrossRef]
5. Lundberg, S.M.; Nair, B.; Vavilala, M.S.; Horibe, M.; Eisses, M.J.; Adams, T.; Liston, D.E.; Low, D.K.-W.; Newman, S.-F.; Kim, J.; et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2018**, *2*, 749. [CrossRef] [PubMed]
6. Kessides, I.N. The future of the nuclear industry reconsidered: Risks, uncertainties, and continued promise. *Energy Policy* **2012**, *48*, 185–208. [CrossRef]
7. Conti, M.; Donadel, D.; Turrin, F. A Survey on Industrial Control System Testbeds and Datasets for Security Research. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 2248–2294. [CrossRef]
8. Shin, H.-K.; Lee, W.; Yun, J.-H.; Kim, H. HAI 1.0: HIL-Based Augmented ICS Security Dataset. In Proceedings of the 13th USENIX Workshop on Cyber Security Experimentation and Test (CSET 20), Online, 10 August 2020; USENIX Association: Washington, DC, USA, 2020.
9. Fovino, I.N.; Masera, M.; Guidi, L.; Carpi, G. An experimental platform for assessing SCADA vulnerabilities and countermeasures in power plants. In Proceedings of the 3rd International Conference on Human System Interaction, Rzeszow, Poland, 13–15 May 2010.
10. Davis, C.M.; Tate, J.E.; Okhravi, H.; Grier, C.; Overbye, T.J.; Nicol, D. SCADA Cyber Security Testbed Development. In Proceedings of the 2006 38th North American Power Symposium, Carbondale, IL, USA, 17–19 September 2006.
11. Mallouhi, M.; Al-Nashif, Y.; Cox, D.; Chadaga, T.; Hariri, S. A testbed for analyzing security of SCADA control systems (TASSCS). In Proceedings of the ISGT 2011, Anaheim, CA, USA, 17–19 January 2011.
12. Hui, H.; Maynard, P.; McLaughlin, K. ICS Interaction Testbed: A Platform for Cyber-Physical Security Research. In Proceedings of the 6th International Symposium for ICS & SCADA Cyber Security Research 2019 (ICS-CSR), Greece, Athens, 10–12 September 2019.
13. Joel Barnard, C.S.; What is anomaly detection? IBM. 12 December 2023. Available online: https://www.ibm.com/topics/anomaly-detection (accessed on 21 July 2024).
14. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 1–58. [CrossRef]
15. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* **2021**, *54*, 1–38. [CrossRef]
16. Jiang, M.; Hou, C.; Zheng, A.; Hu, X.; Han, S.; Huang, H.; He, X.; Yu, P.S.; Zhao, Y. Weakly Supervised Anomaly Detection: A Survey. *arXiv* **2023**, arXiv:2302.04549.

17. Jiang, M.; Hou, C.; Zheng, A.; Han, S.; Huang, H.; Wen, Q.; Hu, X.; Zhao, Y. ADGym: Design Choices for Deep Anomaly Detection. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks, New Orleans, LA, USA, 10–16 December 2023.
18. Ahmad, Z.; Khan, A.S.; Nisar, K.; Haider, I.; Hassan, R.; Haque, M.R.; Tarmizi, S.; Rodrigues, J.J.P.C. Anomaly Detection Using Deep Neural Network for IoT Architecture. *Appl. Sci.* **2021**, *11*, 7050. [CrossRef]
19. Gaggero, G.B.; Caviglia, R.; Armellin, A.; Rossi, M.; Girdinio, P.; Marchese, M. Detecting Cyberattacks on Electrical Storage Systems through Neural Network Based Anomaly Detection Algorithm. *Sensors* **2022**, *22*, 3933. [CrossRef] [PubMed]
20. Chemali, E.; Kollmeyer, P.J.; Preindl, M.; Ahmed, R.; Emadi, A. Long Short-Term Memory Networks for Accurate State-of-Charge Estimation of Li-ion Batteries. *IEEE Trans. Ind. Electron.* **2017**, *65*, 6730–6739. [CrossRef]
21. Catillo, M.; Pecchia, A.; Villano, U. CPS-GUARD: Intrusion detection for cyber-physical systems and IoT devices using outlier-aware deep autoencoders. *Comput. Secur.* **2023**, *129*, 103210. [CrossRef]
22. Wang, Y.; Peng, H.; Wang, G.; Tang, X.; Wang, X.; Liu, C. Monitoring industrial control systems via spatio-temporal graph neural networks. *Eng. Appl. Artif. Intell.* **2023**, *122*, 106144. [CrossRef]
23. Laxminarayana, N.; Mishra, N.; Tiwari, P.; Garg, S.; Behera, B.K.; Farouk, A. Quantum-Assisted Activation for Supervised Learning in Healthcare-Based Intrusion Detection Systems. *IEEE Trans. Artif. Intell.* **2024**, *5*, 977–984. [CrossRef]
24. Gong, L.H.; Ding, W.; Li, Z.; Wang, Y.Z.; Zhou, N.R. Quantum K-Nearest Neighbor Classification Algorithm via a Divide-and-Conquer Strategy. *Adv. Quantum Technol.* **2023**, *7*, 2300221. [CrossRef]
25. Gong, L.H.; Pei, J.J.; Zhang, T.F.; Zhou, N.R. Quantum convolutional neural network based on variational quantum circuits. *Opt. Commun.* **2024**, *550*, 129993. [CrossRef]
26. Senokosov, A.; Sedykh, A.; Sagingalieva, A.; Kyriacou, B.; Melnikov, A. Quantum machine learning for image classification. *Mach. Learn. Sci. Technol.* **2024**, *5*, 015040. [CrossRef]
27. Tuli, S.; Casale, G.; Jennings, N.R. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *arXiv* **2022**, arXiv:2201.07284. [CrossRef]