Assessing Cybersecurity Risks in AI-Based Nuclear Power Plant Operator Support System: A Study on Potential Attacks and Human-Machine Interaction Implications

Young Ho Chae[a*], Seung Geun Kim[a] , Seo Ryong Koo[a]

[a]Korea Atomic Energy Research Institute, Daejeon, Republic of Korea

Abstract: The adoption of artificial intelligence (AI) in nuclear power plant(NPP) operator support systems has the potential to enhance efficiency and safety. However, the increasing reliance on data-driven approaches introduces new cybersecurity risks. This study aims to assess the potential threats and attack methods that AI-based nuclear power plant diagnostic support systems may face, focusing on the Diagnosis module which is being developed in Korea Atomic Energy Research Institute(KAERI). By examining the impact of attacks such as Poisoning, Evasion, Extraction, and Inference, to compare the relative importance of data and identify potential risks. The study employs Explainable AI methodologies to evaluate critical variables and their domains, and explores the feasibility of inducing misdiagnosis using adversarial attacks. Evasion attacks, including Threshold and Decision-based/Boundary-based attacks, and Poisoning attacks using Hidden Trigger Backdoors, are investigated. The effectiveness of mitigation strategies, such as General Adversarial Training, Mix-up, and Label Smoothing, is also studied. The future findings are expected to guide the development of robust cybersecurity frameworks and contribute to the safe and reliable operation of nuclear power plants in the face of evolving cyber threats.

Keywords: Deep Learning Model, Operator Support System, Evasion Attack, Mitigation Strategy.

1. INTRODUCTION

The instrumentation and control(I&C) systems of nuclear power plants have undergone a continuous evolution and advancement over the decades. The transition from analog to digital systems has marked a significant milestone in this progression. Presently, researchers are actively investigating methods to harness the vast quantities of data amassed by these digital systems to develop artificial intelligence (AI) solutions that can further augment the efficiency and safety of nuclear power plant operations. Researches have demonstrated the efficacy of AI methodologies in accurately diagnosing abnormal and emergency situations, underscoring the potential advantages of integrating AI into nuclear power plant control systems [1-6].
However, the increasing digitalization and dependence on data have also introduced a novel set of challenges and risks. One of the most pressing concerns is cybersecurity. The nuclear industry received a stark reminder of the vulnerability of these systems to cyber threats in the early 2000s when the Stuxnet attack targeted nuclear facilities. This incident catalyzed the cybersecurity community to focus on identifying and safeguarding critical digital assets within nuclear power plants. Researchers have been diligently working on developing robust protective measures and countermeasures to mitigate the risks associated with cyber attacks on these critical systems.

It is anticipated that AI methodologies will find their primary application in operator support systems for nuclear power plants. AI-based systems rely heavily on data-driven approaches, which represents a significant shift from traditional software-based systems. Historically, risk mitigation efforts for critical digital assets primarily focused on securing the software and devices themselves. However, with the advent of data-driven AI methodologies, the importance of data integrity, as well as the model's learning process and method, is expected to come to the forefront.

This transition from inevitably expands the attack surface, necessitating the definition and protection of critical data assets alongside the traditional focus on critical digital assets. As AI systems become more integrated into nuclear power plant operations, ensuring the security and integrity of the data used to train and operate these systems will be of paramount importance.

In light of these developments, this study aims to provide a preliminary assessment of the potential threats and attack methods that AI-based nuclear power plant diagnostic support systems may face. By examining the

impact of these attacks, We tried to compare the relative importance of different data types and identify the potential risks associated with such attacks. This analysis will contribute to a more comprehensive understanding of the cybersecurity in the context of AI-based softwares in NPP and help inform the development of effective strategies to mitigate these risks.

By elucidating the unique challenges and vulnerabilities associated with AI-based systems in the nuclear industry, this research will help guide the development of robust cybersecurity frameworks and best practices. Furthermore, the insights gained from this study will contribute to the ongoing efforts to ensure the safe and reliable operation of nuclear power plants in the face of evolving cyber threats.

In conclusion, this study represents a crucial step towards understanding and addressing the cybersecurity risks associated with the adoption of AI methodologies in nuclear power plant operator support systems. By examining potential attack vectors and their impacts, we aim to provide valuable insights that will inform the development of effective risk mitigation strategies and contribute to the overall safety and security of nuclear power plant operations in the digital age.

2. AI-BASED OPERATOR SUPPORT SYSTEM (TARGET SYSTEM)

The schematic diagram of developing AI-based Operator Support System is shown in Figure. 1. The target system is being developed by the Korea Atomic Energy Research Institute (KAERI) since 2022. The objective of developing this system is to diagnose abnormal situations in power plants at an early stage and provide accurate and prompt guidance for appropriate actions. After the system is developed, it will be applied and evaluated using the simulation testbed of Korea Hydro & Nuclear Power's nuclear power plants. The objective of the developing The AI-based operator support system for nuclear power plants (NPPs) consists of several key modules, each designed to enhance the efficiency and safety of plant operations. The primary objective of the signal verification and validation (V&V) and recovery module is to ensure data integrity, which is crucial for the proper functioning of systems with deep neural networks. The module employs a Variational Auto-Encoder (VAE)-based learning algorithm, trained on diverse datasets, to identify and address faulty or untrained signals. If the data is recognized as previously trained, the measurement signal is directly transmitted to the subsequent modules. If the data appears untrained, the signal's integrity is further evaluated, and a signal reconstruction algorithm is employed to restore the signal's integrity if necessary.

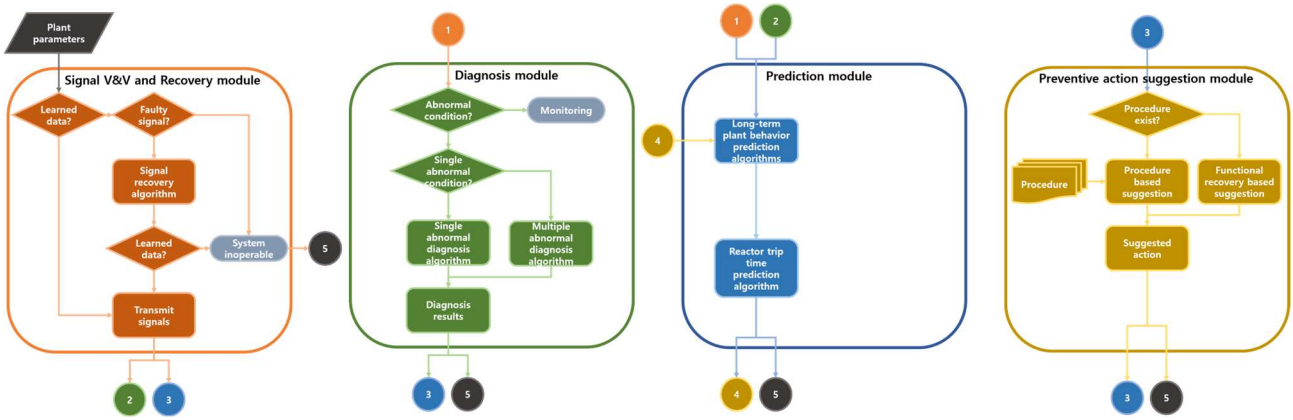

Figure 1. AI-based Operator Support System

The diagnosis module aims to accurately diagnose the condition of the power plant while achieving data-independent characteristics. To ensure the successful deployment of the proposed system in an actual NPP, it is essential to establish robustness against data variations between simulator-derived data and real-plant data. The knowledge distillation technique is being considered to achieve data-independent characteristics, and its viability is currently being assessed by applying the training technique across diverse nuclear power plant simulators.

The progress prediction module consists of two algorithms: a long-term plant behavior prediction algorithm and a reactor trip time prediction algorithm. The former anticipates the prolonged behavior of the plant based on instrumentation data and current status diagnostic results, while the latter estimates the remaining reactor trip time derived from the long-term behavior prediction results. These algorithms provide operators with

insights into the plant's anticipated behavior and the actionable time remaining, facilitating more informed decision-making.

The preventive action suggestion module recommends optimal procedures for operators during abnormal conditions. For single abnormal events, the system utilizes supervised learning based on Abnormal Operating Procedures (AOPs) to promptly propose courses of action. For multiple abnormalities where no standard procedures are prescribed, the module suggests proper actions emphasizing functional recovery, employing reinforcement learning with the nuclear power plant as the learning environment.

3. POETNTIAL ATTACKS ON OSS AND MITIGATION METHODS

We focus on evaluating potential attacks and threats to the Diagnosis module within the AI-based Operator Support System (OSS). Attacks on deep learning systems can be categorized into four main types: Poisoning, which involves contaminating the dataset to induce erroneous inferences; Evasion, which targets the learning process; Extraction, which involves extracting the model to analyze vulnerabilities; and Inference attacks.

Generally, the attacks on neural network can be classified into white-box attack and black-box attack. White-box attacks have full access to the target model, including its architecture, weights, and parameters. Attackers can exploit this information to craft adversarial inputs by calculating precise gradients that maximize the model's error. Techniques like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are commonly used in white-box attacks, allowing attackers to generate highly effective adversarial examples due to their complete knowledge of the model.

Black-box attacks, on the other hand, operate without any direct knowledge of the target model's internal structure or parameters. Instead, attackers can only observe the inputs and outputs of the model, such as predictions or confidence scores. This scenario is more realistic in practical applications since most deployed systems do not expose their internal workings. Black-box attacks rely on query-based methods, surrogate models, or transfer attacks, where adversarial examples generated from one model are used against another.

The lack of internal information makes these attacks more challenging but also more applicable to real-world scenarios. As an attack method, White-box attacks are excluded. Because the White-box attack methods require extensive knowledge of the developed model and the data used, are considered highly unlikely in the context of NPP software. Therefore, this study primarily focuses on black-box attacks, specifically Poisoning and Evasion attacks.

Potential way the OSS could jeopardize the power plant's safety through the Human-Machine Interface (HMI) is by providing erroneous diagnostic information with high confidence levels, which may confuse the operators and lead to human errors. However, excessive data manipulation may raise suspicion among the operators. Therefore, this study aims to assess the feasibility of inducing misdiagnosis using minimal data attacks, such as Evasion attacks.

For Evasion attacks, Threshold attacks and Decision-based/Boundary-based attacks[7] will be utilized. Threshold attacks focus on manipulating the input data to cross the decision boundary and alter the model's output. Given an input sample x and a target class y_t, the attacker aims to find a perturbed sample x' such that $f(x') = y_t$, where f is the targeted model. The optimization problem for threshold attacks can be formulated as:

$$\min_{x'} d(x, x') \text{ subject to } f(x') = y_t \tag{1}$$

where d is a distance metric, such as Euclidean distance or L_p norm, measuring the distortion between the original and perturbed samples.
Decision-based/Boundary-based attacks, on the other hand, iteratively modify the input data to find the minimal perturbation required to change the model's decision. The attacker starts with a large perturbation and gradually reduces it while ensuring that the perturbed sample crosses the decision boundary. The optimization problem for decision-based attacks can be expressed as:

$$\min_{x'} d(x, x') \text{ subject to } \mathrm{f}(x') \neq f(x) \tag{2}$$

Poisoning attacks, to be effective, must be able to pass the Verification and Validation (V&V) process in the supply chain. Hidden trigger backdoor attacks will be employed to evaluate the efficacy of such attacks. In hidden trigger backdoor attacks[8], the attacker injects a specific trigger pattern Δ into a subset of the training data D_p, causing the model to misclassify samples containing the trigger while maintaining high accuracy on clean data. The poisoned dataset D_p is generated by:

$$D_p = \{((x + \Delta, y_t)) \mid (x, y) \in D_s\} \quad (3)$$

where D_s is a subset of the clean training data, and y_t is the target class chosen by the attacker.
To assess the effectiveness of mitigation strategies, the study plans to incorporate General Adversarial Training[9], Mix-up[10], and Label Smoothing techniques[11] into the model design and observe their impact on the model's resilience against attacks. General Adversarial Training involves training the model on a combination of clean and adversarial examples to improve its robustness. The objective function for adversarial training is:

$$min_\theta \, E_{\{(x,y) \sim D\}} \left[max_{\{\delta \in S\}} \, L(\theta, x + \delta, y) \right] \quad (4)$$

where θ represents the model parameters, D is the training dataset, L is the loss function, and S is the set of allowed perturbations.
Mix-up is a data augmentation technique that combines pairs of input data and their corresponding labels to create new training samples. Given two input samples (x_i, y_i) and (x_j, y_j), mix-up generates a new sample (x', y') as follows:

$$x' = \lambda x_i + (1 - \lambda) x_j \text{ and } y' = \lambda y_i + (1 - \lambda) y_j \quad (5)$$

where $\lambda \in [0, 1]$ is a mixing coefficient sampled from a Beta distribution.
Label Smoothing is a regularization technique that replaces hard target labels with a weighted average of the original label and a uniform distribution, making the model less sensitive to small input perturbations. The smoothed label y_s is computed as:

$$y_s = (1 - \alpha) y + \alpha / K \quad (6)$$

where y is the original one-hot encoded label, α is the smoothing factor, and K is the number of classes.
By conducting experiments with these attack scenarios and mitigation strategies, the study aims to provide insights into the vulnerabilities of the AI-based OSS and propose effective countermeasures to enhance its security and reliability in the context of nuclear power plant operations. The mathematical formulations presented above offer a rigorous foundation for understanding and implementing these attacks and mitigation methods, enabling a comprehensive analysis of the AI system's robustness against potential threats.

Also, explainable AI methodologies have been employed to evaluate the critical variables and their respective domains that significantly contribute to the power plant state diagnosis. The feature map, generated based on these important variables and their domains, provides insights into the key factors influencing the diagnostic process. The Figure 2 shows Layer-wise relevance propagation method based analysis results focusing on three distinct emergency scenarios: Loss of Coolant Accident (LOCA), Steam Generator Tube Rupture (SGTR), and Excess Steam Demand Event (ESDE). The figure presents three 3D surface plots, each representing the XAI analysis of these scenarios, visualizing the relationship between Time, Var (variables), and Relevance score.

The plots reveal unique patterns of relevance scores for each emergency scenario:
LOCA (left plot): Shows a relatively uniform distribution of relevance scores with some localized peaks, suggesting a balanced consideration of multiple variables over time.
SGTR (middle plot): Displays more pronounced variations in relevance scores, indicating a dynamic decision-making process where certain variables become significantly more important at specific time points.
ESDE (right plot): Features a prominent peak in relevance scores, representing a critical variable or time point that strongly influences the AI's diagnosis in this scenario.

Critically, this analysis highlights that variables with high relevance scores play a crucial role in the neural network-based operator diagnosis support system. Attacking or manipulating these high-relevance variables could potentially compromise the system's ability to accurately diagnose and respond to emergency situations. For instance, in the ESDE scenario, the prominent peak suggests a particularly vulnerable point where targeted interference could significantly impact the AI's diagnostic capabilities.

This vulnerability underscores the importance of robust security measures and fail-safe mechanisms in AI-driven nuclear power plant management systems. It also emphasizes the need for continuous monitoring and validation of the AI's decision-making process, particularly focusing on high-relevance variables identified through XAI analysis.

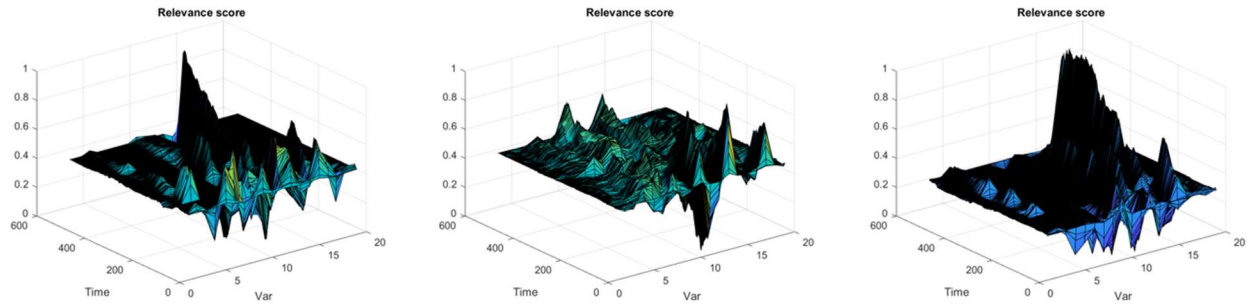

Figure 2. Pilot Experiment Results for Critical Data Asset Finding

4. CONCLUSION AND FURTHER WORK

The integration of AI in nuclear power plant operator support systems presents both opportunities and challenges. While AI-based systems have the potential to enhance the efficiency and safety of plant operations, the increasing reliance on data-driven approaches expands the attack surface and introduces new cybersecurity risks. This study has provided a preliminary assessment of the potential threats and attack methods that AI-based nuclear power plant diagnostic support systems may face, focusing on the Diagnosis module.

Through the examination of Evasion attacks, such as Threshold and Decision-based/Boundary-based attacks, and Poisoning attacks using Hidden Trigger Backdoors, the study has highlighted the feasibility of inducing misdiagnosis and compromising the integrity of the AI system. The employment of Explainable AI methodologies has facilitated the identification of critical variables and their domains, providing insights into the key factors influencing the diagnostic process.

To mitigate these risks, the study has explored the effectiveness of mitigation strategies, including General Adversarial Training, Mix-up, and Label Smoothing. These techniques aim to improve the robustness of the AI system against potential threats by training the model on a combination of clean and adversarial examples, augmenting the training data, and reducing the model's sensitivity to small input perturbations.

The findings of this study have significant implications for the cybersecurity for AI-based software. By elucidating the unique challenges and vulnerabilities associated with AI-based systems in the nuclear industry, this research contributes to the development of robust cybersecurity frameworks and best practices. The insights gained from this study will inform the ongoing efforts to ensure the safe and reliable operation of nuclear power plants in the face of evolving cyber threats.

However, further research is necessary to fully understand and address the cybersecurity risks associated with the adoption of AI methodologies in nuclear power plant operator support systems. Future work should focus on refining the attack scenarios, exploring additional mitigation strategies, and validating the findings through real-world experiments and case studies.

In conclusion, this study represents a crucial step towards understanding and addressing the cybersecurity risks associated with AI-based systems in the nuclear industry. By providing a preliminary assessment of potential threats and mitigation methods, the research laid the foundation for the development of effective risk mitigation

strategies and contributed to the overall safety and security of nuclear power plant operations in the digital and AI era.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2022-00144150)

References

[1] Bae, J., Park, J.W., Lee, S.J., Limit surface/states searching algorithm with a deep neural network and Monte Carlo dropout for nuclear power plant safety assessment. Applied Soft Computing 124, 109007. 2022.
[2] Chae, Y.H., Lee, C., Han, S.M., Seong, P.H., 2022. Graph neural network based multiple accident diagnosis in nuclear power plants: Data optimization to represent the system configuration. Nuclear Engineering and Technology 54, 2859–2870. 2022.
[3] Lee, G., Lee, S.J., Lee, C.,. A convolutional neural network model for abnormality diagnosis in a nuclear power plant. Applied Soft Computing 99, 106874, 2021.
[4] Ridluan, A., Manic, M., Tokuhiro, A., EBaLM-THP – A neural network thermohydraulic prediction model of advanced nuclear system components. Nuclear Engineering and Design 239, 308–319. 2009.
[5] Saeed, A., Rashid, A., Development of Core Monitoring System for a Nuclear Power Plant using Artificial Neural Network Technique. Annals of Nuclear Energy 144, 107513. 2020.
[6] Lee, H.-J., Lee, D., Kim, J., Event diagnosis method for a nuclear power plant using meta-learning. Nuclear Engineering and Technology S1738573324000056. 2024
[7] Brendel, W., Rauber, J. and Bethge, M., Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. 2017.
[8] Saha, A., Subramanya, A. and Pirsiavash, H., Hidden trigger backdoor attacks. In Proceedings of the AAAI conference on artificial intelligence Vol. 34, No. 07, pp. 11957-11965. 2020.
[9] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R.,. Intriguing properties of neural networks. 2013.
[10] Zhang, H., Cisse, M., Dauphin, Y.N. and Lopez-Paz, D., mixup: Beyond empirical risk minimization. 2017.
[11] Shafahi, A., Najibi, M., Xu, Z., Dickerson, J., Davis, L.S. and Goldstein, T., April. Universal adversarial training. In Proceedings of the AAAI Conference on Artificial Intelligence , Vol. 34, No. 04, pp. 5636-5643. 2020.