Application of Explainable Artificial Intelligence (XAI) for Nuclear Power Plant Operations: Classification and Practical Cases

Daehyung Lee, Yongju Cho, Chungwon Seo, Byung Jo Kim [a*]
[a] *KEPCO E&C, Nuclear Technology Research Dept., 269 Hyeoksin-ro, Gimcheon-si, 39660*
**Corresponding author: bjokim@kepco-enc.com*

Keywords : explainable AI(XAI), operation support, machine learning, SHAP, LIME

1. Introduction

Nuclear power plants (NPPs) are critical infrastructures that demand the highest levels of safety and reliability. To detect potential risks and respond swiftly, various artificial intelligence (AI) techniques have been introduced in nuclear plant operations. Recently, big data analytics and deep learning-based approaches have improved maintenance efficiency and accident prevention, which have sparked significant interest in applying AI to the nuclear industry.

However, complex AI models such as deep neural networks often behave as "black boxes," making it challenging to understand or trust their decision-making processes. In a domain where transparent safety justifications are essential, this opacity poses a critical challenge. As a solution, an approach called Explainable Artificial Intelligence (XAI) has emerged. XAI aims to provide interpretable reasons for AI outputs, thereby enhancing the transparency and trust in AI systems. This paper outlines the concept and classification of XAI, reviews its key applications in nuclear power plant operations, and discusses future directions.

2. Concept and Classification of XAI

2.1 Concept of XAI

Explainable Artificial Intelligence (XAI) encompasses a set of techniques and methods that enable humans to understand how and why AI models arrive at specific outputs. While traditional statistical or linear models were relatively interpretable due to their simpler structures, the advent of deep learning—with its numerous parameters and layers—has significantly reduced model interpretability. Consequently, XAI research has focused on either designing inherently interpretable models (ante-hoc approaches) or providing post-hoc explanations to clarify the reasoning behind a model's predictions.

2.2 Classification of XAI

2.2.1 Classification by Model Development Stage

- Ante-hoc Approach: Models are designed to be inherently interpretable from the outset. Decision trees and rule-based systems are typical examples, as their transparent structures allow immediate explanation without additional processing after training.

- Post-hoc Approach: Explanations are generated externally after the model has been trained. Techniques such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations) are widely used post-hoc methods applicable to complex black-box models like deep neural networks and random forests.

2.2.2 Classification by Explanation Scope

- Local Methods: These provide explanations for individual predictions or small regions of the input space. For example, LIME and SHAP clarify why a specific input leads to a particular outcome.

- Global Methods: These aim to describe the overall reasoning structure of the model, including the importance of each feature across the entire dataset, helping domain experts understand the broader patterns the model has learned.

2.2.3 Classification by Explanation Format

- Feature Importance: Methods in this category rank or quantify the contribution of each input variable to the final prediction. LIME and SHAP excel at producing such importance scores.

- Visual Explanations: Tools such as Gradient-weighted Class Activation Mapping (Grad-CAM) highlight the most relevant regions in an image or key segments in a time-series signal that influence the model's decision.

- Rule/Knowledge-based Explanations: These methods extract if-then rules or domain knowledge from the model, allowing users to understand the logical inference path used by the AI system.

These classification schemes enable practitioners to select XAI methods that are most suitable for their domain and objectives. In high-reliability sectors such as nuclear energy, where both interpretability and accuracy

are critical, a multifaceted approach to XAI is often necessary.

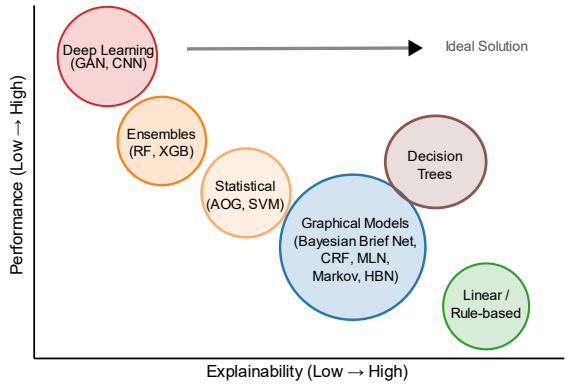

Fig. 1. Trade-off between performance and explainability of widely used AI models. The ideal solution should have both high explainability and high performance. [1]

3. XAI Applications in Nuclear Power Plant Operations

3.1 Anomaly Detection and Failure Prediction

Nuclear power plants deploy thousands of sensors that collect operating data in real time. By applying deep learning models, early detection of abnormal patterns is possible, thereby reducing the risk of severe accidents. For instance, Park et al. (2022) [2] developed an anomaly detection model based on a GRU-AE combined with LightGBM and employed the SHAP explanation method to visualize the key contributing variables. This visualization helps operators identify which sensor readings have the most influence on the detection of anomalies.

In addition to GRU-AE-based approaches, other recurrent models such as Bi-directional LSTM (Bi-LSTM) have been utilized to capture temporal patterns in nuclear plant operation data. For example, a Bi-LSTM-based anomaly detection model was applied to identify early signs of faults in simulator-generated data, and feature attribution was visualized using SHAP to explain which variables triggered anomaly detection decisions [3].

3.2 Operator Decision Support

During complex or emergency situations, AI-generated recommendations can support operator decisions; however, unexplained decisions may be met with skepticism from experienced staff. By providing case-specific explanations using local methods (e.g., LIME or SHAP) and supplementing them with global or visual explanations (e.g., Grad-CAM highlighting critical data segments), operators gain a clearer understanding of the AI's rationale, thereby enhancing their situation awareness (SA) and trust in the system.

An AI-guided reasoning-based operator support system has been developed using Answer Set Programming (ASP) to represent nuclear power plant knowledge through logic rules. This system assists operators by providing fault identification, scenario analysis, and control options, thereby improving decision-making processes during complex events [4].

3.3 Predictive Maintenance and Safety Assurance

Predictive maintenance (PdM) leverages machine learning to forecast component failures in advance, which allows for optimized repair scheduling [5]. Integrating XAI methods such as SHAP clarifies which indicators drive the failure predictions, enabling engineers to prioritize maintenance tasks more effectively. This not only reduces operational costs and improves safety but also fosters trust among regulatory bodies by transparently communicating the underlying reasoning behind the predictions.

Amin et al. (2022) [6] proposed a novel approach to integrate explainable AI into prognostics and health management (PHM) systems for civil nuclear power plants. Using SHAP, they explained predictions from black-box models related to asset degradation. To improve usability among non-ML experts such as plant operators and regulators, they developed algorithms that translate SHAP visualizations into human-readable text. This translation improves transparency and fosters trust in AI-driven maintenance decisions, supporting safer and more accountable deployment of predictive models in the nuclear industry.

Table I: Categorization of XAI Methods

XAI Method	**Stage**	**Scope**	**Format**
	Key Features		
Decision Tree	Ante-hoc	Global	Rules/FI*
	Transparent; explicit if-then rules		
Rule-based Sys.	Ante-hoc	Global	Rules
	Domain-driven; clear logical inference		
LIME	Post-hoc	Local	FI
	Local linear model approximation		
SHAP	Post-hoc	Local/Global	FI
	Shapley values quantify feature impact		
Grad-CAM	Post-hoc	Local	Visual
	Highlights key regions using gradients		

*FI: Feature Importance

4. Discussion and Future Work

4.1 Reliability Verification and Regulatory Compliance

Even with XAI, verifying the accuracy and robustness of explanations remains essential in high-reliability sectors such as nuclear power. Post-hoc explanations

might not fully reflect the internal mechanisms of complex models, potentially leading to misleading conclusions. Thus, it is necessary to cross-validate explanations from multiple methods and to contrast them with physics-based knowledge (e.g., thermal-hydraulics, neutron physics) to ensure that the explanations align with the actual system behavior. Regulatory organizations such as the NRC and IAEA emphasize that AI systems used in safety-critical applications must be both interpretable and verifiable [7].

4.2 Technical Advancements

To simultaneously achieve high interpretability and accuracy, researchers are exploring advanced XAI approaches, including causal reasoning, counterfactual analysis, and simulator-based data augmentation. In areas like nuclear power operations, where real abnormal data are rare, simulators and augmented datasets can be used to generate extreme scenarios for testing both AI models and their corresponding XAI techniques.

4.2.1 Integration of PINN and SciML

Hybrid models like Physics-Informed Neural Networks (PINNs) and Scientific Machine Learning (SciML) combine data-driven approaches with physical laws to improve reliability in extreme or unseen conditions. PINNs, for instance, embed partial differential equations (PDEs) into the learning process, ensuring predictions align with physical principles. In nuclear power, PINNs can enhance thermal-hydraulics and core physics analysis, providing more reliable predictions in data-scarce situations.

4.2.2 Ensuring Robustness Across Diverse Scenarios

For Explainable AI (XAI) to be effective, the model must be robust. If the model is vulnerable to adversarial inputs or performs poorly in untrained scenarios, the explanations lose value. Using simulators or data augmentation to generate diverse scenarios, especially for rare events, and cross-checking with established physical knowledge can help ensure consistent performance and reliable explanations.

4.2.3 Regulatory and Legal Considerations

AI systems in the nuclear industry must meet regulatory standards and provide interpretable results. Bodies like the NRC require explainable AI in safety-critical areas. As there is no universal standard for the level of explanation needed, industry collaboration with regulators is essential to create clear guidelines. A phased introduction, starting with non-safety-critical applications, can help build trust and validate performance.

5. Conclusions

In nuclear power plant operations, Explainable Artificial Intelligence (XAI) plays a vital role in ensuring safety and establishing trust in AI systems. This paper has reviewed the fundamental concepts and classifications of XAI, illustrated its application in anomaly detection, operator decision support, and predictive maintenance, and discussed future directions. The integration of physics-based models such as PINN and SciML with XAI, along with user-friendly interface designs and regulatory compliance, is anticipated to pave the way for robust, high-reliability AI systems in the nuclear industry. Such advances will ultimately enhance operational safety as well as overall trust in AI among industry stakeholders and the general public.

Acknowledgment

This work was supported by Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government (MOTIE) (No. 20224B10100130, Development of operational state simulator for operating nuclear power plant and commercialization technology for artificial intelligence decision-making support system to prevent human error in accident operation)

REFERENCES

[1] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, Information Fusion, Vol.77, pp.29–52, 2022.
[2] J. H. Park, et al., A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP, Nuclear Engineering and Technology, Vol.54, pp.1271-1287, 2022.
[3] A. Chaudhary, J. Han, S. Kim, A. Kim, and S. Choi, Anomaly Detection and Analysis in Nuclear Power Plants, Electronics, vol. 13, no. 22, p. 4428, Nov. 2024.
[4] B. Hanna, T. C. Son, and N. Dinh, AI-Guided Reasoning-Based Operator Support System for the Nuclear Power Plant Management," Annals of Nuclear Energy, vol. 154, p. 108079, May 2021.
[5] C. M. Walker, et al., Demonstration and Evaluation of Explainable and Trustworthy Predictive Technology for Condition-based Maintenance, Idaho National Laboratory (INL), Idaho Falls, ID, 2024.
[6] O. Amin, B. Brown, B. Stephen, and S. McArthur, A Case-study Led Investigation of Explainable AI (XAI) to Support Deployment of Prognostics in industry, 2022.
[7] NRC, Considerations for Developing AI Systems in Nuclear Power Plants, NRC Report ML24241A252, 2024.
[8] A. Hall, et al., Human-centered and explainable artificial intelligence in nuclear operations, Proceedings of NPIC&HMIT, 2023.
[9] S. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, Nov. 25, 2017.