Operator-Centric Visualization of Explainable Diagnosis for Abnormal Situations Using Multilevel Flow Modeling

Ji Hyeon Shin[1,*], Seo Ryong Koo[1], Seung Jun Lee[2]

[1] *Korea Atomic Energy Research Institute, 111, Daedeok-daero 989 beon-gil, Yuseong-gu, Daejeon, Republic of Korea*
[2] *Ulsan National Institute of Science and Technology, 50, UNIST-gil, Ulsan 44919, Republic of Korea*
[*] *jhshin0127@kaeri.re.kr*

ABSTRACT

In nuclear power plants, the rapid and accurate diagnosis of abnormal states is crucial to ensure safe operations and mitigate the risk of accidents. This study introduces an operator-centered approach for explainable diagnostics that emphasizes human understandability, making complex information accessible to operators when it matters most. By utilizing Multilevel Flow Modeling (MFM), this method significantly enhances the interpretability of diagnostic results, enabling operators to grasp underlying system behaviors and causal relationships during abnormal states. By integrating explainable artificial intelligence (XAI) techniques with MFM, this approach provides visual, intuitive explanations that bridge the gap between advanced AI outputs and operator decision-making needs. Through this, operators are better equipped to understand diagnostic insights, fostering a sense of trust and confidence in the model's conclusions, even in challenging situations.
The proposed approach was tested using simulation data from various abnormal scenarios in nuclear power plants, validating model performance with a primary focus on usability and operator comprehension. This operator-centric design is expected to lead to improved clarity in understanding diagnostic results, which in turn should increase operator confidence when responding to critical situations. By prioritizing human-centered explainable AI applications, this work seeks to support safer, more effective nuclear plant operations. Ultimately, it aims to create diagnostic systems that operators can rely on fully during high-stakes decision-making, thus promoting more resilient plant management in the face of potential risks.

Keywords: Nuclear power plant, Abnormal state diagnosis, Explainable artificial intelligence, Multilevel flow modeling, Visualization

I. INTRODUCTION

In nuclear power plants (NPPs), abnormal conditions occurring in individual components can influence surrounding monitoring parameters such as flow rate, temperature, and pressure. Operators play an important role in diagnosing such situations as specific abnormal events based on alarms and symptoms triggered by changes in these parameters. To support operators in this diagnostic task, various classification models based on artificial neural networks (ANNs) have been developed. These models also perform diagnoses by learning the patterns of parameter changes associated with abnormal events in the input data. However, due to the inherent black-box nature of neural networks, it is difficult to understand how these models make their decisions. For operators to trust and effectively utilize the model outputs, it is required that the diagnostic results be clearly explained in terms of which parameter changes, which symptoms, led to a specific diagnosis.

ANNs can address part of their black-box problem by using appropriate explanation methods, enabling them to provide users with explanations for their outputs. However, the explanations generated by advanced artificial intelligence (AI) models do not necessarily align with the reasoning processes traditionally used by human operators. This mismatch can lead to a gap in understanding, thereby limiting the practical usability of such models in operational contexts. Accordingly, an effective approach is needed to enhance operator understandability. To address this challenge, the present study aims to explain neural network–based abnormal state diagnosis models using explainable AI (XAI) techniques and to present the resulting explanations in a manner that is meaningful to human operators. Specifically, the study compares and analyzes explanations for an abnormal event: (1) explanations derived from the diagnostic model via XAI, (2) symptom-based explanations traditionally used by operators, and (3) causal explanations based on system structure and physical flow using Multilevel Flow Modeling (MFM). Through this comparison, we identify parameters that are commonly emphasized across the different

explanations as well as those that are not, gaining insight into which aspects are most understandable from the operator's perspective. Based on this comparison, we propose strategies for bridging the gap between AI-generated diagnostic reasoning and operator understandability. To clarify the objective of this study, we pose the research question and insight that how operators can make use of the explanations provided by XAI techniques. Ultimately, this research aims to support the development of diagnostic systems that operators can trust and use effectively, even under abnormal conditions in NPPs.

II. EXPLAINABLE DIAGNOSIS MODEL FOR ABNORMAL EVENT

In this study, we purpose to identify the symptoms associated with specific abnormal events by leveraging both XAI techniques and MFM. The contributed symptoms corresponding to the abnormal events diagnosed by the neural network model can be analyzed using appropriate XAI methods. Meanwhile, MFM allows for the understanding of physically connected flow resulting from abnormalities in specific components or functions. The following sections describe these two methodologies as applied in this study.

II.A. Explainable Artificial Intelligence

XAI refers to techniques that provide human-understandable explanations for the decisions made by AI models. In other words, XAI helps users understand why a model produced a certain output and which input features contributed most significantly to that result. By providing this interpretability, XAI enhances users' ability to understand and trust the model's decisions, which is particularly important in safety-prioritized domains such as NPP operation. Among them, post-hoc explanation methods are commonly used to explain the outputs of already-trained models when given specific input data. These methods aim to explain model behavior without altering the internal structure of the model. Among the various post-hoc approaches, the XAI technique introduced below has been employed in previous studies to explain the outputs of neural network-based diagnostic models in the context of NPP condition monitoring.

II.A.1. Deep Explainer SHAP

Shapley Additive exPlanations (SHAP) is a method that explains model predictions by decomposing them into the contributions of individual features [1]. A Shapley value, originally derived from cooperative game theory, represents the average marginal contribution of a feature across all possible combinations of features, offering a fair way to quantify each feature's importance in a prediction. Among its variants, Deep Explainer SHAP is specifically designed for application to neural networks. It approximates Shapley values by leveraging the structural characteristics of deep learning models. Specifically, it extends the principles of the Deep Learning Important FeaTures (DeepLIFT) algorithm, calculating feature contributions through activation differences between input data and a baseline across the model's layers [2]. This integration allows it to produce explanations more efficiently than sampling-based SHAP methods.

II.B. Multilevel Flow Modeling

Multilevel Flow Modeling (MFM) is a framework utilized to represent the structure and behavior of complex industrial systems—such as NPPs—by capturing their functional objectives, means-end relationships, and causal dependencies within mass flow systems and energy flow systems [3, 4]. It facilitates qualitative assessment of system performance by reflecting the underlying physical flow mechanisms. In this study, the modeling and analysis of MFM were conducted using MFMSuite, a software platform originally developed by the Technical University of Denmark, in conjunction with a graphical editor provided by IFE Harden.

III. CASE STUDY

III.A. Experimental Setup

This study provides diagnostic information that is understandable and trustworthy to human operators by integrating the previously introduced methods. The experimental procedure consists of the following sequential steps, through which the analysis and discussion are conducted based on the resulting observations:

(1) The abnormal event diagnosis model training and application to a scenario for case study: A neural network-based classification model is trained to diagnose various abnormal conditions in NPPs. The trained model is then applied to case scenarios to generate diagnostic outputs.

(2) Application of the XAI method to the model's diagnostic results: To enhance interpretability, the XAI technique are applied to the output of the trained model. This step aims to identify which input features most significantly influenced the diagnosis of the case abnormal scenario.
(3) Comparison with actual alarms and symptoms: The diagnostic outcomes and their explanations are compared against actual alarm signals and symptoms from the scenario, enabling assessment of how well the AI model's reasoning aligns with human-understandable indicators.
(4) Analysis of MFM-based consequent trees for the diagnosed event: For the event diagnosed by the model, MFM is used to generate and examine the corresponding consequent tree, which captures the causal propagation of functional disruptions throughout the system.
(5) Visualization of diagnostic explanations: The results of the XAI analysis are visualized with highlighting the key system by considering physical flow in step (4) [5].

III.A.1. Abnormal Event Diagnosis Model

In this case study, the diagnosis model was trained using abnormal scenario data obtained from the 3KEYMASTER NPP simulator [6, 7]. The model was trained on 23,375 data to classify 15 kinds of abnormal events. The Deep Explainer SHAP technique is model-agnostic and can be applied regardless of the model structure. Therefore, the choice of model does not affect the applicability of this explanation method. The simple convolutional neural network (CNN) was adopted in this study. The structure of the abnormal event diagnosis model used in this case study is as follows.

TABLE I. The Abnormal Event Diagnosis Model Structure

Layer name	Layer type	Output shape
Conv1d	Conv1D	(None, 391, 16)
Conv1d_1	Conv1D	(None, 391, 32)
Max_pooling1d	MaxPooling1D	(None, 195, 32)
Conv1d_2	Conv1D	(None, 195, 64)
Max_pooling1d_1	MaxPooling1D	(None, 97, 64)
Flatten	Flatten	(None, 6208)
Dense	Dense	(None, 16)
activation	Activation	(None, 16)

III.A.2. Multilevel Flow Modeling for Abnormal Events

In this case study, we aim to understand the physical flow associated with each abnormal event by analyzing the corresponding consequent trees. To facilitate this, an MFM representation was developed based on the system structure of the 3KEYMASTER NPP simulator and 15 abnormal events that were used to training data for the diagnostic model. As summarized in the table below, the constructed MFM includes a total of six systems, comprising three mass flow systems and three energy flow systems.

TABLE II. Systems and Functions including in the MFM Model

Type	System	Included components about each function
Mass flow system	Reactor coolant system and chemical volume control system	Cold-leg, reactor pressure vessel, hot-leg, reactor heat remover, pressurizer, pressurizer relief tank, steam generator U-tube, letdown heat exchanger, letdown demineralizer, volume control tank, charging pump, regenerator, reactor coolant pump
	Steam generator and secondary system	Steam generator, main steam system, turbine bypass valve, condensate storage tank, turbine, condenser, condensate pump, low-pressure feedwater heater, main feedwater pump, high-pressure feedwater heater, moisture separator reheater, steam generator blowdown
	Containment spray system	Refueling water storage tank, containment pump, containment spray
Energy flow system	Reactor coolant system heat remover	Fuel, reactor pressure vessel, pressurizer, pressurizer relief tank, steam generator U-tube, steam generator, main steam system, turbine, turbine bypass valve, condenser, circulating water system

	Primary component cooling system	Reactor heat removal, letdown heat exchanger, component cooling water system, essential service water system
	Electrical system	Turbine, main generator protection, switchyard, high-voltage 13.8kV, 1E medium voltage 4.16kV, diesel generators control, medium voltage 4.16kV

III.B. Case Study with Pressurizer Spray Valve Positioner Failure

In Section III.A, we compare the explanations derived from the CNN model by Deep Explainer SHAP and the MFM-based model in the context of abnormal event occurrences. We then discuss whether the explanations generated by the neural network are comprehensible to plant operators. The consistency between the AI-derived interpretation and the physically grounded reasoning provided by the MFM model is examined. Based on this comparison, we identify which information should be conveyed to operators and which can be excluded, and propose a visualization method to effectively present the finalized diagnostic information.

III.B.1. Explanation Results

In this case study, the abnormal scenario for case study is a Pressurizer Spray Valve Position Failure. This event was selected as the case study because, although the event itself is relatively simple, it affects several key parameters such as pressurizer pressure and water level. These changes can propagate to other components, making it useful for analyzing system-wide behavior. While not one of frequent failures, it is representative of control component malfunctions that can lead to cascading effects. When this event occurs, the primary symptom is a noticeable decrease in pressurizer pressure. In addition, secondary symptoms such as the temperature of the pressurizer spray line and the operation of the pressurizer heaters are also affected. Based on these peripheral symptoms and corresponding alarms, plant operators are expected to identify the abnormal condition and follow appropriate tasks as prescribed in the operating procedure. The table below presents the explanation results from the abnormal event diagnosis model trained in this study by Deep Explainer SHAP, illustrating how the model recognizes and explains the pressurizer spray valve failure scenario.

TABLE III. The Explanation Result of Deep Explainer SHAP

Monitoring parameter	Location	Description	Feature importance
hmi_RCSLT459A_VALUE	Pressurizer	Pressurizer level	1
hmi_RCSLT459_VALUE	Pressurizer	Pressurizer level	0.7738
hmi_RCSTT463A_VALUE	Cold-leg	RCL-2B delta temperature	0.5072
hmi_NBXIT6_VALUE	Medium voltage 4.16kV	LC feeder breaker NBX209 current	0.4248
movBBHV8000B.avpVlvPos	Reactor coolant system	HV8000B valve position	0.3779
movBBHV8000A.avpVlvPos	Reactor coolant system	HV8000A valve position	0.3614
vmodEGTV0030.avpVpos	Component cooling water system	TV30 valve position	0.3536
hmi_NBXWT15_VALUE	Medium voltage 4.16kV	Engineered safety features transformer XNB02 power	0.3431
hmi_CVCTT130A_VALUE	Letdown heat exchanger	Letdown heat exchanger outlet temperature	0.3362
hmi_RCSLT502_VALUE	Steam generator	Steam generator 2 wide-range level	0.3354

The model diagnosed the event as a Pressurizer Spray Valve Position Failure with a high predicted value, and the Deep Explainer SHAP analysis identified pressurizer level as the most important parameter in the diagnosis. In addition, the model also utilized other key parameters such as the opening of surrounding valves near the pressurizer spray valve and the cold-leg temperature, indicating that it considered a range of contributed parameters in reaching its conclusion. It is worth noting that the NPP simulator used in this study is designed to simulate a generic pressurized water reactor, and thus may not fully capture the precise symptom progression observed in an actual plant. Nevertheless, the results reveal that the model relies on a set of parameters that differ from those typically used by human operators, highlighting a potential gap in understandability and trust.

We analyze the possible end consequence from the constructed MFM model's consequence tree that most closely aligns with the feature importance identified by the abnormal event diagnosis model. Figure 1 below illustrates the branch of this

possible end consequence, which includes the pressurizer, reactor coolant system, and charging system. While the MFM consequence tree does not capture all potential symptoms resulting from the pressurizer spray valve malfunction, it effectively represents the surrounding flow relationships involving the pressurizer. In this way, the MFM model conveys a flow-based understanding that aligns with the operator's understandable model and supports human-centered interpretability.

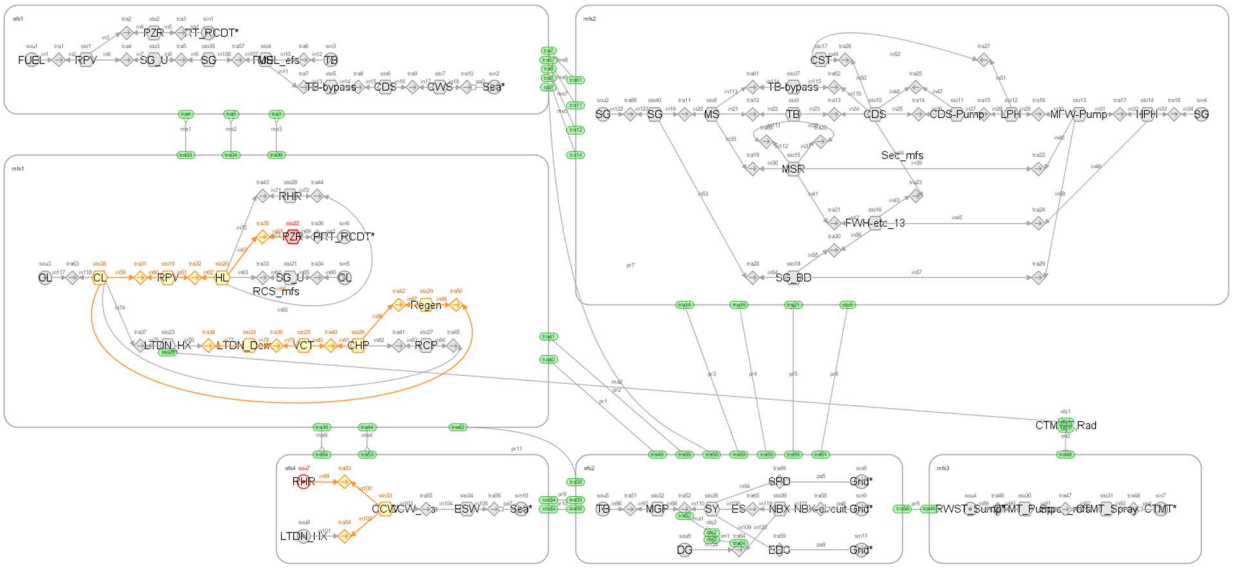

FIGURE 1. The Consequence Tree Result of the MFM Model

III.B.2. Explanation Visualization

In this study, only the explanation results from the Deep Explainer SHAP results that correspond to operator-understandable flow paths are visualized. Specifically, parameters that are not represented in the MFM model or do not align with the most relevant consequence tree branch were excluded from the visualization. The resulting explanation visualization, shown below, highlights components such as the pressurizer, reactor coolant system, and chemical and volume control system, while excluding parameter such as the LC feed breaker current, which are less relevant from a flow-based understandability perspective.

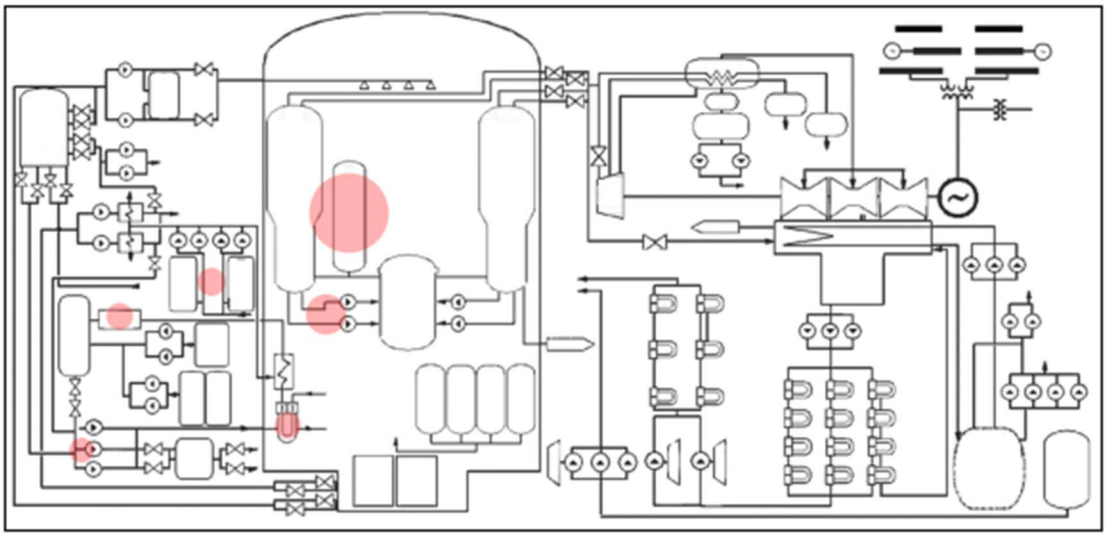

FIGURE 2. Explanation Visualization with Selected Information

IV. CONCLUSIONS

AI models have been recently researched to support operator tasks in NPP main control rooms. However, to ensure operator acceptance of AI-based diagnostic information, it is need to provide not only the results but also the reasoning in a form that is understandable to human operators. XAI techniques offer a way to explain model outputs by identifying key input features that influence the model's predictions. In the context of abnormal event diagnosis, several previous studies have sought to apply XAI to provide explainability alongside neural network-based decisions. Nevertheless, the parameters identified by XAI as contributing to a diagnosis may not always align with operators' understanding. Presenting such mismatched information can lead to confusion rather than clarity. To address this problem, this study examined the diagnostic reasoning of an CNN model through the perspective of operator understanding. Using Deep Explainer SHAP, we explained the model's decisions and analyzed the understandability of the resulting feature importance from an operator's perspective. Additionally, we applied MFM to derive the consequence branch for a specific abnormal event scenario, allowing us to cross-check the relevance of Deep Explainer SHAP-identified parameters with physical flow-based reasoning. By filtering out provided information that did not conform to operator-acceptable flow logic, we visualized a refined set of diagnostic information that better aligns with human understandability. This study contributes to propose the human-centered XAI by demonstrating how XAI-based explanations can be adapted to the operator understanding in the context of abnormal event diagnosis. While this study qualitatively explores explanation understandability, no user study has yet been conducted to quantitatively assess improvements in operator understanding and trust. Therefore, future work will involve user-centered experiments to assess how the proposed visualization method influences understandability, trust, and usability in practice. Furthermore, it is necessary to test various abnormal event cases, not just a single one, to examine the potential for generalization.

ACKNOWLEDGMENTS

This research was supported by the National Research Council of Science & Technology(NST) grant by the Korea government (MSIT) (No. GTL24031-000).

REFERENCES

[1] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
[2] Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." In International conference on machine learning, (2017): 3145-3153. PMlR.
[3] Lind, Morten. "An introduction to multilevel flow modeling." Nuclear safety and simulation 2, no. 1 (2011): 22-32.
[4] Kang, Jung Sung, and Seung Jun Lee. "Concept of an intelligent operator support system for initial emergency responses in nuclear power plants." Nuclear Engineering and Technology 54, no. 7 (2022): 2453-2466.
[5] Shin, Ji Hyeon, Jung Sung Kang, Jae Min Kim, and Seung Jun Lee. "Concept of understandable diagnostic cause visualization with explainable AI and multilevel flow modeling." Nuclear Engineering and Technology 57, no. 8 (2025): 103589.
[6] F. Western Service Corporation, MD, USA, 3KEYMASTER Simulator (2013)
[7] Lee, Gyumin, Seung Jun Lee, and Changyong Lee. "A convolutional neural network model for abnormality diagnosis in a nuclear power plant." Applied Soft Computing 99 (2021): 106874.