Nuclear Engineering and Technology 57 (2025) 103589

Contents lists available at ScienceDirect

Nuclear Engineering and Technology

journal homepage: www.elsevier.com/locate/net

Original Article

Concept of understandable diagnostic cause visualization with explainable AI and multilevel flow modeling

Ji Hyeon Shin [a], Jung Sung Kang [a], Jae Min Kim [b], Seung Jun Lee [a,*]

[a] Department of Nuclear Engineering, Ulsan National Institute of Science and Technology, 50, UNIST-gil, Ulsan, 44919, Republic of Korea
[b] Korea Atomic Energy Research Institute, Yuseong-gu, Daejeon, 34057, Republic of Korea

ARTICLE INFO

Keywords:
Nuclear power plants
Abnormal event diagnosis
Feature rearrangement
Explainable artificial intelligence
Multilevel flow modeling

ABSTRACT

In nuclear power plants, operators can face cognitive workloads when diagnosing abnormal events due to the need to monitor numerous parameters and consider hundreds of potential scenarios. Artificial intelligence technologies have been proposed to support this process by providing diagnostic results; however, their lack of transparency can lead to out-of-the-loop unfamiliarity and distrust, hindering effective decision-making. To address these challenges, this study introduces a novel concept to enhance the understandability and trustworthiness of diagnostic support systems through Explainable Artificial Intelligence (XAI). The first method in the proposed concept rearranges monitoring parameters based on system structures to reflect parameter relationships. The second method refines explanations from XAI using Multilevel Flow Modeling (MFM) to ensure consistency with physical flow, and it visualizes diagnostic cause components on a plant map. By filtering out incomprehensible information and visualizing intuitive diagnostic causes, the system enables operators to identify expected causes of diagnostic results directly on the NPP map at the component or system level. This approach provides explainable and comprehensible support information, fostering trust in the system and improving diagnostic efficiency in abnormal situations.

1. Introduction

When an abnormal situation occurs, the operator in a nuclear power plant (NPP) diagnoses an event that matches the situation based on the alarms and symptoms described in the abnormal operating procedure. After that, the operator can perform the given procedure according to the diagnosed abnormal event and alleviate the situation in the NPP. However, the process of monitoring numerous changing parameters and considering hundreds of potential events to diagnose a single matching event can impose a high-level workload on operators. Recently, to support operators in diagnosing abnormal events, artificial intelligence (AI) technologies have been studied. Kim et al. proposed a two-stage structure of Gated Recurrent Unit for operating procedures and subprocedures to classify abnormal events [1]. Lee et al. proposed a diagnosis algorithm that classifies abnormal scenarios and emergency scenarios based on Robust AI utilizing a feature extractor [2]. Dong et al. detected anomalies using a 1-dimensional Convolutional Neural Network (CNN) and a soft attention mechanism in an anomaly case dataset for a high-temperature gas-cooled reactor [3]. Lin et al. identified events with sensor faults using deep learning-based schemes [4]. However, if the model in the diagnostic support system only provides results, the operator will not be able to fully understand the situation of the plant with the information provided. This transparency problem of the model may lead to out-of-the-loop unfamiliarity, which may prevent proper decision making [5]. In addition, this may lead to distrust in the information provided by the diagnostic support system itself [6].

To address this issue, some prior research has aimed to develop technologies that utilize Explainable Artificial Intelligence (XAI) to provide operators with both the model's diagnostic results and their interpretations for abnormal situations. Park et al. proposed providing diagnostic evidence to operators by explaining the diagnosis model using a Gated Recurrent Unit–Autoencoder and Light Gradient Boosting Machine, combined with SHapley Additive exPlanations (SHAP) [7]. Reddy et al. introduced a method to identify model errors by measuring uncertainty quantification for incident identification in NPPs and explaining the features contributing to this uncertainty through SHAP [8]. Kim et al. proposed a new methodology to determine appropriate perturbation values in perturbation analysis, aiming to identify an XAI method suitable for NPP accident diagnosis models and enhance model trustworthiness through explanations [9]. Zhang et al. developed a

[*] Corresponding author.
E-mail address: sjlee420@unist.ac.kr (S.J. Lee).

https://doi.org/10.1016/j.net.2025.103589
Received 23 December 2024; Received in revised form 14 February 2025; Accepted 15 March 2025
Available online 17 March 2025

novel neural network architecture that combines Long Short-Term Memory and attention mechanisms for rapid abnormal event detection, improving interpretability through Derivative Dynamic Time Warping Methods [10]. Providing explanations for the model's results addresses transparency challenges and enables diagnostic support systems to offer operators clear reasoning behind the model's decisions. Prior research has adopted the model's explanations as they are, but the following additional considerations regarding explanations are necessary. First, it should be noted that if the operator does not understand the explanation, excessive information may lead to confusion [11]. Furthermore, the trustworthiness of the diagnostic support system may not be resolved solely by providing model explanations [12].

This study aims to improve the understandability of diagnostic support information for operators performing abnormal event diagnosis tasks and, ultimately, to enhance the trustability of diagnostic support systems. We propose a concept that provides understandable diagnostic cause visualization from a support system that includes XAI. First, the monitoring parameters in the dataset are rearranged based on the positions of their respective components or systems to reflect parameter relationships through the characteristics of convolution operations. Next, the proposed concept includes a method to refine the explanation of the model's results by using Multilevel Flow Modeling (MFM) to ensure that only explanations consistent with the physical flow are retained. These explanations are visualized on the plant map at the component level and provided to the user as diagnostic cause information. Finally, we implemented an interface based on the proposed concept. Users can be provided with the diagnostic result from the model along with the anticipated diagnostic cause information used to infer that result. Through the proposed concept, diagnostic cause information is provided so that the user can understand it without departing from the physical flow and can intuitively recognize the information through visualization.

2. Concept development

In this section, we propose two approaches to provide understandable diagnostic cause visualization. The first approach introduces a model that can better learn parameter relationships by preprocessing and rearranging the dataset parameters based on their component positions. The second approach involves selective cause visualization using intuitive relevance calculation with MFM at the component level through XAI. These concepts are shown in Fig. 1 below.

First, the development environment for the concept proposed in this study is based on the 3KEYMASTER Simulator by Western Service Corporation [13]. It is a full-scope simulator for a generic pressurized water reactor with a 2-loop structure and includes the functions of most components in an NPP.

2.1. Abnormal event diagnostic model

In this study, we propose a model that can learn the characteristics of related parameters in NPP abnormal situations. In NPPs, abnormal situations refer to conditions where the systems, components, or operating parameters of an NPP deviate from the normal operational state or exhibit an unexpected symptom. When an abnormal event occurs in a specific component in an NPP, monitoring parameters that depend on other components may remain unaffected. In contrast, emergency situations are accompanied by rapid cascading changes across the reactor coolant system, turbine system and electrical system throughout the plant following a reactor trip. In other words, abnormal scenarios have to consider impacts on local parameters by targeting a relatively narrow range, unlike emergency scenarios. In this aspect, this study suggests an explainable model by rearranging features within the datasets to emphasize local patterns of abnormal events, and utilizing a CNN to fully leverage their locality and spatial relationship characteristics.

2.1.1. Two-channel convolutional neural networks

A CNN is a deep learning model primarily used for processing and analyzing image data [14]. In CNN, convolutional layers perform convolution operations on input data using small filters [15]. This allows the network to focus on smaller regions of the data through local connectivity rather than analyzing the entire dataset at once. Consequently, the network can effectively capture subtle features, even when abnormal events affect only a small subset of monitoring parameters in NPPs. Lee et al. showed that a two-channel CNN outperforms a single-channel model in classifying abnormal events in an NPP [16]. The two-channel input data structure is designed to account for both current states and temporal changes. The dataset, structured as a two-dimensional parameter value-by-time matrix, is transformed into snapshots of parameters at each time point for input into the CNN. In this structure, each feature in the input data corresponds to a single parameter point. The first channel contains the current values of monitoring parameters, while the second channel includes the variations in these parameters over a 5-s interval to enable rapid diagnosis. Tracking both the current

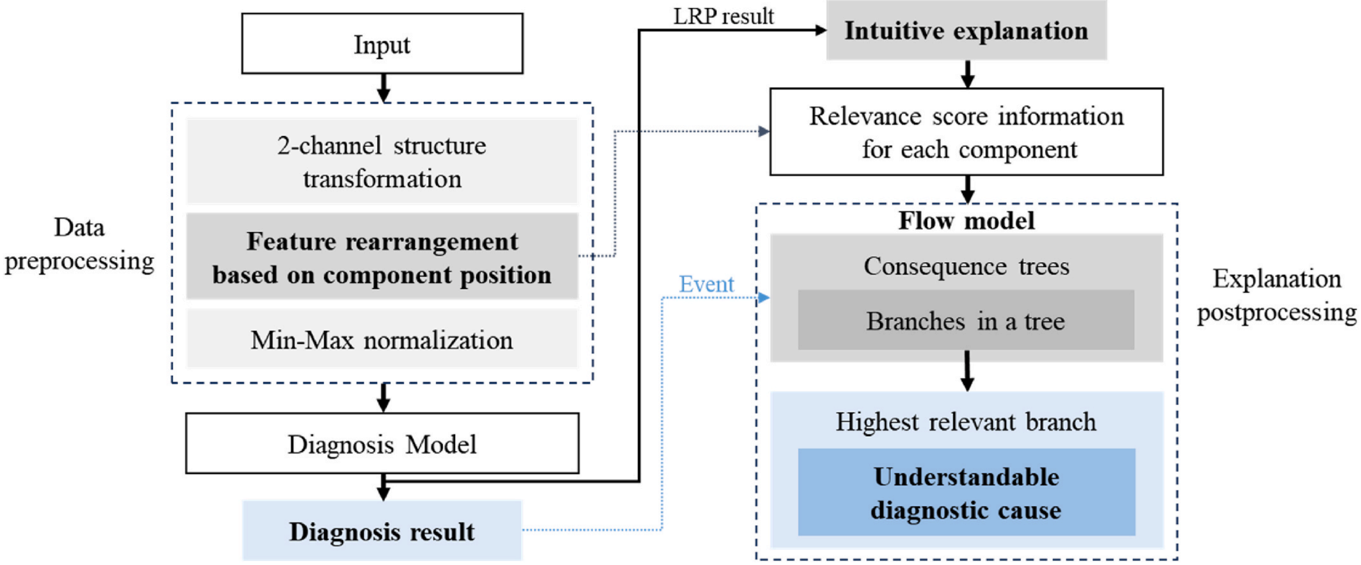

Input
2-channel structure transformation
Feature rearrangement based on component position
Min-Max normalization
Data preprocessing
Diagnosis Model
Diagnosis result
LRP result
Intuitive explanation
Relevance score information for each component
Flow model
Consequence trees
Branches in a tree
Event
Highest relevant branch
Understandable diagnostic cause
Explanation postprocessing

Fig. 1. Proposed concept for diagnostic results with understandable cause.

values and their variations allows the model to more effectively identify abnormalities and sudden shift in the data. Based on this rationale, this study utilizes a two-channel CNN as an abnormal event diagnosis model. An example of the values represented by the features within each channel is shown in Fig. 2 below.

2.1.2. Feature rearrangement based on component position

The data preprocessing for the given data set is essential to ensure optimal model performance. A CNN is capable of learning localized features in data while preserving spatial information. For instance, when input parameters of a specific component are physically close to each other, the network can effectively capture the relationships between these parameters using local filters [17]. Additionally, by detecting characteristic patterns such as abnormal values occurring in specific regions, a CNN is advantageous for understanding how a particular component within a system interacts with other components. It means that the feature rearrangement directly influences training efficiency and the distribution of weights for the CNN model [18,19]. Feature rearrangement in the input data helps the CNN model better learn relationships between adjacent features by leveraging its spatial pattern recognition capability. As a result, more persuasive explanations with a higher concentration of relevance on key feature groups can be provided to operators. In contrast, without feature rearrangement, the model may need to use deeper layers or larger kernels to capture inter-feature relationships, which can result in the weights to be dispersed across multiple locations. Consequently, it becomes more challenging to identify what the model is focusing on in its explanations. Considering these aspects, this study rearranges each monitoring parameter according to the position of the system or component it depends on. This rearrangement facilitates training by capturing systematic information through the restructured pattern of the final dataset for each abnormal event. Consequently, the one-dimensional parameter information at each time step is transformed into a two-dimensional image structure, as shown in Fig. 3. In this arrangement, each parameter is assigned a position reflecting the topology of its dependent component.

For data preprocessing, we selected 30 systems or components along with 391 associated monitoring parameters, as detailed in Table 1 below. These parameters were reorganized into a 1900-pixel image with dimensions of 38 b y 50, arranged systematically. Some parameters were redundantly used to fill the component-specific regions of the mapped image. The approximate spatial arrangement of components within the system, such as trains, pumps, and valves, was not considered during this process.

The data from the NPP simulator samples changes in monitoring parameters over time, with each parameter exhibiting unique characteristics. Some parameters have discrete binary values, such as pump operating status, while others have continuous values, like valve openings between 0 and 1. Additionally, parameters representing thermodynamic properties, such as temperature and pressure, vary in range. These differences can result in inconsistent value ranges, potentially degrading model performance if not properly processed. To address this, the data was normalized using the maximum and minimum values for each parameter. Fig. 4 shows an example of the first channel after the proposed preprocessing was applied.

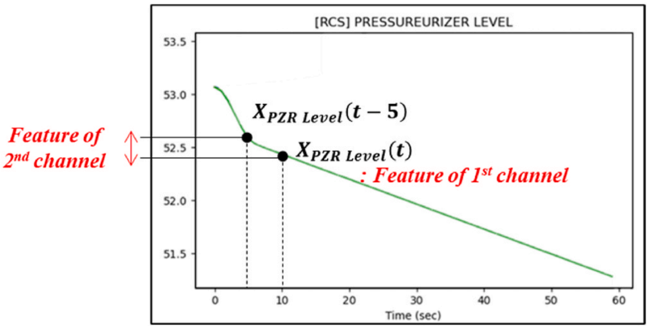


Fig. 2. The example of features within each channel.

The final two-channel input data, transformed by feature rearrangement, will be used to train CNN. This approach leverages the plant map-based images to provide snapshot-like patterns, enabling the model to utilize convolution operations effectively.

2.2. Understandable diagnostic cause visualization

Multi-class classification models commonly use the softmax activation function to learn separable features between classes [20,21]. In other words, the nodes and weights activated during classification reflect the features that separate the data into specific classes. Using XAI techniques, the classification contribution of input features can be analyzed, providing users with valuable insights into how the model makes decisions and the factors influencing its predictions. Similarly, operators provided with diagnostic results can gain insights from the parameters contributing to the diagnosis model's classification using XAI. This diagnostic cause information can support operators in forming their own diagnostic inferences. However, the parameters identified as contributing factors by the model may not always align with the operator's knowledge, as they can be influenced by factors such as the kinds of events the model classifies, used training scenarios, and the structure of the model itself. In such cases, the operator might struggle to understand the provided information, leading to potential confusion. Prior research has attempted to enhance the understandability of explanations by integrating XAI methods with domain knowledge, and another study has demonstrated that users can better evaluate model performance when XAI methods and domain knowledge are integrated [22,23]. Additionally, a filtering-out approach using guidelines for potential areas has been proposed to improve explanations [24]. In this study, we propose using a flow model to present the expected cause information along with the diagnosis results in a way that is both clear and useful to the operator.

2.2.1. Parameter relevance score

This study aims to provide operators with inference support and insights obtained from the model's diagnostic process through explanations. Therefore, instead of model-agnostic explanations that rely solely on input-output relationships, we employ model-specific explanations that utilize the model's internal information directly. Among them, Layer-wise Relevance Propagation (LRP) was chosen as the XAI technique to gain insights into the diagnostic model, as it reflects the activation of the model's internal weights during calculations. LRP maximizes the use of internal information by propagating the model's output relevance backward through its layers, redistributing relevance scores to input features based on their influence on the final prediction [25,26]. In other words, this process involves decomposing the output relevance layer by layer and propagating relevance to calculate the contribution of each neuron to the output. An ε term is added to the denominator to prevent issues when the denominator approaches zero to ensure computational stability. The main Eq. (1) for LRP is:

$$R_j^{(l-1)} = \sum_i \frac{a_j^{(l-1)} w_{ji}}{\sum_k a_k^{(l-1)} w_{ki} + \varepsilon} R_i^{(l)} \tag{1}$$

Using this explanation technique to calculate the diagnosis results allows for the determination of relevance scores for each feature. For image data, relevance scores are typically visualized as heatmaps. However, this study used Eq. (2) to enhance clarity before generating heatmaps [27,28]. It provides explanations only for the parts that have made positive contributions to the model's output. In other words, this equation eliminates negative explanations of the model's output, allowing operators to focus on the relevant causal factors without experiencing confusion from conflicting information. Next, the

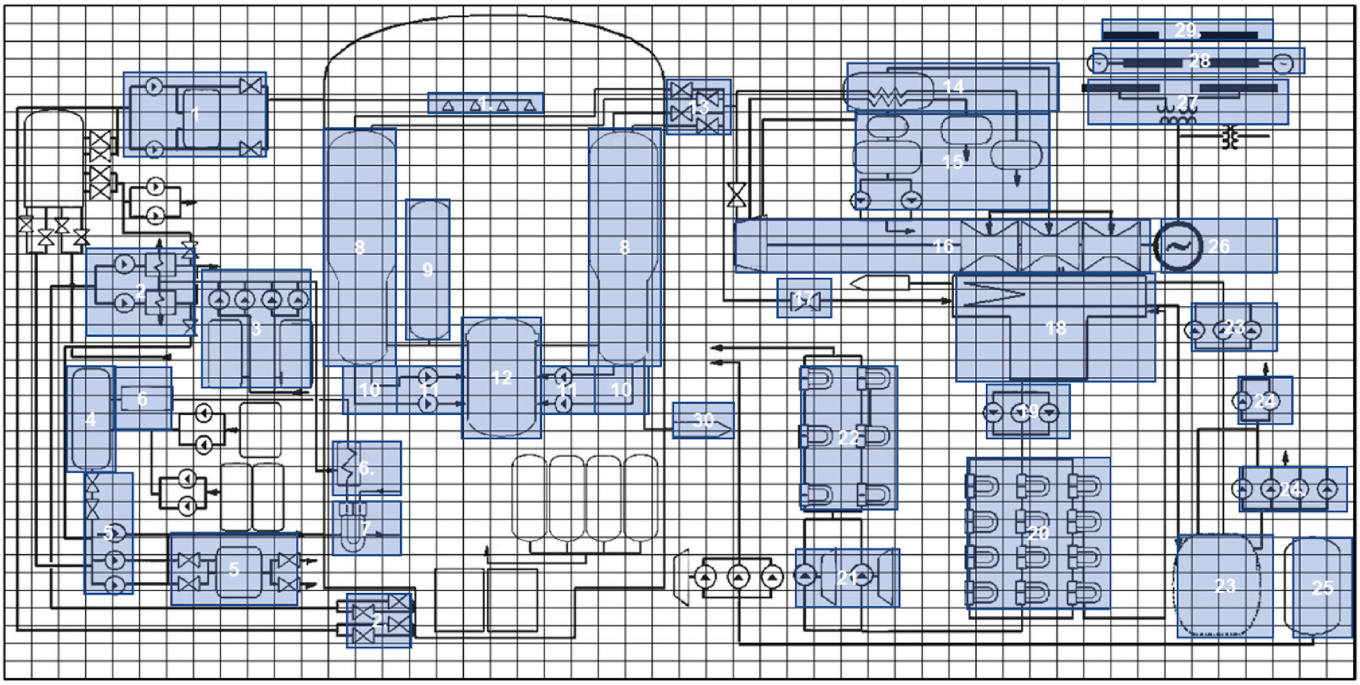


Fig. 3. Feature rearrangement based on component position.

Table 1
Component or system with selected parameters.

Component num.	Component/system	Num. Of parameters
1	Containment spray system	11
2	Residual heat remover system	3
3	Component cooling water system	20
4	Volume control tank	4
5	Charging system	14
6	Letdown system	9
7	Regenerative heat exchanger	4
8	Steam generator	10
9	Pressurizer/pressurizer relief tank	13
10	Reactor coolant system	12
11	Reactor coolant pump	18
12	Reactor/Reactor rod control system	6
13	Main steam system	12
14	Moisture and reheat steam	24
15	Feedwater heater extraction, drains, and vents system	48
16	Main turbine	20
17	Turbine bypass valve, steam dump system control	9
18	Condensate system	14
19	Condensate pump	13
20	Low-pressure feedwater heater	15
21	Main feedwater pump	16
22	High-pressure feedwater heater	14
23	Circulating water system	25
24	Essential service water system	6
25	Condensate storage tank	8
26	Main generator	15
27	Electric system (high-voltage 13.8 KV)	5
28	Diesel generators control system	8
29	Electric system (Medium-voltage 4.16 KV)	5
30	Steam generator blowdown	10
Total number of parameters		391

relevance scores are normalized to a range between 0 and 1, defined by the maximum value R_{max} and minimum value R_{min}. By focusing only on features with high relevance, this approach reduces the information presented to the operator, making it more concise and easier to identify.

The corresponding Eq. (2) and Eq. (3) are as follows.:

$$R_j^{(l-1)} = \sum_i \left(\frac{\left(a_j^{(l-1)} w_{ji}\right)^+}{\sum_k \left(a_k^{(l-1)} w_{ki}\right)^+ + \varepsilon} \right) R_i^{(l)} \tag{2}$$

$$R_{normalized}(x) = \frac{R(x) - R_{min}}{R_{max} - R_{min}} \tag{3}$$

2.2.2. *Diagnostic cause selected based on physical flow model*

This study introduces a selective diagnostic cause visualization method based on physical flow model. Diagnostic cause selection focuses on isolating causal information that operators can easily comprehend. This process filters out unnecessary details that may still have non-zero relevance scores by Section 2.2.1 but are not aligned with the operator's understanding or do not require confirmation. For this, this study proposes using MFM to exclude the information that diverges from the system's physical flow.

MFM is a methodology designed for modeling complex industrial systems, such as NPPs, by explaining goals and functions, means-end, and causal relations within mass and energy flow systems. This approach enables qualitative reasoning about task success or failure by incorporating the physical flow of the system. For this study, MFMSuite, a tool developed at the Technical University of Denmark and integrated with an editor created by IFE Harden, was used for designing and analyzing MFM [29]. Our model includes all 30 systems and components introduced in Section 2.1.2, encompassing three energy flow systems and three mass flow systems, detailed as follows.

1. Energy flow system
   A Reactor coolant heat remover system
   B Primary component cooling system
   C Electrical system
2. Mass flow system
   A Reactor coolant system and chemical volume control system
   B Steam generator and secondary system
   C Containment spray system

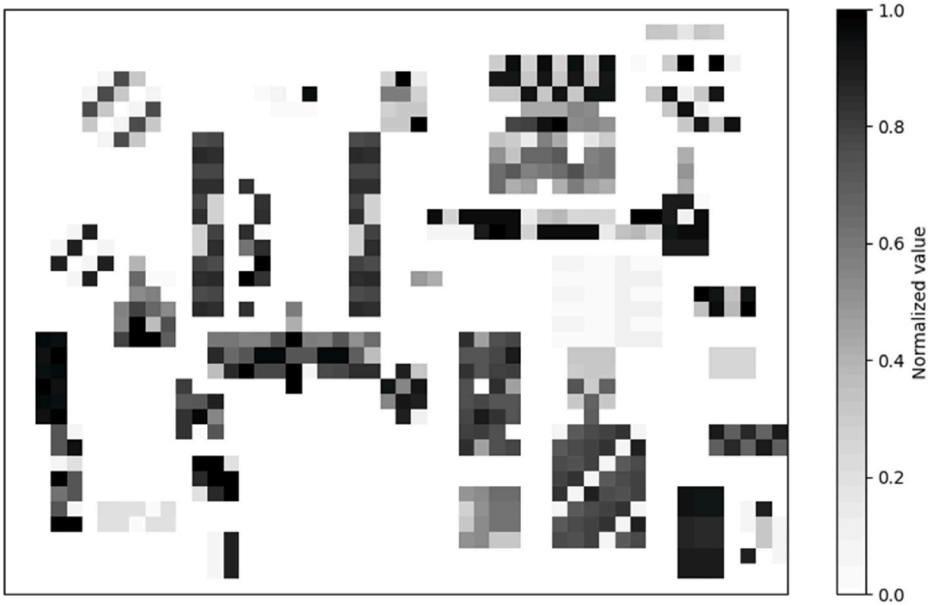


Fig. 4. An example of the first channel of the CNN model after feature rearrangement based on component position.

This model was designed to reflect the flow in 3KEYMASTER simulator, with primary and secondary materials categorized into distinct mass flow systems, ensuring circulation within each system. Additionally, energy transfer between components is represented through the energy flow system. For instance, the transfer of heat energy from the primary system to the secondary system via the steam generator U-tube exemplifies this concept. The resulting model of the NPP simulator used in this study is shown in Fig. 5.

To represent only the overall flow of the NPP systems, each physical component is simplified in the model as one or more flow functional elements, such as source functions or transport functions. In the designed MFM, an appropriate trigger state is selected for the MFM flow function to simulate the abnormal event trained in the diagnosis model. This trigger causes the mass or energy state of the flow function to be low or high, which is then propagated to the next flow structure [30]. Consequence analysis is conducted by injecting an abnormal event as a trigger state into a specific flow function in the MFM and activating the prognosis function. The injected state is analyzed as shown in Fig. 6, generating a consequence tree that includes all branches representing potential paths where physical flow can impact other flow functions which correspond to other systems or components. Each branch outlines the flow path until it reaches a possible end consequence, as shown in Fig. 7.

Through this process, identifying the component associated with the flow path makes it possible to determine which component can influence or be influenced by the specific abnormal event. The parameters

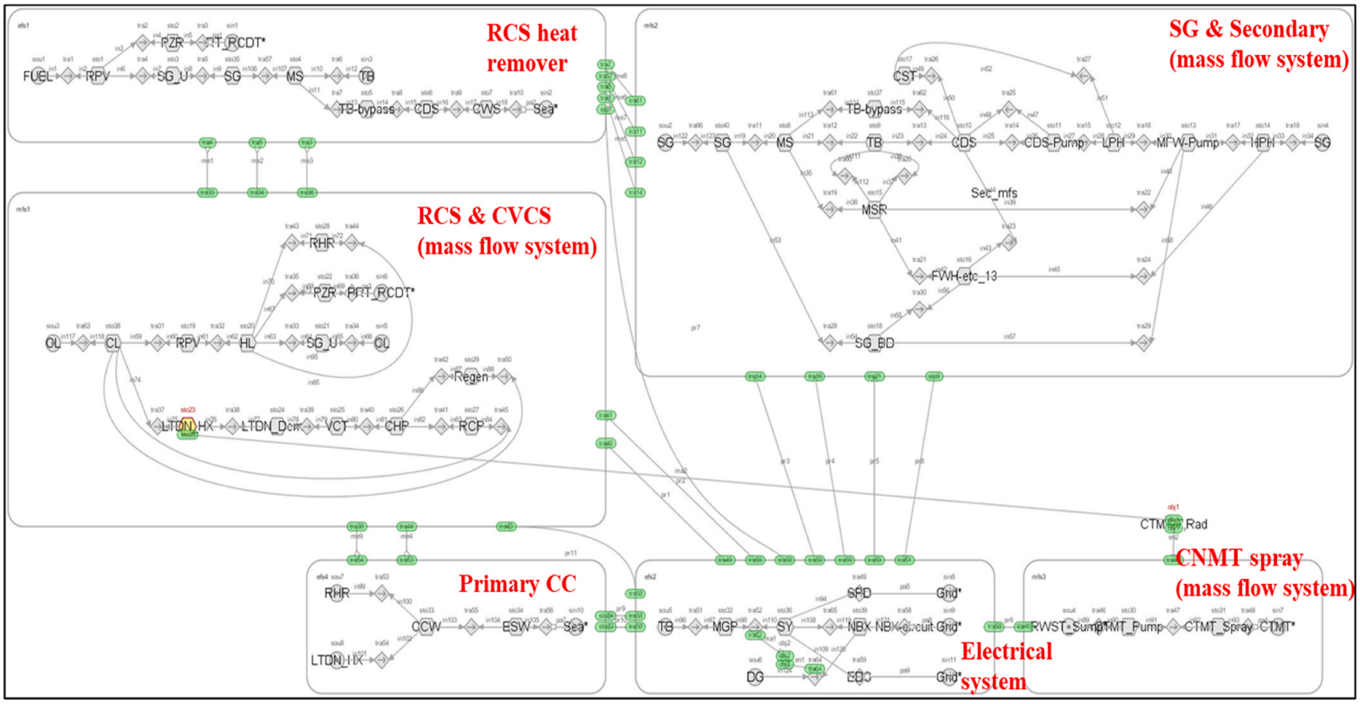
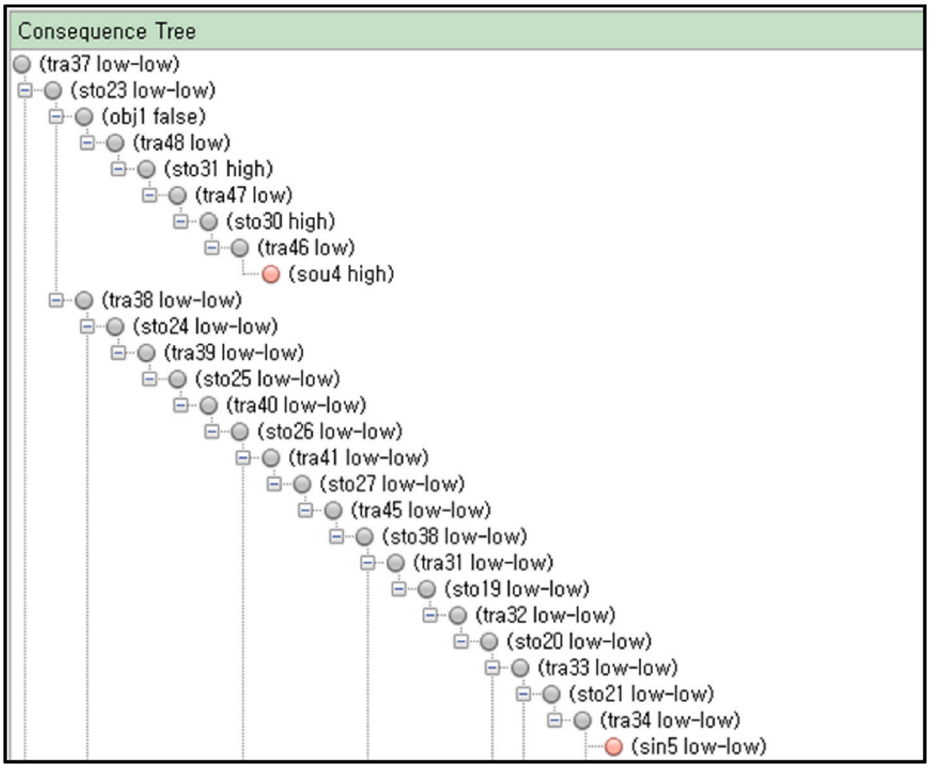


Fig. 5. MFM designed about 3KEYMASTER simulator.



Consequence Tree

- (tra37 low-low)
- (sto23 low-low)
  - (obj1 false)
    - (tra48 low)
      - (sto31 high)
        - (tra47 low)
          - (sto30 high)
            - (tra46 low)
              - (sou4 high)
  - (tra38 low-low)
    - (sto24 low-low)
      - (tra39 low-low)
        - (sto25 low-low)
          - (tra40 low-low)
            - (sto26 low-low)
              - (tra41 low-low)
                - (sto27 low-low)
                  - (tra45 low-low)
                    - (sto38 low-low)
                      - (tra31 low-low)
                        - (sto19 low-low)
                          - (tra32 low-low)
                            - (sto20 low-low)
                              - (tra33 low-low)
                                - (sto21 low-low)
                                  - (tra34 low-low)
                                    - (sin5 low-low)

Fig. 6. Example of the consequence tree.

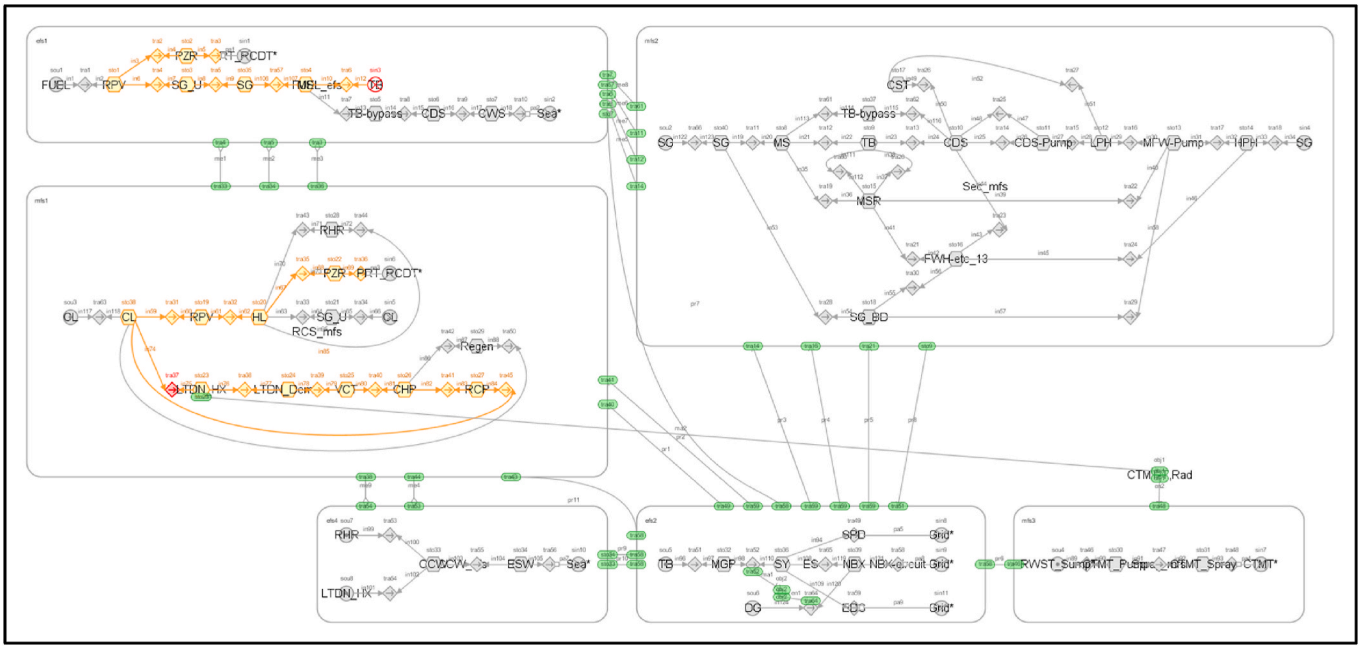


Fig. 7. Example of the potential path by branch.

related to each component are identified in Table 1, and the relevance of each parameter can be confirmed through the results of LRP. This relevance is provided to the operator to aid in inference. The information provided through this approach can offer insights into verifying the model's diagnosis, much like how symptoms and alarms from operating procedures are utilized.

To provide the operator with information on the most relevant branch, the following steps are performed.

1. The highest relevance score among the parameters associated with each component is identified.
2. The maximum relevance scores for all components within each branch corresponding to the diagnosed event are summed.
3. The branch with the highest total relevance score is selected.
4. The operator is presented with a list of components included in the branch with the highest total relevance score.
5. Parameters associated with components outside this branch are excluded and not provided.

In this study, explanations are confined to the MFM consequence tree of the diagnosed event to support the operator's understanding. To achieve this, (1) branches directly related to the event are prioritized, and (2) components included in the explanation are limited to those within a single branch to maintain consistency in the physical flow. Even if an individual component has a high relevance score, the explanation is provided based on the overall flow of the branch with the highest cumulative relevance score to prevent operator confusion. The explanations derived from XAI techniques undergo postprocessing to generate relevance scores understandable as physical flows. Furthermore, the highest relevance score for each component is extracted and used as the relevance score for that component. This approach enables the provision of component-level relevance scores, offering the potential for visualization on a plant map image.

3. Implementation with user interface

When explaining the classification of image data, relevance scores are typically visualized as a heat map. In this study, abnormal event diagnosis is conducted using image data that reflects component positions obtained through preprocessing. The result without Section 2.2 can be represented as a heat map, as shown in Fig. 8.

This example shows a heatmap of relevance scores corresponding to the 'component cooling water service loop header leak to auxiliary atmosphere' abnormal event mapped onto the system layout. The red regions highlight areas where the parameter relevance scores are the highest, suggesting that the primary causes are likely centered around the component cooling water system. However, several diagnostic causes visualized appear in other areas, where their connection to the abnormal event is less clear. This indicates that the current visualization may overwhelm operators with irrelevant information, making it difficult to pinpoint the root cause. By comparison, the proposed diagnostic cause visualization based on component level only with Section 2.2, as shown in Fig. 9, aims to address this.

In Fig. 9, the parameters contributing to the model's diagnosis of the abnormal event are mainly distributed within the component cooling water system and the Residual heat remover system. The proposed diagnostic cause visualization intuitively represents the relevance of components or systems by indicating the size of the red-highlighted areas according to their degree of relevance to the diagnosis results. Additionally, this approach delivers causal information consistent with the physical flow, enabling operators to comprehend the insights provided, even when they deviate from existing entry conditions. The user interface is designed to incorporate this visualization generated by the proposed approach, allowing operators to efficiently comprehend the diagnostic information, as shown in Fig. 10.

To ensure simplicity and clarity, the main interface displays only essential diagnostic information while excluding additional details derived from internal models or techniques, such as prediction values from the diagnostic model and relevance scores calculated through explanation methods. Diagnostic cause is visualized on an NPP map to help operators intuitively understand the information. The map has to be designed to resemble the layout operators are accustomed to in the control room, including the existing large display panel. For this study, the overview map from the 3KEYMASTER simulator, used in the development environment, was adopted.

The user interface includes the following functions.

1. Abnormal Event Diagnosis Results: This feature displays the abnormal events identified by the diagnosis model. Operators can use this function to monitor abnormal events currently occurring in the NPP.
2. Heatmap Display: This function visualizes the relevance scores calculated for abnormal event diagnosis on the plant map. Only components or systems relevant to the event are highlighted in red, with the marker size indicating the degree of relevance. Additionally, the names of relevant components or systems are activated using a button function, and the most relevant ones are marked in black. This feature supports operators by intuitively identifying the components involved, enabling them to diagnose provided events independently.
3. Real-Time Trend Plotting: When an operator selects an activated component or system, the interface plots real-time trends for parameters with high relevance scores, offering additional insights into the event's progression.

4. Case study

We conducted a case study on the proposed concept and its interface as follows. First, a training dataset for building a model to diagnose abnormal events was generated and used to train the model according to the approach introduced in Section 2.1. Several representative cases were diagnosed by the model, and relevance scores for diagnosis results were output using the XAI technique. These relevance scores were refined following the approach detailed in Section 2.2 and subsequently visualized through the user interface.

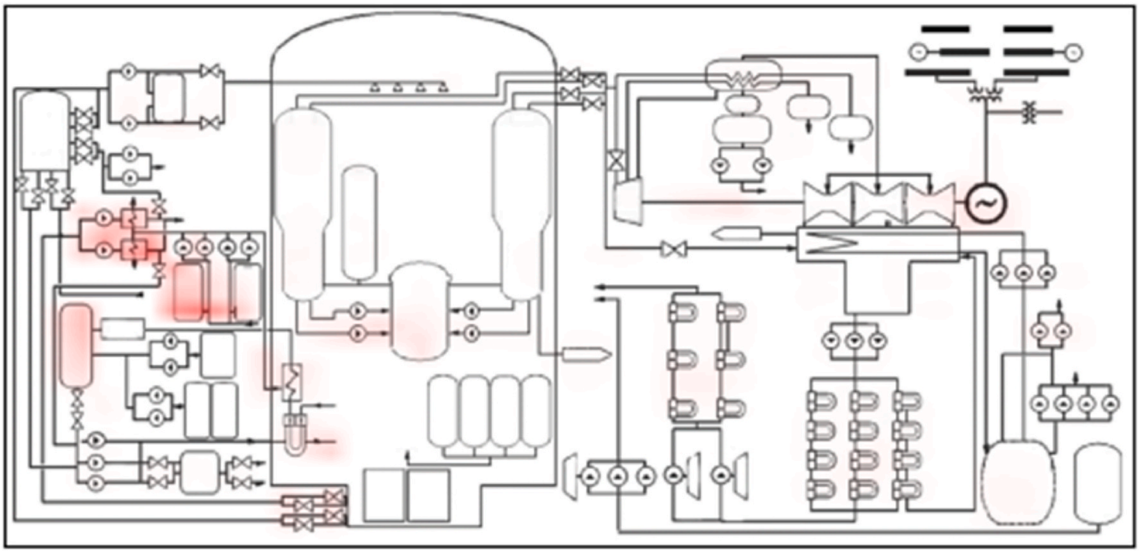


Fig. 8. Example of the heatmap with raw relevance scores.

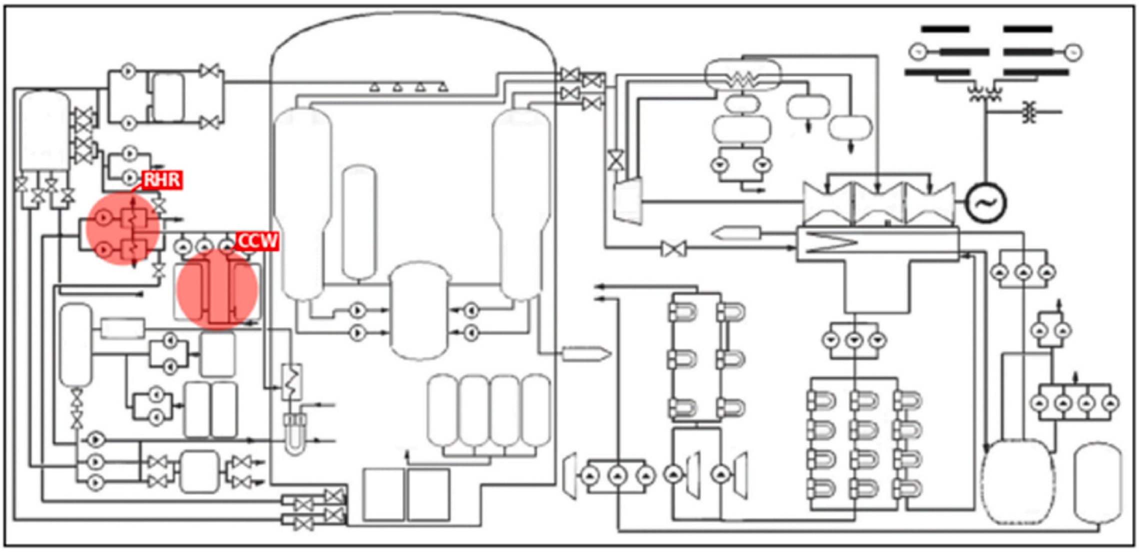


Fig. 9. Example of the heatmap with understandable diagnostic cause.

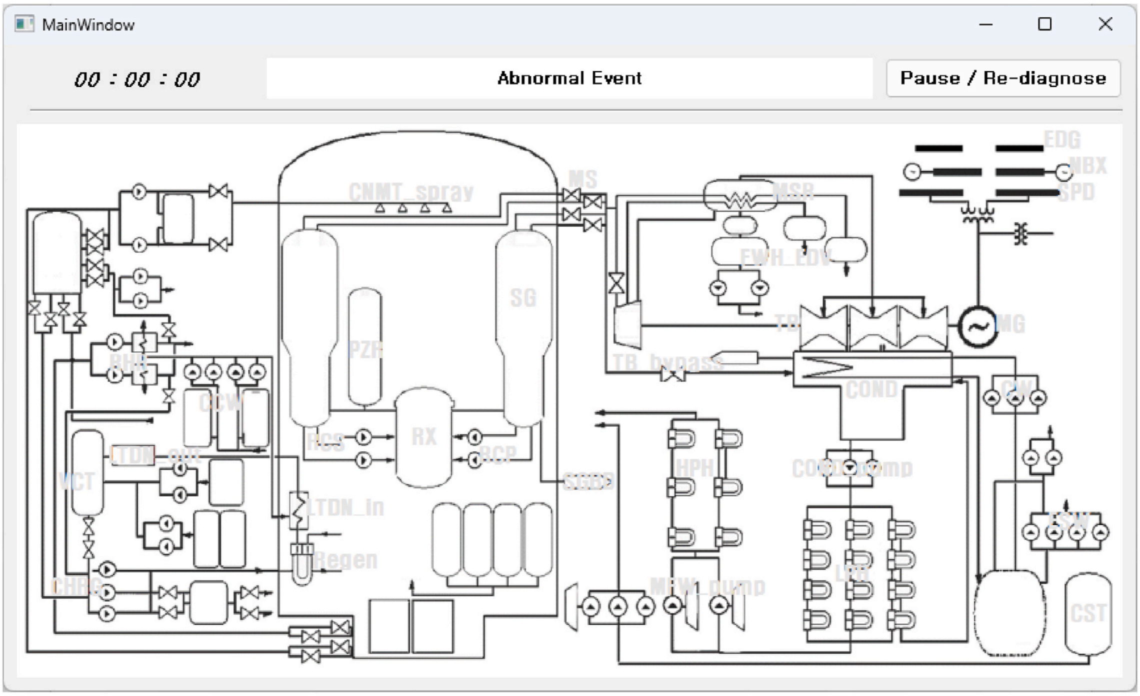


Fig. 10. User interface for proposed concept.

4.1. Abnormal event diagnostic model training

The dataset consisted of one normal and 15 abnormal situations, comprising 391 monitoring parameters sampled every second over 60-time steps. Through the NPP simulator, a total of 425 scenarios were generated for model training using scenario scripts written with evenly spaced intervals within the malfunction (MF) fraction range. Table 2 provides a description of the abnormal events included in the training dataset.

The constructed dataset underwent preprocessing, including feature rearrangement, and was used to train a 2-channel CNN model for abnormal event diagnosis. It was trained using this dataset divided into a training set and a validation set at a 7:3 ratio. During model training, early stopping was applied with the validation loss monitored and a patience of 20 epochs. The structure and hyperparameters of the model are shown in Table 3.

The model achieved a log loss of 9.81e-05 on the training dataset and 8.19e-05 on the validation dataset. Fig. 11 shows the model's training process across epochs.

4.2. Comparative study of the proposed approach

This section compared explanations with and without the proposed feature rearrangement in this study. Additionally, it examines whether the selected LRP method effectively identifies features that are highly relevant to the model's diagnosis. To evaluate the explanations, the deletion metric was employed as a quantitative evaluation technique. The deletion metric quantifies the effect of removing features with the

Table 2
Abnormal event description about used datasets.

Num.	Abnormal event	MF fraction	
		Minimum	Maximum
1	Steam generator A tube leakage	4	10
2	Charging line break upstream	10	100
3	Letdown line leakage inside containment	100	1000
4	Loss of condenser vacuum	45	50
5	Pilot-operated safety relief valve leakage	0.2	1
6	Circulating water tube leakage in low-pressure condenser	65	100
7	Main steam isolation valve positioner failure	0	0.3
8	Loss of reactor coolant pump seal injection water	0	0.03
9	Main steam header steam leakage	2	3
10	Pressurizer spray valve positioner failure	70	100
11	Component cooling water service loop header leakage	10	100
12	Low-pressure feedwater heater 1A tube break	10	100
13	High-pressure feedwater heater 5A tube break	55	90
14	Main feedwater pump recirculation valve positioner failure	0.45	0.7
15	HP turbine control valve positioner failure	0	0.25

Table 3
Abnormal event diagnosis model structure and training hyperparameter.

Layer	Parameters	Value
Input	–	–
Conv2D	Filters: 16, Kernel: (3, 3)	Activation function: Softplus [31]
Conv2D	Filters: 32, Kernel: (3, 3)	Activation function: Softplus
MaxPooling2D [32]	Pool size: (2,2)	–
Conv2D	Filters: 64, Kernel: (3, 3)	Activation function: Softplus
MaxPooling2D	Pool size: (2, 2)	–
Flatten	–	–
Dense	Units: 16	Activation function: Softmax
Model training	Optimizer	Adam [33]
	Learning rate	0.001
	Loss function	Categorical cross-entropy
	Batch size	64
	Patience for early stopping	20

highest relevance scores from the input data sequentially and measures the performance degradation of the model. To minimize distortion, we replaced the highest relevant parameter value with the normal-state value at the same simulation time step. This metric is primarily used to evaluate fidelity, determining whether the explanation method correctly identifies the features that are important for the model's prediction.

4.2.1. *Comparative study on the influence of feature rearrangement*

We aim to evaluate how the preprocessing step that involves rearranging features affects the explanations provided by the proposed approach. For comparison, we utilized two models: one where the same set of 391 features was randomly arranged in a one-dimensional sequence and another where the 391 features was rearranged according to the order of components or systems. Both models employed a one-dimensional convolutional layer with a kernel size of 3, while all other hyperparameters remained identical to those described in Section 4.1. The results for 315 new scenarios for test, which included the abnormal events listed in Table 3, indicated that all three models were successfully trained with high performance as shown in Table 4.

To evaluate the explanation results of the three models, we performed the deletion metric by sequentially removing up to ten of the most relevant parameters identified by each model in the test scenario dataset. The results of the deletion metric for the models are shown in Fig. 12.

In Fig. 12, Comparative model 1, which does not consider feature rearrangement, exhibits minimal performance degradation, whereas the proposed model shows the most significant decrease in performance. Additionally, the Area Under the Curve (AUC) values for each model are 9.38, 9.69, and 9.43 over ten steps, respectively. This suggests that the explanations output by the proposed model provide the highest fidelity to the model's diagnosis. Fig. 13 shows results of the Kernel Density Estimate and the Cumulative Distribution Function, which illustrate the distribution of the normalized relevance scores (scaled between 0 and 1) for each dataset.

These results indicate that the explanations provided by the proposed model are captured at the highest proportion within low

Table 4
Test results about three models.

	Proposed model	Comparative model 1	Comparative model 2
Feature rearrangement	Component-based	Randomization	Component-based
Accuracy	0.9998	0.9999	0.9998

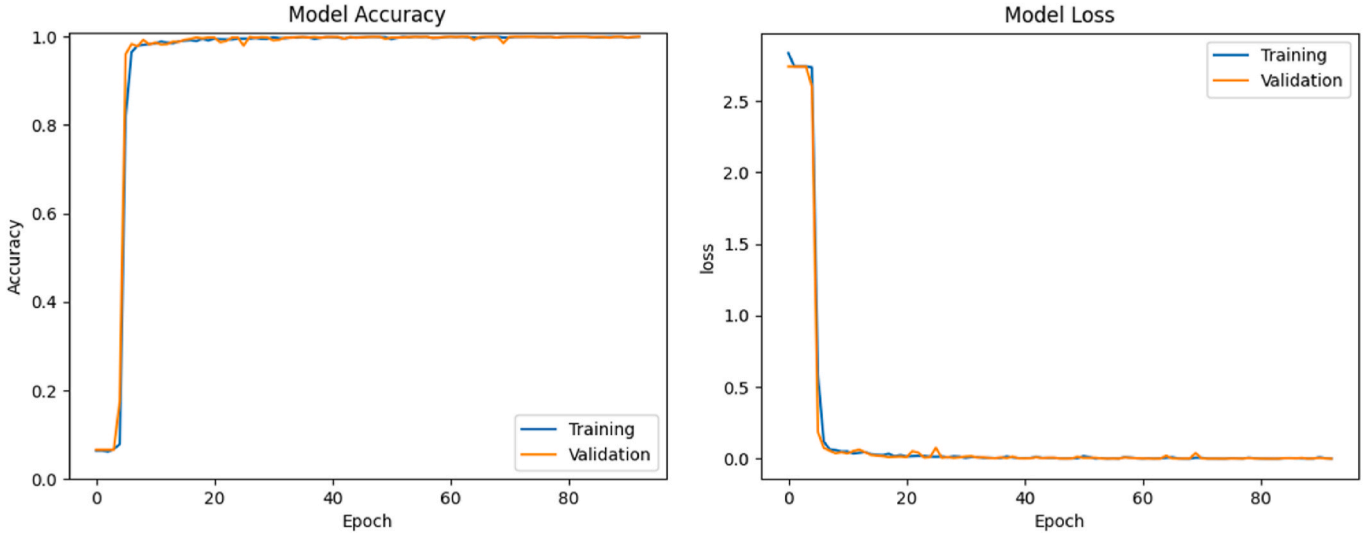


Fig. 11. Learning curves about accuracy (left) and loss (right).

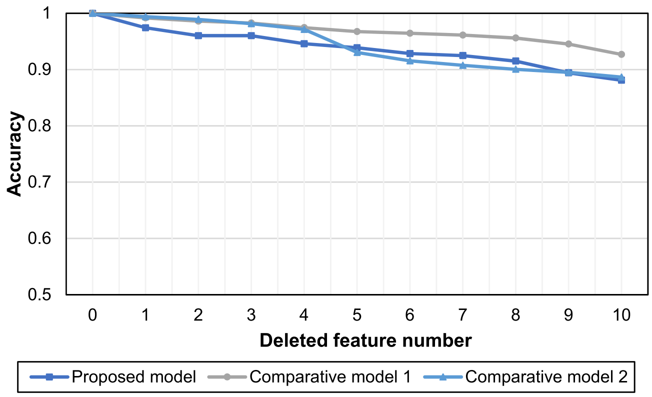


Fig. 12. Results of the deletion metric comparing the explanations of the three models.

relevance scores. Through this comparison, feature rearrangement not only enables the model to focus more on localized feature patterns during training but also facilitates providing clearer explanations with high relevance.

4.2.2. Comparative study of explanation methods

In this section, we evaluate the fidelity of the LRP method used for explanation. For comparison, we selected the Saliency Mapping and Gradient-weighted Class Activation Mapping (Grad-CAM) as comparative explanation methods. Similar to LRP, these methods can be applied to explain the contribution of input features to the model's output in CNNs, relying on the internal structure of the neural network and utilizing gradient-based computations. Using the models trained in Section 4.1, we performed the deletion metric with these three explanation methods. The results are shown in Fig. 14.

The AUC values for the three explanation methods, evaluated based on the 10 most relevant parameters, are 9.38, 9.96, and 9.60, respectively. These results demonstrate that the LRP method achieves the highest fidelity in explaining the model's diagnosis outputs compared to other model-specific explanation methods.

4.3. Results with implemented concept

This section aims to conduct several case studies to verify the understandable diagnostic cause visualization for the trained abnormal event diagnosis model. We determine how the output of LRP is reconstructed into information presented in the user interface through the physical flow model. Additionally, we compared the final provided cause information with the existing entry conditions used for diagnosis. The states of the MFM simulation corresponding to each abnormal event are as shown in Table 5 below. For example, in an MFM, to simulate an abnormal event involving letdown line leakage inside the containment, a low-state trigger can be set for transport 37, a flow function responsible for transferring fluid mass from the cold-leg to the letdown heat exchanger.

The case studies were conducted as follows, focusing on representative abnormal events: that occurring in the steam generator, where heat transfer takes place from the primary to the secondary system; the abnormal event in the primary system; and the abnormal event in the secondary system. The abnormal scenarios for each case study were designed with different MF fractions from those in the training dataset to ensure distinct conditions.

1. Case Study on steam generator tube leakage event

For the scenario in Case Study 1, the left side of Fig. 15 presents a visualization of the scores of the most relevant features for each component in the explanation of Comparison model 1, as described in Section 4.2.1. The right side of Fig. 15 presents a visualization of the proposed concept integrated into the user interface.

Comparative Model 1 shows that the relevance scores are distributed across multiple components, as evaluated in Section 4.2.1. In contrast,

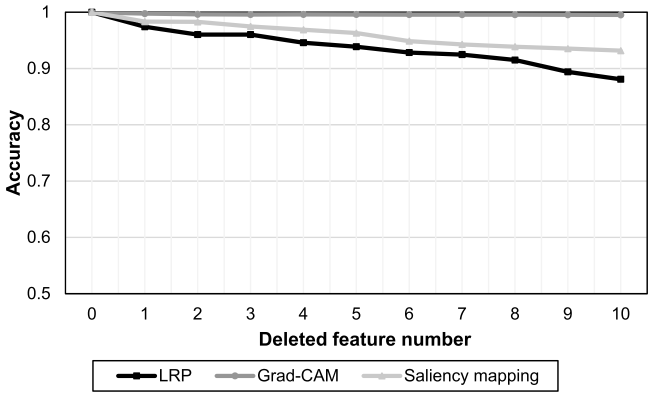


Fig. 14. Deletion metric results for comparing the three explanation methods.

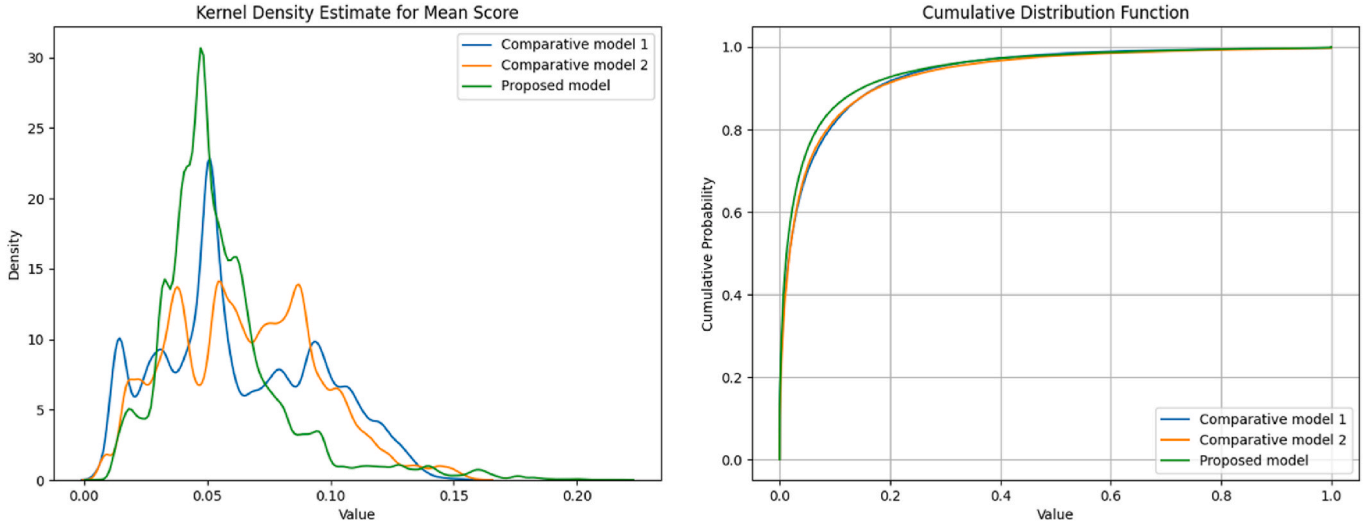


Fig. 13. Results of kernel density estimate (left) and cumulative distribution function (right).

Table 5
MFM simulation description about each abnormal event.

Abnormal event	MFM simulation	
	Flow function	State
Steam generator tube leakage	tra[a]66	High
Charging line break upstream	tra42	Low
Letdown line leakage inside containment	tra37, thr1	Low
Loss of condenser vacuum	sto10, sto6	Low
Pilot-operated safety relief valve leakage	tra36 (tra3)	High
Circulating water tube leakage in low-pressure condenser	sto[a]7	Low
Main steam isolation valve positioner failure	sto8	Low
Loss of reactor coolant pump seal injection water	sto27	Low
Main steam header steam leakage	tra11	Low
Pressurizer spray valve positioner failure	sto22, sto2	Low
Component cooling water service loop header leakage	sto33	Low
Low-pressure feedwater heater tube break	sto12	High
High-pressure feedwater heater tube break	sto14	High
Main feedwater pump recirculation valve positioner failure	sto13	Low
High-pressure turbine control valve positioner failure	tra12	Low

[a] tra: transport; sto: storage.

for the same scenario, the user interface visualized the steam generator, pressurizer, and low-pressure feedwater heater as diagnostic causes related to the detected steam generator tube leakage. This demonstrates that the proposed concept can visualize results in a way that is understandable to operators. The following Table 6 compares three key cause information based on components or systems for the event in Case Study 1, including existing entry conditions, the output of LRP, and the final output derived from the flow model. For comparison with the 391 parameters used in this study, the entry conditions in Table 6 represent the general conditions for an abnormal reactor coolant system leak to the outside of the containment due to a steam generator tube leakage event.

It shows that cause information related to the essential service water system, which falls outside the operators' scope of understanding, was successfully excluded during the final process with the flow model. Furthermore, the proposed concept identified the low-pressure feedwater heater system as one of the key causes, in addition to the steam generator, unlike the existing entry conditions. Each parameter with the highest relevance score, related to the steam generator and low-pressure feedwater heater, was plotted, as shown in Fig. 16.

These results suggest that, even when the identified causes—such as the steam generator level and the low-pressure feedwater heater temperature—differ from the operators' conventional knowledge, the model can provide insights that align with physical flow, contributing to diagnostic reasoning.

2. Case Study on loss of reactor coolant pump seal injection water

The user interface clearly visualized the reactor coolant pump as the diagnostic cause of the loss of reactor coolant pump seal injection water event, as shown in Fig. 17. Furthermore, Table 7 demonstrates that, consistent with the actual entry conditions used for diagnosis, the proposed concept maintained the pressurizer as a key diagnostic cause from the intermediate output through to the final output derived from the flow model.

When the system was selected by the user, the interface displayed the trend of the reactor coolant pump seal water inlet temperature and its flow to the user, as shown in Fig. 18.

Additionally, it can be observed that cause information about moisture and reheat steam, which is relatively less relevant to this event, was removed from the user interface through the flow model.

3. Case Study on loss of condenser vacuum

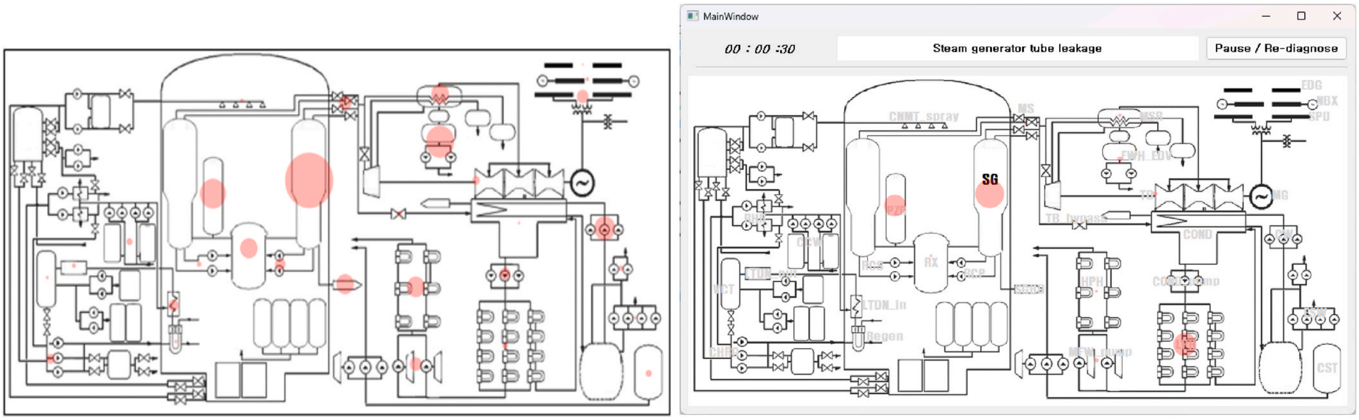
In the third case study, the proposed concept diagnosed the event as a loss of condenser vacuum while simultaneously visualizing the diagnostic causes, including the condensate system, low-pressure feedwater heater, and circulating water system, as shown in Fig. 19 and Table 8. Among them, it identified the condensate system as the most significant component contributing to the diagnostic cause.

Although the essential service water system may be relevant to the model's diagnosis, it does not directly contribute to condenser cooling and may interfere with operators' reasoning. The results indicate that, through the flow model, the essential service water system was excluded from the cause information in the final output. Furthermore, the findings of this study show that the low-pressure feedwater heater has a high relevance score. However, unlike the existing entry condition, which considers the increase in low-pressure turbine exhaust hood temperature, the proposed concept reasonably diagnosed based on the increase in low-pressure feedwater heater inlet temperature, as shown in Fig. 20. Although this parameter differs from the conventionally used entry conditions, its increasing trend can provide new insights.

These results show that the interface developed based on the

Table 6
Entry condition and explanation results for Case Study 1.

	Entry condition	Output (LRP)	Output (User interface)
Component/ system	- Pressurizer - Reactor coolant system - Volume control tank - Charging system	- Essential service water system - Steam generator - Low-pressure feedwater heater	- Steam generator - Low-pressure feedwater heater - Pressurizer



00 : 00 :30

Steam generator tube leakage

Pause / Re-diagnose

Fig. 15. Visualization of explanation results by comparative model 1 (Left) and user interface (Right).

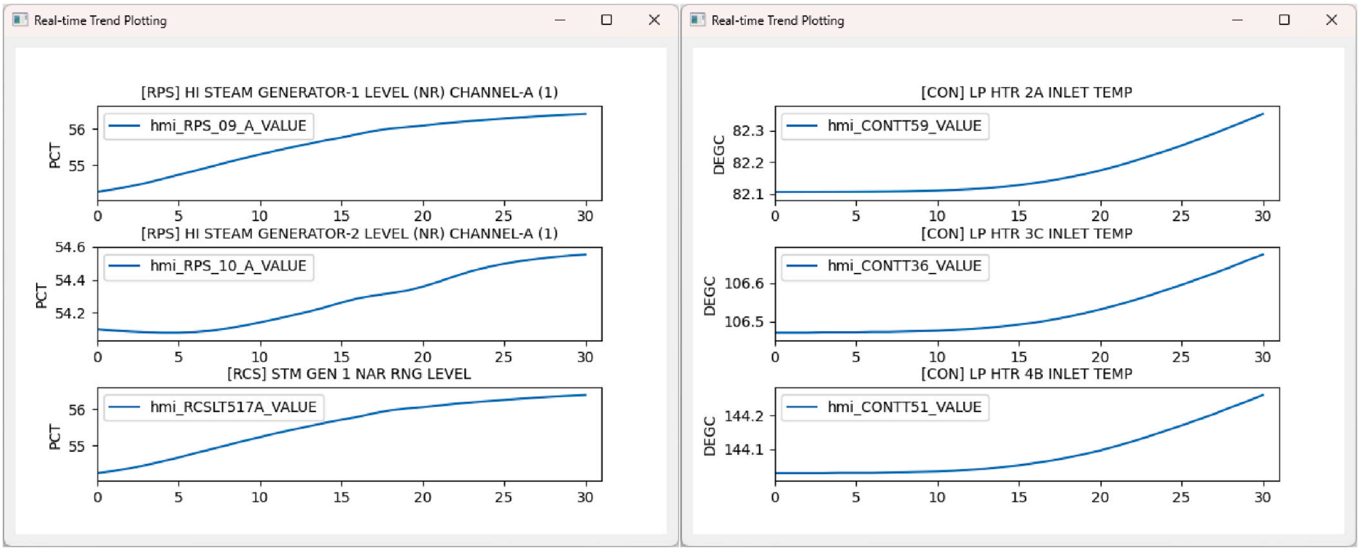
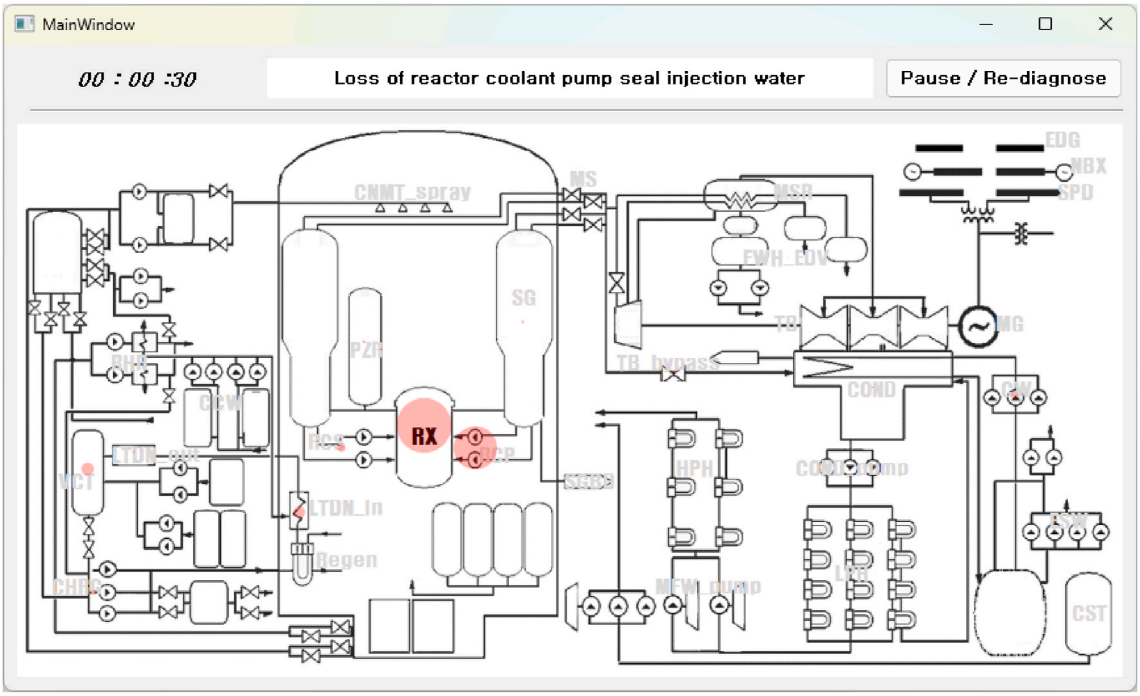


Fig. 16. Highest relevant parameter trends for Case Study 1.


00 : 00 :30

Loss of reactor coolant pump seal injection water

Fig. 17. User interface result for Case Study 2.

Table 7
Entry condition and explanation results for Case Study 2.

	Entry condition	Output (LRP)	Output (User interface)
Component/ system	- Reactor coolant pump	- Reactor - Reactor coolant pump - Moisture and reheat steam	- Reactor - Reactor coolant pump - Volume control tank

proposed concept provides users with understandable cause information alongside accurate events from the model diagnosis. This indicates that the essential cause information required for operators to diagnose events is preserved and delivered, provided it aligns with the flow defined by the MFM.

5. Conclusion

Numerous AI technologies have been explored to assist operators in diagnosing abnormal situations in NPPs. For these AI technologies to be effectively utilized as diagnostic support systems, it is essential to provide operators with clear explanations of how the models derive their diagnostic results. However, explanations generated by artificial neural networks are typically based on learned distinctions between classes, which may differ from operators' prior knowledge or the entry conditions outlined in existing abnormal operating procedures. If operators fail to understand such discrepancies, it can hinder their comprehension of the explanations, thereby limiting the applicability of the information

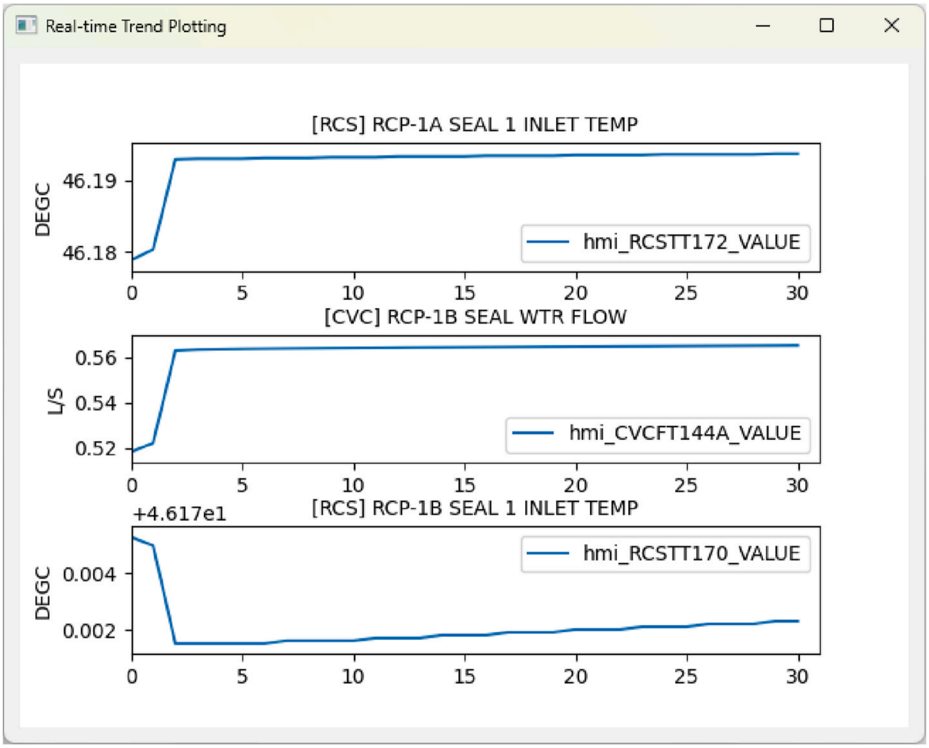
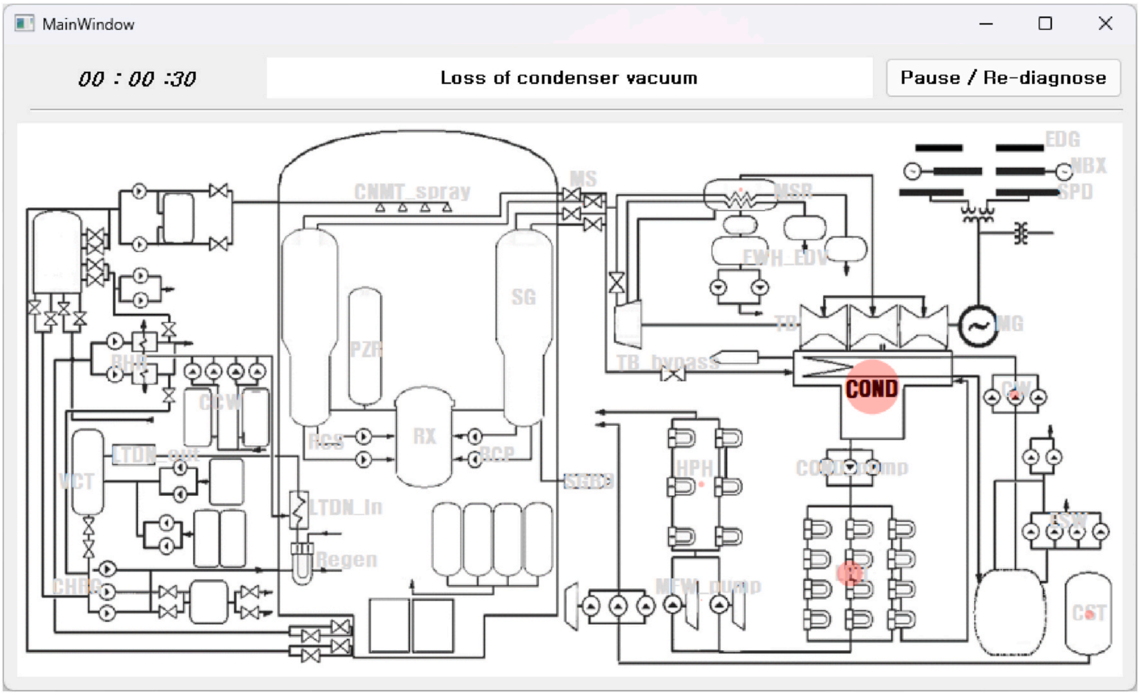


Fig. 18. Highest relevant parameter trend for Case Study 2.


MainWindow

00 : 00 :30

Loss of condenser vacuum

Fig. 19. User interface result for Case Study 3.

provided by the diagnostic support system.

This study proposes a concept to address the challenges associated with AI-based diagnostic support systems in providing results that operators can understand. First of the proposed concept, monitoring parameters within the training dataset obtained from the NPP simulator are rearranged based on the positions of each system on the plant map. This approach enables the CNN to learn abnormal patterns localized to specific regions, allowing the explanations of the model's diagnostic results to be derived based on these learned patterns. Next, the concept combines the results of LRP-ε as the XAI technique with MFM to provide operators with understandable explanations. Even if the explanations generated by LRP-ε differ from the operator's understanding, this method ensures that only results consistent with the physical flow are presented. To enhance operator cognition, the user interface was

Table 8
Entry condition and explanation results for Case Study 3.

	Entry condition	Output (LRP)	Output (User interface)
Component/ system	- Condensate system - Low-pressure Turbine exhaust hood	- Condensate system - Low-pressure feedwater heater - Essential service water system	- Condensate system - Low-pressure feedwater heater - Circulating water system

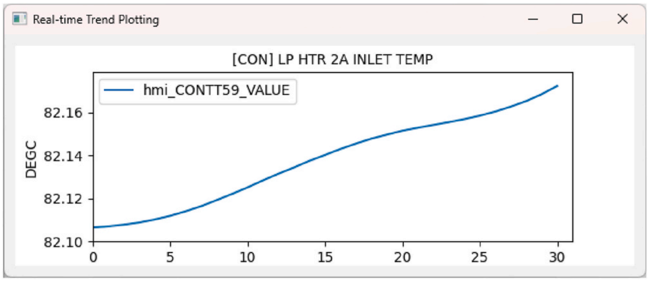


Fig. 20. Highest relevant parameter trend for Case Study 3.

simplified, and the explanations were visualized at the component level. A case study demonstrated that the proposed concept not only provides appropriate diagnostic results but also effectively visualizes diagnostic causes at a level comprehensible to the operator.

The developed concept aims to provide operators with insights into the diagnostic results based on the internal information learned by the model, delivering explanations at a comprehensible level. Additionally, these insights enable operators to provide feedback on the model's diagnostic outcomes, ultimately fostering trust in the diagnostic support system. By leveraging this concept, operators are expected to perceive the system's information as helpful, thereby enhancing the system's overall trustability. However, while AI models provide insights for classifying abnormal events based on their learning, operators typically diagnose events by considering their broader characteristics described with entry conditions. Nevertheless, the approach offers the advantage of presenting distinctive features for classification, optimized for the identification of abnormal events [34]. Therefore, future work should leverage these advantages by expanding the range of abnormal events covered in the proposed framework, enabling the provision of optimized separable diagnostic causes across the entire spectrum of abnormal scenarios. This would support more accurate differentiation of individual abnormal events within a diagnostic support system. Moreover, this study focuses on providing explanations of the model's output in a way that is understandable to operators. However, operators must process not only the new information provided by the system but also the information they already need to perceive for their existing tasks. Therefore, the information provided should be designed with ergonomic intuitiveness in mind, and the amount of information should be optimized from the operator's perspective. To achieve this, a usability evaluation should be conducted to ensure that the proposed concept can be effectively applied as a diagnostic support system.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Ji Hyeon Shin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jung Sung Kang:** Validation, Resources. **Jae Min Kim:** Conceptualization. **Seung Jun Lee:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (No. GTL24031-400). Also, this work was supported by Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government (MOTIE) (RS-2024-00403194, Next-Generation Nuclear Technology Creation IP-R&D Talent (Human Resources) Development Project).

References

[1] J.M. Kim, G. Lee, C. Lee, S.J. Lee, Abnormality diagnosis model for nuclear power plants using two-stage gated recurrent units, Nucl. Eng. Technol. 52 (9) (2020) 2009–2016.
[2] H.-J. Lee, D. Lee, J. Kim, Event diagnosis method for a nuclear power plant using meta-learning, Nucl. Eng. Technol. 56 (6) (2024) 1989–2001.
[3] F. Dong, S. Chen, K. Demachi, M. Yoshikawa, A. Seki, S. Takaya, Attention-based time series analysis for data-driven anomaly detection in nuclear power plants, Nucl. Eng. Des. 404 (2023) 112161.
[4] T.-H. Lin, T.-C. Wang, S.-C. Wu, Deep learning schemes for event identification and signal reconstruction in nuclear power plants with sensor faults, Ann. Nucl. Energy 154 (2021) 108113.
[5] S.-J. Lee, P.-H. Seong, Development of an integrated decision support system to aid cognitive activities of operators, Nucl. Eng. Technol. 39 (6) (2007) 703–716.
[6] N. Ekanem, A. Mosleh, S.-H. Shen, M. Ramos, Phoenix–A model-based human reliability analysis methodology: data sources and quantitative analysis procedure, Reliab. Eng. Syst. Saf. 248 (2024) 110123.
[7] J.H. Park, H.S. Jo, S.H. Lee, S.W. Oh, M.G. Na, A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP, Nucl. Eng. Technol. 54 (4) (2022) 1271–1287.
[8] B. Reddy, E. Gursel, K. Daniels, A. Khojandi, J. Baalis Coble, V. Agarwal, R. Boring, V. Yadav, M. Madadi, Uncertainty-aware and explainable human error detection in the operation of nuclear power plants, Nucl. Technol. 210 (12) (2024) 2312–2330.
[9] S.G. Kim, S. Ryu, K. Jin, H. Kim, Quantitative Comparison of Explainable Artificial Intelligence Methods for Nuclear Power Plant Accident Diagnosis Models, Available at: SSRN 4907597.].
[10] T. Zhang, Q. Jia, C. Guo, X. Huang, Abnormal event detection in nuclear power plants via attention networks, Energies 16 (18) (2023) 6745.
[11] G. Skraaning, G.A. Jamieson, Human performance benefits of the automation transparency design principle: validation and variation, Hum. Factors 63 (3) (2021) 379–401.
[12] D. Ahn, A. Almaatouq, M. Gulabani, K. Hosanagar, Will we trust what we don't understand? Impact of model interpretability and outcome feedback on trust in AI, arXiv preprint (2021) arXiv:2111.08222.
[13] F. Western Service Corporation, MD, USA, 3KEYMASTER Simulator, 2013.
[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).
[15] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[16] G. Lee, S.J. Lee, C. Lee, A convolutional neural network model for abnormality diagnosis in a nuclear power plant, Appl. Soft Comput. 99 (2021) 106874.
[17] R. Ghosh, A.K. Gupta, Investigating convolutional neural networks using spatial orderness, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, 0-0.
[18] A. Farahat, F. Effenberger, M. Vinck, A novel feature-scrambling approach reveals the capacity of convolutional neural networks to learn spatial relations, Neural Netw. 167 (2023) 400–414.
[19] O. Bazgir, R. Zhang, S.R. Dhruba, R. Rahman, S. Ghosh, R. Pal, Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks, Nat. Commun. 11 (1) (2020) 4391.
[20] R.D. Luce, Individual Choice Behavior, Wiley, New York, 1959.

[21] M. Grandini, E. Bagli, G. Visani, Metrics for Multi-Class Classification: an Overview, 2020 arXiv preprint arXiv:2008.05756.

[22] M. Szelążek, S. Bobek, G.J. Nalepa, Improving understandability of explanations with a usage of expert knowledge, in: European Conference on Artificial Intelligence, Springer, 2023, pp. 36–47.

[23] I. Feustel, N. Rach, W. Minker, S. Ultes, Enhancing model transparency: a dialogue system approach to XAI with domain knowledge, in: Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2024, pp. 248–258.

[24] H. Yuan, C. Hong, P.-T. Jiang, G. Zhao, N.T.A. Tran, X. Xu, Y.Y. Yan, N. Liu, Clinical domain knowledge-derived template improves post hoc AI explanations in pneumothorax classification, J. Biomed. Inf. (2024) 104673.

[25] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140.

[26] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise Relevance Propagation: an Overview, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 2019, pp. 193–209.

[27] W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, K.-R. Müller, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer Nature, 2019.

[28] E.M. Kenny, C. Ford, M. Quinn, M.T. Keane, Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies, Artif. Intell. 294 (2021) 103459.

[29] M. Lind, An introduction to multilevel flow modeling, Nuclear safety and simulation 2 (1) (2011) 22–32.

[30] J.S. Kang, S.J. Lee, Concept of an intelligent operator support system for initial emergency responses in nuclear power plants, Nucl. Eng. Technol. 54 (7) (2022) 2453–2466.

[31] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[32] B. Graham, Fractional max-pooling. arXiv Preprint arXiv:1412.6071, 2014.

[33] D.P. Kingma, Adam: a method for stochastic optimization. arXiv Preprint arXiv:1412.6980, 2014.

[34] J.H. Shin, J. Bae, J.M. Kim, S.J. Lee, An interpretable convolutional neural network for nuclear power plant abnormal events, Appl. Soft Comput. 132 (2023) 109792.