Abnormal Diagnosis using eXplainable Artificial Intelligence in Nuclear Power Plants

Ji Hun Park, So Hun Yun, Ye Ji An, Man Gyun Na*
Department of Nuclear Engineering, Chosun Univ., 309 Pilmun-daero, Dong-gu, Gwangju, Korea 61452
**Corresponding author: magyna@chosun.ac.kr*

1. Introduction

Various events can occur in nuclear power plants (NPPs), and operators must identify them and take actions. If an accident occurs, the operator performs accident identification based on the symptoms presented in the procedure and takes actions to mitigate the accident with the contents specified in the procedure. However, the identification and actions are carried out within a short time, which can increase the operator's mental load and human error. This paper primarily aims to reduce human error by providing accident identification results to the operator using artificial intelligence (AI) techniques. However, from the information provided primarily, the AI can inform the operator of which accident occurred, but it is not known how the accident was diagnosed. This only raises questions about the diagnostic background, which makes AI completely unreliable. Therefore, in this paper, not only the accident diagnosis result but also the diagnosis background is provided to the operator through explainable artificial intelligence (XAI). Explaining the causal relationship to the diagnosis result to the operator is expected to increase the trust in AI. Besides, it is expected that it will contribute to reducing human error if the provided results are presented to the operator by configuring a more understandable interface.

2. Abnormal Diagnosis Algorithm

An algorithm was implemented to improve the reliability of the operator in AI by introducing XAI to the previously studied paper [1]. The algorithm was constructed by integrating a diagnostic module for abnormal diagnosis through AI and a verification module for increasing operator's reliability. The integrated algorithm is shown in Fig. 1, divided into Group 1 and Group 2, and each group represents a module. Also, the green color showed in Fig. 1 means the output for each step. Additionally, the task for each step and the used methodology are shown in Table I. To be more specific about algorithm, the task of step 1 is important because the diagnosis of AI is limited in the case of untrained data. Therefore, only the trained data in step 1 is advanced to the next step. In step 2, the abnormality of the data is diagnosed, and if it is in a normal state, no further diagnosis is made. In step 3, the scenario of abnormal data is determined. The diagnosed result is transferred to the verification module to improve the operator's reliability. Step 4 verifies whether the diagnosis succeeded or failed. Step 5 shows whether the expected symptoms were satisfied in the diagnosed scenario. Step 6 provides the basis for the diagnosis.

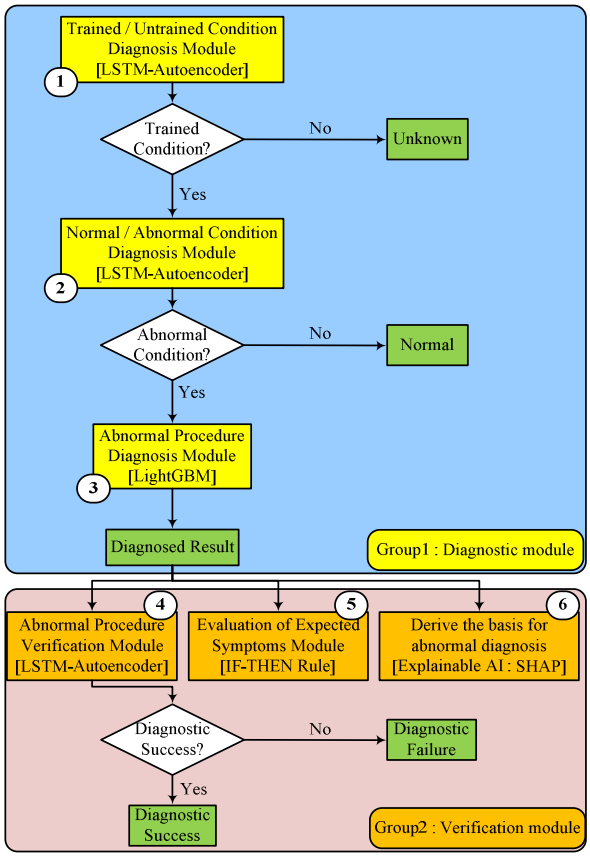

Fig. 1. Algorithm overview for abnormal diagnosis.

Table I: Task and method for each step in the algorithm.

Group	Step	Task	Methodology
Group1	1	Trained conditions or not	LSTM-Autoencoder
	2	Abnormal states or not	LSTM-Autoencoder
	3	Scenario diagnosis	LightGBM
Group2	4	Diagnosis successful or not	LSTM-Autoencoder
	5	Expected symptoms satisfied or not	IF-THEN rule
	6	Derive the basis for diagnosis	XAI : SHAP

3. Group1: Diagnostic Module

This section describes the diagnostic module in detail. The used AI methodology, data, and finally results are listed.

3.1 LSTM-Autoencoder

LSTM-Autoencoder is a combined model of LSTM for time series processing and an Autoencoder that copies the input to the output [2]. The model of LSTM-Autoencoder is shown in Fig. 2.

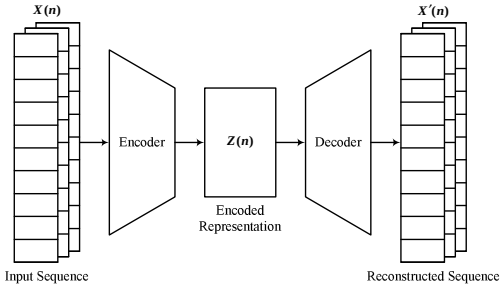

Fig. 2. LSTM-Autoencoder structure.

LSTM-Autoencoder performs good reconstruction on the trained data but the reconstruction fails on the untrained data. Using these characteristics, equation (1) is used to calculate the reconstruction error. In other words, the reconstruction error is low when trained data is input, and the reconstruction error is high when untrained data is input. In equation (1), x means real data, and x' means reconstructed data.

$$L(x-x') = \left\| x-x' \right\|^2 \tag{1}$$

The threshold is calculated with a 99.7% confidence interval based on the mean and standard deviation of the reconstruction error. Then binary classification is performed based on the calculated threshold. This methodology is used in steps 1, 2, and 4 in the algorithm.

3.2 LightGBM

LightGBM model is one of the decision tree methodologies, which is a machine learning technique with fast learning time and high performance [3]. The reason for high performance is that it uses a technique called gradient boosting decision tree to combine weak classifiers to create a strong classifier, which results in high accuracy. The reason for the fast learning time is that the existing decision tree series generates unneeded branch points in a level-wise algorithm, but the proposed model generates only the necessary points in a leaf-wise algorithm to have quick learning time. This methodology is used in step 3 in the algorithm.

3.3 Applied data for Abnormal Diagnosis

To obtain abnormal data from NPPs, a compact nuclear simulator that simulates Westinghouse-900 was used. Also, in order to determine the abnormal scenarios to be collected, the operational performance information system (OPIS) for nuclear power plant in which NPPs information is recorded was analyzed, and 20 scenarios were selected. The collected abnormal scenarios are shown in Table II.

Table II: List of collected abnormal scenarios.

Num.	Name of abnormal scenarios
Ab21-01	Pressurizer pressure channel failure (High)
Ab21-02	Pressurizer pressure channel failure (Low)
Ab20-01	Pressurizer level channel failure (High)
Ab20-04	Pressurizer level channel failure (Low)
Ab15-07	Steam generator level channel failure (High)
Ab15-08	Steam generator level channel failure (Low)
Ab63-04	Control rod fall
Ab63-02	Continuous insertion of control rod
Ab63-03	Continuous withdrawal of control rod
Ab21-12	Pressurizer PORV opening
Ab19-02	Pressurizer safety valve failure
Ab21-11	Pressurizer spray valve failed opening
Ab59-01	Charging pump failure stop
Ab80-02	Stopped 2/3 of the main feed water pump turbines
Ab64-03	Main steam line isolation
Ab60-02	Rupture of the front end of the regenerative heat exchanger
Ab23-03	Leakage from CVCS to RCS
Ab59-02	Leakage at the rear end of the charging flow control valve
Ab23-01	Leakage from CVCS to CCW
Ab23-06	Steam generator u-tube leakage

As input variables, 46 variables were extracted by analyzing the expected symptoms for each scenario. In addition, the data was normalized using a min-max scaler. Equation (2) was used as a min-max scaler.

$$std = \frac{x-\min(x)}{\max(x)-\min(x)}$$
$$scaled = std \times (\max(x)-\min(x)) + \min(x) \tag{2}$$

3.4 Results of each step in the Diagnostic Module

The results obtained from the abnormal diagnosis algorithm are as follows. Fig. 3 shows the results for step 1, which corresponds to whether or not it is in the trained conditions. In Fig. 3, the threshold is expressed as a black straight line. If it is below the threshold, it means a trained condition, and if it is above the threshold, it means an untrained condition. The model trained only 16 of the 20 scenarios collected. As a result, when an untrained scenario is given as an input value, it is shown that it is untrained condition beyond the threshold.

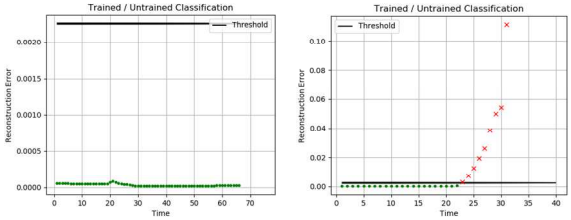

Fig. 3. Diagnosis result of trained conditions (left: trained condition, right: untrained condition).

Fig. 4, corresponding to step 2, indicates whether or not the data is abnormal, and classifies normal and abnormal based on the threshold as shown in Fig. 3. This model trained only normal data. As a result, when an abnormal scenario is given as an input value, it is shown that the abnormal condition exceeds the threshold.

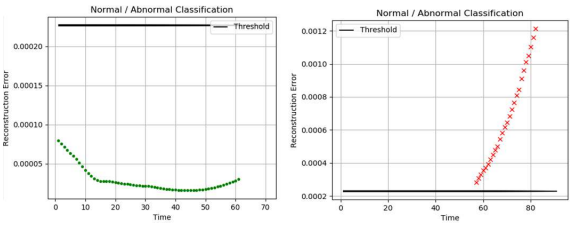

Fig. 4. Diagnosis result of abnormal states (left: normal condition, right: abnormal condition).

Fig. 5, corresponding to step 3, shows the scenario diagnosed with the LightGBM model. In Fig. 5, the pressurizer PORV opening scenario is given as an input value, and it can be seen that the corresponding scenario number Ab21-12 is correctly matched.

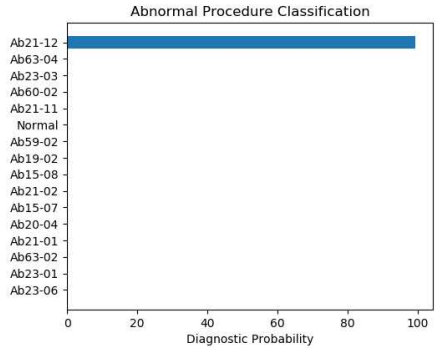

Fig. 5. Abnormal scenario diagnosis result.

4. Group2: Verification Module

This section describes the verification module in detail. It describes the results of step 4 in the verification module and the XAI methodology.

4.1 Results for step 4 in the Verification Module

In step 4, the LSTM-Autoencoder described in section 3.1 is used, and diagnosis failure or success is classified based on the threshold. At this time, there are 20 scenarios to be diagnosed, and 20 models corresponding to each diagnosis scenario are created. As a result, when the diagnosed result and the corresponding model are combined, the diagnosis is classified as success because it does not exceed the threshold, and when combined with the non-corresponding model, the diagnosis fails beyond the threshold. Fig. 6, corresponding to step 4, indicates whether or not the diagnosis is successful, and classifies success and failure based on the threshold as shown in Fig. 3.

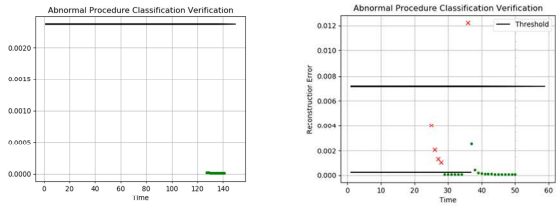

Fig. 6. Diagnosis result of whether diagnosis succeeds or not (left: diagnosis success, right: diagnosis failure).

4.2 Shapley additive exPlanations (SHAP)

In the verification module, the XAI technique was used to improve the reliability of AI. First, the core of XAI is the interpretability. More specifically, interpretability is the process of finding the rationale for why you should or shouldn't trust the model, why the model made a certain decision, and determine what outcomes are expected. Therefore, the algorithm used the Shapley additive exPlanations (SHAP) [4], one of the XAI methodologies, to derive the reason for making a specific decision. The SHAP methodology calculates the contribution of each variable by taking a specific value called the Shapley value. The Shapley value is represented by equation (3).

$$\phi_i(v) = \sum_{S \in N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!}(v(S \cup \{i\}) - v(S)) \qquad (3)$$

The definition of the variable used in equation (3) is shown in Table III below.

Table III: Definition of variable used in equation (3).

Variable	Describe
ϕ_i	Shapley value for i data
n	Total number of variables
S	All sets except i variable in the total group
$v(S)$	The contribution of the set excluding the i variable to the result
$v(S \cup \{i\})$	The contribution of the set containing the i variable (overall contribution)

That is, in equation (3), the contribution of the i variable is a value obtained by subtracting the sum of the contributions of the set excluding the i variable from the total set contribution. This methodology is used in step 6 of the verification module.

5. Interface Application

In order to more easily provide the result derived from the algorithm to the operator, an interface was constructed. The interface was designed by composing 4 panels, and each panel consists of diagnosis results, monitoring of major variables of the diagnosed scenario,

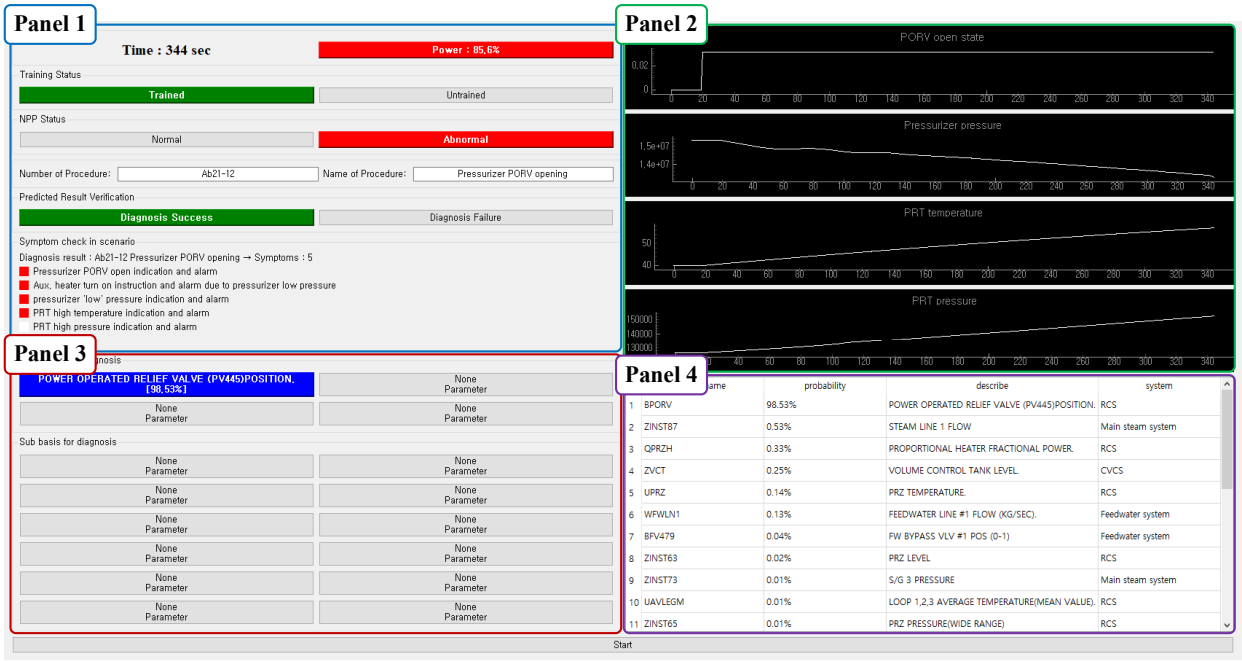

Fig. 7. Integrated abnormal diagnostic algorithm interface.

a list of major evidence for diagnosis, and detailed explanations of the diagnosis evidence. Fig. 7 shows each panel resulted from the pressurizer PORV opening scenario. First, panel 1 consists of steps 1 to 5 within the algorithm. Panel 1 shows the time and NPP's power in real-time in the first line. The second line corresponds to step 1, which expresses its status in an alarm format. The third line corresponds to step 2 and is similarly expressed in an alarm format. The fourth line corresponds to step 3 and displays the number and name of the diagnosed abnormal scenario. The fifth line corresponds to step 4 and is expressed in an alarm format. Finally, line 6 corresponds to step 5 and displays symptoms depending on the diagnosed scenario. If the symptom is satisfied, a red alarm is displayed. Second, panel 2 graphically represents the symptom variables presented in step 5. This allows operators to monitor variables related to the scenario in real-time. Thirdly, panel 3 is associated with step 6 and shows the calculated contribution (Shapley value). In panel 3, the upper part presents variables with a contribution of 10% or more, and the lower part shows variables with a contribution of 1% to 10%. Finally, panel 4 shows a table with a more detailed description of panel 3.

6. Conclusions

In this paper, an abnormal diagnosis algorithm was constructed to support the operator when abnormal conditions occur. This will reduce the human error and operator's mental load by providing useful information to the operator using artificial intelligence (AI) for abnormal in the nuclear power plant. Besides, the reliability of the operator for AI can also be improved by providing diagnostic evidence using explainable artificial intelligence (XAI). Additionally, the interface was designed to make it easier to present all the results of the algorithm. If the operator is provided with evidence for the diagnosis result derived from AI, it is expected that the operators can diagnose rapidly accidents of nuclear power plants based on the evidence and take appropriate actions.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (Grant No. NRF-2018M2B2B1065651).

REFERENCES

[1] H. J. Kim, J. H. Kim, Algorithm of Abnormal Event Diagnosis with the Identification of Unknown Events and Output Confirmation, Transactions of the Korean Nuclear Society Virtual Spring Meeting July 9-10, 2020.
[2] Ali H. Mirza, Selin Cosan, Computer network intrusion detection using sequential LSTM neural networks autoencoders, In:2018 26th signal processing and communications applications conference (SIU), IEEE, 2018. p. 1-4.
[3] Guolin Ke et al, LightGBM: A highly efficient gradient boosting decision tree, In: Advances in neural information processing systems, 2017, p. 3146-3154.
[4] Scott M. Lundberg, Gabriel G. Erion, Su-In Lee, Consistent individualized feature attribution for tree ensembles, arXiv preprint arXiv:1802.03888, 2018.