Anomalies Detection by Unsupervised Learning Using Explainable Artificial Intelligence in Nuclear Power Plants

Sang Won Oh, Hye Seon Jo, Ho Jun Lee, and Man Gyun Na*
Dept. of Nuclear Engineering, Chosun Univ., 309 Pilmun-daero, Dong-gu, Gwangju, 61452
**Corresponding author: magyna@chosun.ac.kr*

1. Introduction

In nuclear power plants (NPPs), failures and accidents can occur due to various causes, such as equipment failure, electrical and instrumental errors, and human errors. In an accident, the NPPs operator should identify the cause of the accident, select an appropriate procedure, and implement mitigation measures; this is called a diagnostic task. However, since many monitoring factors to be judged by the operator change rapidly in an accident situation, the urgent situation that appropriate actions should be taken within a short time causes the human errors of the operator. Human errors, such as misjudgment by the operator, lead to failure of accident mitigation measures and can lead to serious accidents beyond the control area of the NPPs.

Recently, many studies have been conducted to assist operators through anomaly detection and diagnosis using artificial intelligence (AI). Most of the studies are conducted mainly through supervised and unsupervised learning. The supervised learning is a machine learning method that is trained by labeling information about each accident. That is, the supervised learning is a method that provides AI with a problem (accident data) and answer (label) so that the AI can learn the correct answer. Unlike the supervised learning, the unsupervised learning is a method of finding and learning data features based on input data without data labeling when applying an AI model. Since NPPs contain more than about 200 abnormal operating procedures, it is limited in labeling all situations [1]. In addition, the supervised learning has a problem in that it is impossible to diagnose unlearned data.

Therefore, this study developed an anomaly detection algorithm for NPPs using the unsupervised learning. Autoencoder (AE), which is one of the unsupervised learning methods, was used for algorithm development, and the application of the explainable AI (XAI) method was considered to provide reasons for AI judgment results. The Shapley additive explanation (SHAP) method was used to the application of XAI.

The application of XAI has the following advantages. 1) It can build trust between operator and AI. The AI only provides an opinion for judgment. The decision on the operation of a NPPs is solely determined by the operator. Therefore, the absence of explanations of AI opinions can make it more difficult for operator to judge. Reliable information from XAI can help operator make decisions. 2) It helps to improve the performance of the AI model. Too many or unnecessary variables negatively affect model learning. Through XAI, variables that do not contribute to learning can be found, and the performance of the AI model can be improved by solving these problems. Therefore, the anomaly detection model using XAI in this study is expected to show high reliability and high performance.

2. Methods

2.1 Autoencoder

AE was first introduced in the 1980s and has been widely used in deep architectures [2]. This method is an unsupervised learning-based neural network that is trained to generate the same target value (X') as the input data (X). The structure of AE is shown in Fig. 1. AE is composed of the input layer, encoder, latent vector, decoder, and output layer. Also, it has a symmetrical structure. The input data that enter the input layer is compressed through dimensionality reduction through an encoder. The compressed value is restored to the same dimension as the input data through the decoder. This process is expressed in Eq. (1).

$$z = h(X) = W_e X + b_e$$
$$X' = g(h(X)) = W_d z + b_d \quad (1)$$

where W and b are the weight and bias, respectively.

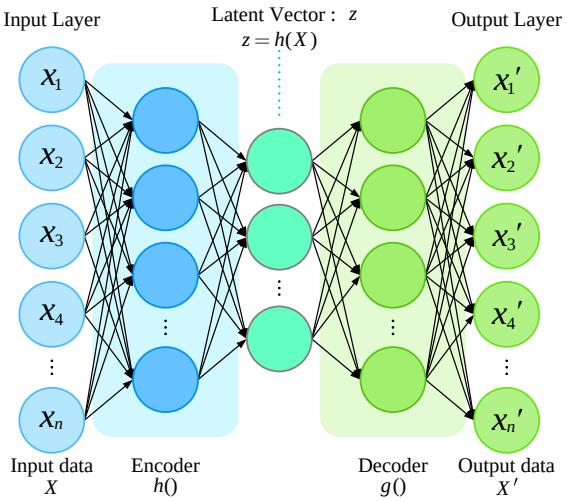

Fig. 1. Overview of the AE structure.

The restored data is not completely reconstructed into the same data as the input data. Therefore, a difference is generated between the reconstructed data and the

input data; here, this residual is calculated as a reconstruction error (RE). Eq. (2) shows how to calculate RE.

$$RE(X, X') = \left\| X - X' \right\|^2 \tag{2}$$

2.2 Shapley additive explanation

SHAP is a method of calculating the Shap value, which is the importance of each variable, using Lloyd Stowell's Shapley value [3]. The Shapley value refers to the contribution that each variable affects the model result. The Shapley value is represented by Eq. (3). Variables used in Eq. (3) are shown in Table I. The calculated Shapley values appear as variables contributing and offsetting to AI results.

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

Table I: Explanation of variables used in Eq. (3)

Variable	Description
ϕ_i	Shapley value for i data
n	Total number of variables
S	All set except i variable in 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

2.3 Optimization of the anomaly detection model

Optimization of the anomaly detection model is performed to improve the performance of the model and prevent overfitting problem. In this study, AE was used as an anomaly detection model, and the model structure and hyperparameters were set for optimization. Hyperparameters were set through several trials using the grid search method. The grid search method is to find hyperparameters with the highest performance after sequentially inputting values that can be put into model hyperparameters. The hyperparameter information set through the grid search is shown in Table II.

Table II: Hyperparameter information used in AE

Layer	Batch size	Activation	Optimizer	Loss
5	64	ReLU	Adam	Mean squared error

Additionally, an early stopping was used for optimal model training. The early stopping is a function to prevent overfitting. That is, when the loss function value for the validation data does not fall below the optimal value by more than the number of patience during model training, training is terminated. In this study, the mean squared error was used as the loss function, and the maximum epoch and patience were set to 500 and 20, respectively.

3. Data collection and pre-processing

In this study, data obtained through the compact nuclear simulator (CNS) were used. The CNS is a simulator designed based on the Westinghouse 3-loop pressurized water reactor. Through the CNS, normal data for anomaly detection model training and abnormal situation data for testing were collected. Since the CNS has a limitation in collecting many normal data, additional data were collected by applying noise. As a result, 64,000 normal data were obtained through the application of noise.

The collected data consists of 2,222 variables. Among them, in this study, variables corresponding to systems and components were divided to detect system anomalies. The variable information of the component corresponding to each system is shown in Table III.

Table III: System and component variable information

System	Component	No. of variables
Reactor coolant system	- Reactor - Pressurizer - Temperature control - Flow control - Reactor coolant pump	63
Chemical and volume control system	- Pump - Valve - Flow control - Temperature control - Volume control tank - Heater	30
Main steam and turbine system	- Turbine - Valve - Flow control	22

4. Experiment

To verify the results of this study, it is necessary to confirm the symptoms of the actual accident scenario. The scenario used for verification is the data in which the pressurizer (PZR) spray valve opens abruptly by 33% due to a failure. When the spray valve opens suddenly, the following symptoms occur; 1) the spray flow increases 2) the pressure of the PZR decreases 3) the heater operates to increase the pressure of the PZR 4) the charging flow increases to supply the spray water resource. Section 4.3 compares the results with the actual accident symptoms.

4.1 System anomaly detection using AE

To evaluate the AE model in this study, anomaly detection through the PZR spray valve failure "open" scenario was considered. The anomaly detection process is shown in Fig. 2. When abnormal data enters the AE model that has been trained on normal data, a RE is generated. If the RE is higher than the threshold, it is judged as an abnormal occurrence. In this study, the threshold was set using the 3-sigma rule with a reliability of 99.7%. A confidence interval of 99.7% means a reconstruction failure of 0.3%. Threshold calculation is expressed by Eq. (4).

$$Threshold = \mu \pm 3\sigma \tag{4}$$

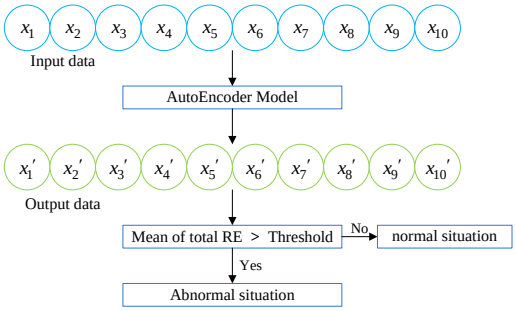

Fig. 2. Schematic explanation of the anomaly detection process.

When an outlier occurs, it is possible to check whether an anomaly is detected for each system. Fig. 3 shows the result of detecting the anomalies in the PZR spray valve open accident. In Fig. 3, the black line is the malfunction injection time (30sec) of the accident. In addition, the blue line is a threshold to distinguish the normal state, and if RE exists below the threshold, it is a normal condition (green color), and if RE exists above the threshold, anomaly condition (red color) is displayed. As a result of the test, a system anomaly was detected immediately by reactor coolant system (RCS) after malfunction injection, and an abnormality was detected before trip occurred in chemical volume control system (CVCS) and main steam / turbine system (MSTS).

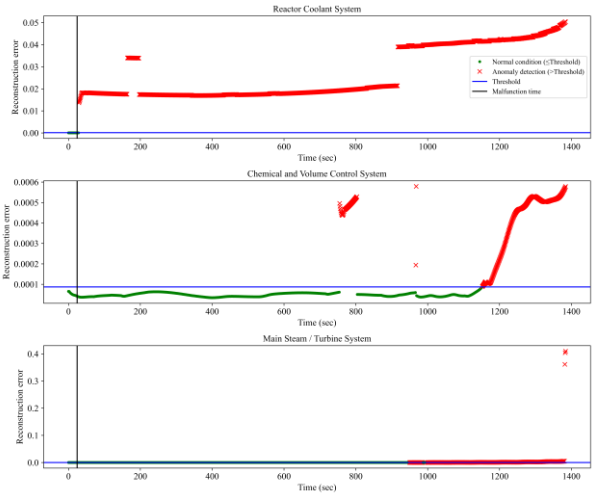

Fig. 3. PZR spray failure "open" anomaly detection test result.

4.2 Component Anomaly detection using relative error

By comparing the RE of the verification scenario, the variables most affecting the abnormal situation were collected. Each variable in CNS data has the characteristic of having a different distribution. Therefore, the comparison of RE was performed by calculating relative error. Relative error calculation is expressed by Eq. (5).

$$\text{Relative error} = \frac{|\text{actual value - reconstructed value}|}{\text{actual value}} \tag{5}$$

Table IV shows the 5 variables with the largest relative error. From the result of relative error in the table, it can be seen that the anomaly of the PZR-related variables is large. That is, it means detecting anomalies in the PZR.

Table IV: Relative error of the top 5 variables

Variable	**Description**	**Relative error (%)**
ZINST66	PZR spray flow	218.37
QPRZH	Proportional heater power	100
UCHGUT	Charging line outlet temp	40.7
BPSV10	Aux. PZR spray valve position	11.067
WCHGNO	Charging flow	10.73

4.3 Anomaly detection explanation using SHAP

Since the PZR spray flow has the largest relative error (218.37%), it is considered that it has the greatest influence on anomaly detection. Fig. 4 shows the results of analysis using SHAP for the PZR spray flow, which had the greatest influence on detecting the anomaly.

This result is explained as follows; 1) as the basis for judging that the spray flow is large, the proportional heater power, the charging line outlet temperature, the charging flow, etc. are shown. 2) as the basis for offsetting, spray flow is indicated.

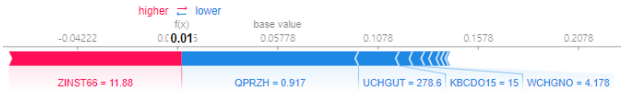

Fig. 4. The results of analysis using SHAP for the PZR spray flow.

5. Conclusions

In this paper, anomaly detection of NPPs was performed using AE, an unsupervised learning model. In addition, reliable AI results were confirmed using SHAP. In this way, it is possible to provide the operator with reliable information about the abnormal situation of each system of the NPPs. Additionally, it was possible to judge which component was abnormal by finding the variable that had the greatest influence on the abnormal situation.

In this study, only RCS, CVCS, and MSTS were used to detect anomalies. In future work, we plan to perform anomaly detection work targeting all systems. In addition, we plan to utilize various AI and XAI methods to improve performance and interpretability.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant, funded by the Korean Government (MSIT) (Grant No. NRF-2018M2B2B106565123).

REFERENCES

[1] Y. G, Kim, D. S. Park, Consideration on the use of Explainable AI in Operator Support System, Proceedings of the Korean Nuclear Society Virtual Autumn Meeting, Dec. 17-18, 2020.
[2] M. Sakurada, T. Yairi, Anomaly detection using autoencoders with non-linear dimensionality reduction, in: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, ACM, p.4, 2014.
[3] L. S. Shapley, A. E. Roth, The Shapley Value: Essays in Honor of Lloyd S. Shapley, Cambridge University Press, 1988.