Application of SHAP Method for Explanation of Abnormal State Diagnosis Model Based on Trend-images in Nuclear Power Plants

Sang Won Oh, Ji Hun Park, and Man Gyun Na*
Dept. of Nuclear Engineering, Chosun Univ., 10, Chosundae 1-gil, Dong-gu, Gwangju, 61452
**Corresponding author: magyna@chosun.ac.kr*

1. Introduction

A variety of abnormal states exist in nuclear power plants (NPPs) due to various causes. In the event of an abnormal state, economic loss and safety can be threatened. The operator of NPPs should identify many monitoring factors to perform diagnosis and action to prevent the deterioration of the abnormal condition. Diagnostic work is carried out in an urgent state, which can cause a human error of the operator.

Recently, many studies have been conducted to support operator decisions by applying artificial intelligence (AI) to the accident diagnosis of NPPs to reduce human error [1]. Many studies have led to the development of high-performance AI, which shows high accuracy in diagnosis. However, AI does not always provide the correct answer, and the operator is responsible for the incorrect diagnosis and action. Therefore, reliability is an important issue in AI application.

In various fields, the development of explainable AI (XAI) to increase the reliability of AI is actively progressing. However, XAI is being developed mainly in the field of the image. That is, there are limitations in its application to time-series data of NPPs [2]. Accordingly, in this study, time-series data of NPPs are converted into images. There are many ways to convert time-series data into image (e.g., recurrence plot, Markov transition field, etc.). These ways are effective in converting time-series characteristics but are difficult to interpret. Therefore, it is converted into an image in the form of a trend plot for intuitive interpretation. Additionally, an abnormal diagnosis model is developed based on the converted image data. After that, the shapely additive explanation (SHAP) of the XAI method is applied to the model to explain the results of AI.

This study proposes to provide the operator with the focus of AI (i.e., the part that AI focuses on making decisions) through the application of XAI.

2. Methods

2.1 Convolutional Neural Network

Convolutional neural network (CNN) was used as the abnormal diagnosis model. CNN is an effective model for extracting features from images and classifying them by class [3]. Several convolution operations and pooling generate features representing patterns in the image. After that, the features are converted into one-dimensional data in the flatten layer and input to the fully connected layer. In a fully connected layer, classification is performed through the activation function softmax. Fig.1 shows the structure of the CNN used as a diagnostic model in this study.

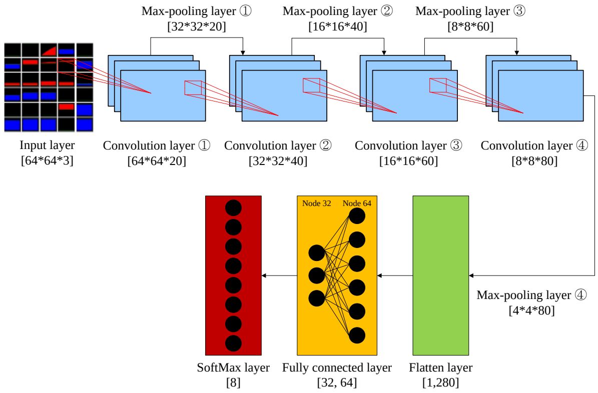

Fig. 1. Overview of CNN structure and hyperparameters used as diagnostic model.

2.2 Shapely Additive Explanation

SHAP is a method that aims to explain the output of a model using shapely values [4]. The Shapely value is the importance value of each feature calculated considering all feature combinations that contribute to the model output. In other words, it calculates the shapely value of each pixel, providing information that contributed to the model's output (i.e., decision). The shapely value ϕ of the feature i is represented by Eq. (1). Parameters information used in Eq. (1) is shown in Table I.

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

Table I: Shapely value parameter description

Parameter	Description
ϕ_i	Shapley value for i data
n	Total number of features
P	All set except i feature in total group
$f(P)$	The contribution of the set excluding the i feature to the result
$f(P \cup \{i\})$	The contribution of the set containing the i feature

3. Generate Image Data

3.1 Data Collection and Pre-Processing

In this study, the training and test data of the NPPs abnormal state diagnosis model were collected through the compact nuclear simulator (CNS). The CNS is a simulator designed based on the Westinghouse 3-loop pressurized water reactor. Among the collected data, 30 parameters are selected through correlation analysis. Additionally, a min-max normalization is applied to normalize the collected data between 0 and 1. The selected parameters are listed in Table II.

Table II: Selected input variables list

No.	Description
1	PRZ pressure safety valve opening state
2	HV6 valve opening state
3	PRZ spray valve opening state
4	PRZ spray flow
5	PORV opening state
6	PRZ level (channel)
7	Normalized PRZ level (process)
8	PRZ pressure (channel)
9	PRZ pressure (process)
10	PRZ temperature
11	PRT pressure (channel)
12	PRT pressure
13	PRT temperature
14	PRT water level
15	Back-up heaters power
16	Proportional heaters power
17	VCT pressure
18	VCT level
19	Containment radiation
20	Containment pressure
21	Containment sump level
22	Containment relative humidity
23	Containment temperature
24	Charging flow
25	Letdown flow
26	Charging line outlet temperature
27	45 GPM orifice valve opening state
28	60 GPM orifice valve opening state
29	75 GPM orifice valve opening state
30	Letdown isolation valve opening state

* PRZ: pressurizer
* PORV: power operated relief valve
* PRT: pressure relief tank
* VCT: volume control tank
* GPM: gallon per minute

3.2 Image Generation

Each variable in the data is converted into an image that plots the trend. The trend images represent the data change of 20 second period as a plot as shown in Fig. 2. Additionally, the plot shows the difference between the steady state and the current state by dividing the two regions based on the average of the normal value. In the image, the green line represents the average of the normal values of the corresponding variables in the NPPs. The red and blue areas represent a higher part than normal and a lower part than normal, respectively.

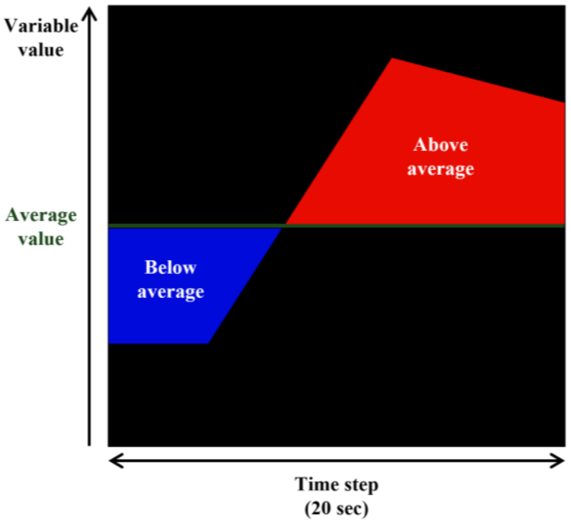

Fig. 2. Example of trend plot image.

The converted image used the form of subplots so that the states of all variables could be included in one image. Fig.3 shows an example of an image, one of the images used as training and test data. Each variable depicted in Table II is positioned to match the number shown in Fig. 3.

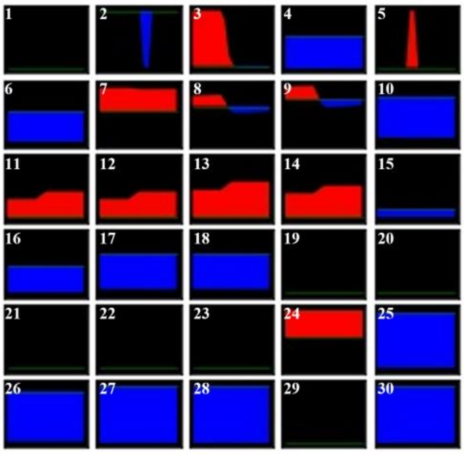

Fig. 3. Example of input image using subplot.

4. Results

4.1 Diagnosis Result of CNN Model

As shown in Table III, this study classifies eight abnormal and normal states related to instrument error, component status abnormality, and pipe leakage.

Diagnostic results are evaluated for accuracy using a confusion matrix. The confusion matrix is a table that shows the comparison of the model's predicted results with the actual true answers. Fig. 4 shows the confusion matrix of the diagnosis model test result. Accuracy is calculated through Eq. (2), and the test accuracy showed about 99.9%. There were some diagnostic failures.

Table III: Collected data and scenario information

No.	Scenario name	No. of train data	No. of test data
1	Normal	7,848	400
2	PRZ spray valve failure "open"	38,185	8,934
3	PRZ pressure channel failure 'high'	2,242	297
4	PRZ level channel failure "low"	7,952	9,090
5	PRZ level channel failure "high"	8,855	1,771
6	PRZ PORV open	26,034	3,659
7	PRZ pressure safety valve failure	21,155	3,598
8	Leakage from RCS to CCW	1,018	169
* RCS: reactor coolant system * CCW: component cooling water			

$$\text{Accuracy} = \frac{\text{number of test data predicted correctly}}{\text{number of total test data}} \quad (2)$$

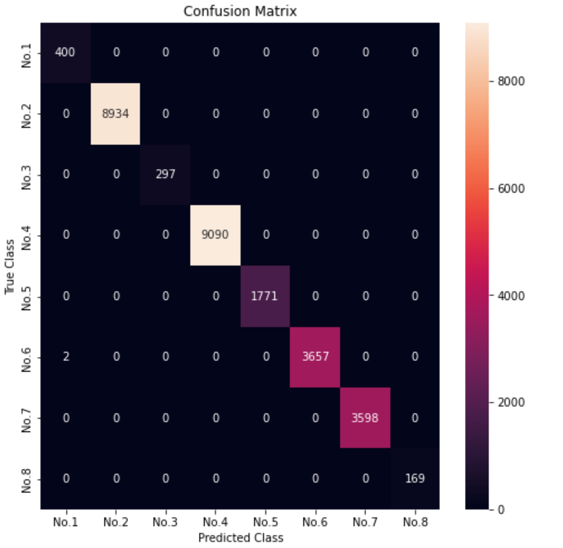

Fig. 4. A confusion matrix representing the test results of the diagnostic model.

4.2 Application of SAHP

XAI results utilize the DeepExplainer method in SHAP, which is effective for deep learning model. DeepExplainer visualizes the positive and negative effects of all output from one input. The features that have a positive effect on the diagnosis are shown in red, and the features that have a negative effect on the diagnosis are shown in blue. Fig. 5 shows the diagnosis result of the 'leakage from RCS to CCW' scenario (i.e., No.8 scenario). AI focuses on the shape of the following variables in the outcome of the No.8 scenario diagnosis; 1) containment radiation and temperature and sump level increase 2) reduced PRZ safety valve opening state.

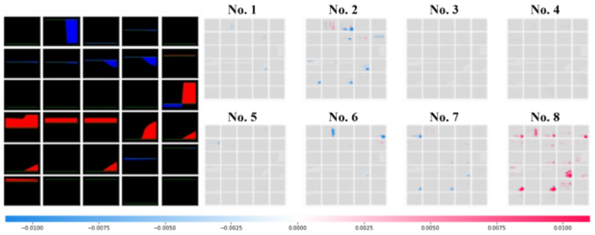

Fig. 5. SHAP application result for 'leakage from RCS to CCW' scenario.

5. Conclusions

In this study, the application of XAI was proposed for the reliable diagnosis of abnormal states in NPPs. For the application of XAI, multivariate time-series data of NPPs were converted into images, and CNN was applied for diagnosis using image data. The diagnostic task was tested by classifying 8 scenarios and showed high accuracy. In addition, it showed the focus of AI in the diagnosis process and the reliable diagnosis results.

In future work, we plan to analyze the focus of AI and identify more scenarios through trend image optimization.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant, funded by the Korean Government (MSIT) (Grant No. NRF-2018M2B2B1065651).

REFERENCES

[1] D. Lee, AM Arigi, and J. Kim, Algorithm for Autonomous Power-increase Operation Using Deep Reinforcement Learning and A Rule-based System. IEEE Access, Vol. 8, pp. 196727-196746, 2020.
[2] R. Saluja, A. Malhi, S. Knapič, K. Främling, and C. Cavdar, Towards a Rigorous Evaluation of Explainability for Multivariate Time Series, arXiv:2104.04075, 2021.

[3] R. Yamashita, M. Nishio, R. K. G. Do, and k. Togashi, Convolutional neural networks: an overview and application in radiology, Insights Imaging, Vol.9, pp. 611-629, 2018.

[4] S. M. Lundberg, and S. I. Lee, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, pp.4765-4774, 2017.