Abnormal State Diagnosis and Operator-centered Explanation Using XAI-based Deep Learning

Young Do Koo [a,b], Sa Kil Kim [a], Seung Geun Kim [a], Yonggyun Yu [a], Man Gyun Na [b*]
[a]*Korea Atomic Energy Research Institute, 989-111 Daedeok-daero, Yuseong-gu, Daejeon, Republic of Korea 34057*
[b]*Department of Nuclear Engineering, Chosun University, 309 Pilmun-daero, Dong-gu, Gwangju, Republic of Korea 61452*
[*]*Corresponding author: magyna@chosun.ac.kr*

***Keywords**: operator-centered explanation, procedure-based rationale, XAI-based deep learning

1. Introduction

Explainable artificial intelligence (XAI) is introduced to resolve the black-box characteristics of artificial intelligence (AI), and to mitigate trade-off between learning performance and explainability of machine learning including deep learning [1]. The need that explainability of XAI should attain a level that end-users can understand or more than a higher level as well as that end-users can interpret AI [2] is recently suggested. Therefore, the studies that explanation from XAI can be user- or human-centered was carried out [3,4]. Here, the meaning of the term 'user/human-centered' is similar to comprehensible or informativeness for end-users (e.g., operator, regulator, or stakeholders).

In a nuclear power plant (NPP) field, XAI is applied to an operator decision-making support technology based on AI, especially, machine learning including deep learning, in order to enhance its usability and reliability. However, it seems to be focused on interpretability of an AI model rather than explainability for operator as an end-user until now. Hence, the study for explanation from the viewpoint of an operator using XAI is necessary.

This study is performed to develop an XAI-based deep learning method providing explanation easily comprehensible to an operator for usability of AI-based operator decision-making support technologies. In the study, operator-centered explanation is defined and adopted. In addition, the model for NPP abnormal state diagnosis and procedure-based rationale is developed using XAI-based deep learning method. Its result is compared with that of a model in a conventional way.

2. Operator-centered explanation

Operator-centered explanation refers to not only interpretable information (i.e., rationale for decision-making of AI) but also comprehensible information (i.e., compatible to mental model) to an operator in an NPP. Further, operator-centered explanation is intended to be toward effectiveness evaluation dimensions related to end-users (e.g., comprehensibility, user satisfaction, and so on [5]) in aspect of XAI field.

In the study, a procedure-based variable is selected as operator-centered explanation for usability of an AI-based operator decision-making support technology. Operators in an NPP normally perform rule-based tasks checking variables described in an operating procedure. Hence, procedure-based rationale of decision-making support information from XAI is able to be comprehensible as well as interpretable in aspect of operators' tasks.

3. Methodology

3.1 CNN with Grad-CAM

Convolutional neural network (CNN) based on gradient-weighted class activation mapping (Grad-CAM) [6] is used to provide high diagnosis accuracy and operator-centered explanation to an operator.

In the NPP field, it is essential that the performance of AI technologies for operator decision-making support (e.g., anomaly detection, state diagnosis or prediction/forecasting) is similar to or better than a human-level. Therefore, deep learning is usually used on account of its higher learning performance despite of its lower explainability. Dilated causal convolutional neural network (DCCNN) [7] is used to diagnose abnormal states in an NPP in the study. Dilated convolution is applied to more extract features from inputs using receptive filter while keeping its filter size. Causal convolution makes convolution layers suitable for time-series analysis.

Rationale for a result of abnormal state diagnosis from DCCNN is visualized using Grad-CAM. Grad-CAM calculates the weights (i.e., importance for prediction) using gradient from the last layer of DCCNN, and then importance is emphasized using the color according to its scale.

3.2 Ensemble Learning

Performance of deep learning is generally higher when more inputs useful to learning are applied. However, high performance cannot be ensured when procedure-based variables are applied as inputs for only explainability. In contrast, explainability of procedure-based explanation is expected to decrease in the event that all the useful variables are applied to deep learning for performance. Hence, stacking of ensemble learning is used to achieve both high accuracy and comprehensible explainability.

Ensemble learning is the method that outputs from multiple weak learner are combined to make a strong learner of which performance is better than an individual weak learner. Stacking is a way of ensemble learning making a meta model learned using outputs from sub models to produce the final output.

3.3 Application Data

The data of NPP abnormal states is acquired by simulating abnormal state scenarios using compact nuclear simulator (CNS) developed by Korea Atomic Energy Research Institute. Twelve scenarios of NPP abnormal states are constructed based on abnormal operating procedures (AOPs).

382 simulated data are used to develop a Grad-CAM-based CNN model, and separated into training, verification, and test datasets for cross-validation. 41 variables of the simulated data are used as input features for developing a model. All the input features for a model are procedure-based variables which are available in CNS among variables described in twelve AOPs. Input features are also used as procedure-based rationale by XAI. The simulated data consist of a normal state followed by an abnormal state.

4. Results

4.1 Developed Model

The developed model consists of twelve sub models and one meta model. Each sub model is trained using the simulated data with input features corresponding to each AOP. Sub models diagnose thirteen abnormal states (i.e., one normal state and twelve abnormal states). Meta model is trained using outputs from sub models. Meta model diagnose thirteen NPP states, which become the final result.

According to the result of abnormal state diagnosis from a meta model, Grad-CAM is applied to the last convolution layer of the corresponding sub model to provide rationale calculating the importance of input features, and then to visualize that. All sub models have the same DCCNN structure.

4.2 Performance of Abnormal State Diagnosis

The developed model has shown accuracy of 0.944 for abnormal state diagnosis when applying the test dataset. The confusion matrix of abnormal state diagnosis is shown in Fig. 1.

4.3 Visualization of Procedure-based Rationale

Visualization of procedure-based rationale by Grad-CAM is presented in Fig. 2 for diagnosis result of one ab60-02 case. That is, the importance of each variable applied to the corresponding sub model over time is expressed. Here, each variable applied to the corresponding sub model is stated in the abnormal-60-02 procedure.

The color on each procedure-based variable differs depending on the importance scale. If the importance scale gets lower, the color become brighter.

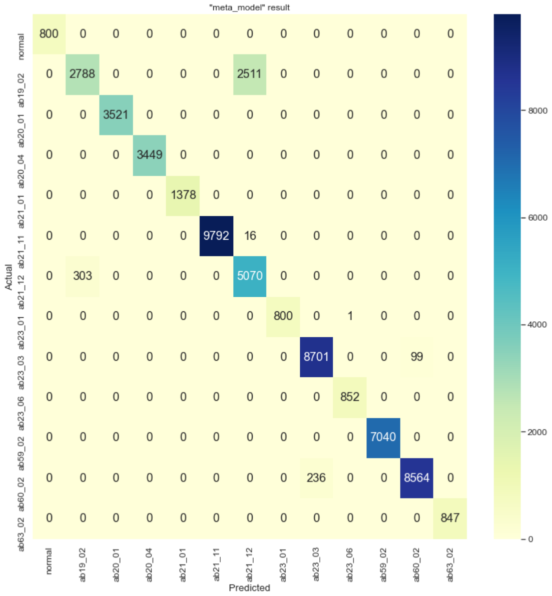

Fig. 1. Confusion matrix for abnormal state diagnosis of developed model

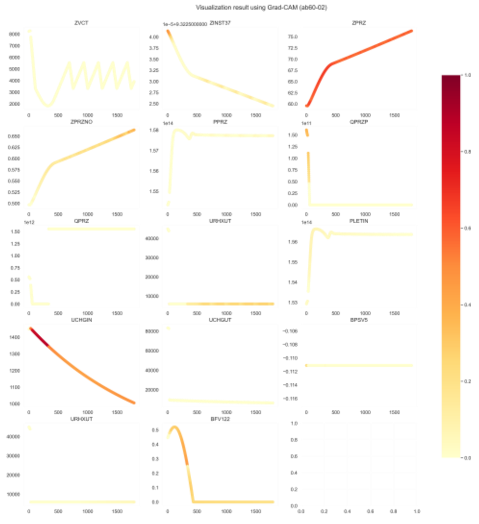

Fig. 2. Visualization explanation of developed model (in case of ab60-02).

4.4 Comparison with Conventional Model

The accuracy and explanation of the developed model are compared with those of a conventional model. The conventional model refers to a model where multiple variables are applied at once. In the study, the conventional model is established applying the same simulated data with 41 procedure variables to the network structure which is the same as sub models of the developed model.

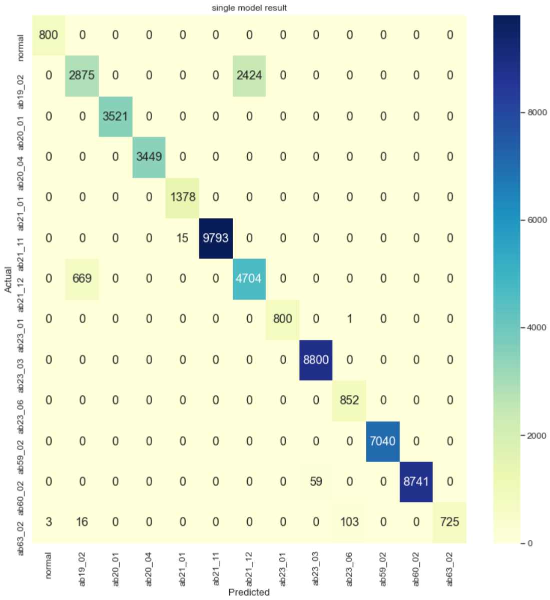

Fig. 3. Confusion matrix for abnormal state diagnosis of conventional model.

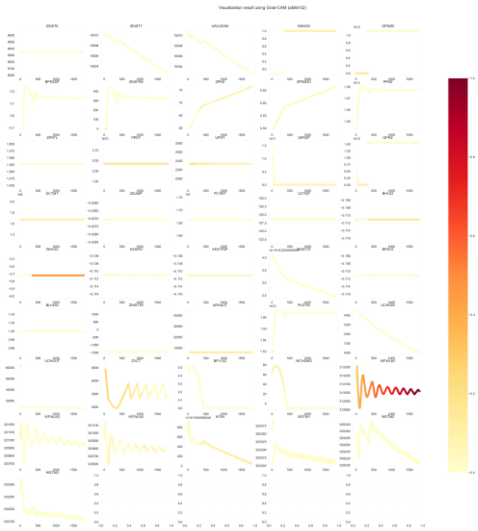

Fig. 4. Visualization explanation of conventional model (in case of ab60-02).

Accuracy of 0.942 is shown from the conventional model. The confusion matrix of abnormal state diagnosis using the conventional model is shown in Fig. 3. Fig. 4 shows visualized explanation of the conventional model using Grad-CAM for diagnosis result of one ab60-02 case. In Fig. 4, the importance of all variables applied to the conventional model over time is expressed. However, the highlighted variables ‘WHV22’ and ‘WFWLN1’ are not indicated in the abnormal-60-02 procedure.

The accuracy of both models is nearly equivalent. However, procedure-based rationale of the developed model seems to be more useful than the conventional model since the importance of input features applied is presented within a range of an AOP. That is, it is more compatible to an operator’s task.

5. Conclusions

The model providing explanation comprehensible from the viewpoint of an operator is developed using XAI-based deep learning method. The accuracy and explainability of the developed model are compared with those of a conventional model. Consequently, the developed model is able to provide operator-centered explanation which is relatively more comprehensible to operator than the conventional model without any accuracy loss in NPP abnormal states.

ACKNOWLEDGEMENT

This research was supported by a grant from the Korea Atomic Energy Research Institute (KAERI) R&D Program (No. KAERI-524450-23) and National Research Foundation (NRF) of Korea grant funded by the Korean Government (MSIT) (No. NRF-2018M2B2B1065651).

REFERENCES

[1] D. Gunning and D. W. Aha, DARPA’s Explainable Artificial Intelligence (XAI) Program, AI magazine, Vol.40, pp.48-58, 2019.

[2] S. Mohseni, N. Zarei, and E. D. Ragan, A Survey of Evaluation Methods and Measures for Interpretable Machine Learning, arXiv:1811.11839, 2018.

[3] T. Kaluarachchi, A. Reis, and S. Nanayakkara, A Review of Recent Deep Learning Approaches in Human-Centered Machine Learning, Sensors, Vol.21, 2514, 2021.

[4] Q. Liao and K. R, Varshney, Human-Centered Explainable AI (XAI): From Algorithms to User Experiences, arXiv:2110.10790, 2021.

[5] A. B. Arrieta et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Vol.58, pp.82-115, 2020.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, 2017 IEEE

International Conference on Computer Vision (ICCV), pp.618-626, 2017, Venice, Italy.
[7] A. Oord et al., WaveNet: A Generative Model for Raw Audio, arXiv:1609.03499, 2016.