Study on Quantitative Comparison between Explainable Artificial Intelligence Methods for Nuclear Power Plant Applications

Seung Geun Kim [a*], Seunghyoung Ryu [a], Hyeonmin Kim [b], Kyungho Jin [b]
[a]Applied Artificial Intelligence Section/[b]Risk Assessment Research Division, Korea Atomic Energy Research Institute,
111, Daedeok-daero,989beon-gil, Yuseong-gu, Daejeon, South Korea, 34057
[*]Corresponding author: sgkim92@kaeri.re.kr

***Keywords :** nuclear power plant, explainable artificial intelligence, perturbation analysis

1. Introduction

Artificial intelligence(AI) technology is actively applied not only for the computer science fields but also for the other engineering fields. In nuclear engineering field, AI models are adopted for various purposes including: signal validation/reconstruction[1, 2], event diagnosis[3], trend prediction[4], and automation/autonomous operation[5]. Most of these studies are based on deep neural network(DNN)-based models, since DNN models have shown best performances all along the data-driven problems.

Although DNN models are showing high performance, their practical applications on critical domains such as nuclear or medical are hindered due to its low explainability. In general, low explainability means that the relation between model's inputs and outputs are not explicitly revealed. Low explainability may negatively affects the overall trustworthiness of the model, leading to the decreased public acceptance and feasibility of practical applications.

To enhance the explainability and trustworthiness of DNN models, various explainable AI(XAI) methods have been proposed[6]. Most of existing XAI methods are developed and verified based on image or natural language data. Although many XAI methods can be applied to the time-series data, studies on their effectiveness and validity are still insufficient. Since nuclear field is one of the representative safety-critical domain and AI application on nuclear field mostly based on time-series data, studies on XAI methods that effectively work for time-series data are essential.

Another problem during the application of XAI method is that, it is difficult to determine which XAI method is better. This problem is due to the inexistence of quantitative standards on 'good explanation'. Accordingly, many studies have been relied on human or domain experts for qualitative explanation evaluation, which could be ambiguous and biased to the conventional knowledge or common sense. Especially, since time-series data is relatively more difficult for intuitive interpretation, relying on human evaluation is even more infeasible.

In this study, we investigated about the proper application methods of perturbation analysis, which enables the quantitative comparison between XAI methods that deduces relative importance of input elements(i.e. attribution score) for deducing specific output. During the perturbation analysis, values of input elements are changed into pre-defined value(i.e. perturbing value) in the order of importance, and corresponding output changes are observed. As there are only few studies on XAI and perturbation analysis applications for time-series data, we developed time-series data-based accident scenario classification model and used it for the further experiments.

Configuration of this paper is as follows. In section 2, description about the overall methods and experiment results will be presented including the development of accident scenario classification model, application of XAI method, application of perturbation analysis, and method for selecting perturbing value. Section 3 will summarize and conclude the paper.

2. Methods and Results

2.1 Development of Accident Scenario Classification Model

The developed accident scenario classification model receives 900 seconds and 19 kinds of instrumentation signals as input (with 17,100-length vector) and deduces classification probabilities corresponding to SGTR(steam generator tube rupture)- and MSLB(main steam line break)-based five accident scenarios (SGTR/A, SGTR/2A, SGTR/4A, MSLB/A, MSLB/2A) as output. Here, 'A' implies the relative break size. Data was acquired through MARS(multi-dimensional analysis of reactor safety) code. During the simulation, operator actions were considered to ensure data diversity.

The model is constructed with fully-connected layer only, to simplify the processes for XAI method application. Developed model consists of five fully-connected layers, and has shown about 93.1% accuracy for 84,625 training data and 91.6% accuracy for 20,000 testing data[7]. Fig. 1 is a schematic of the developed accident scenario classification model.

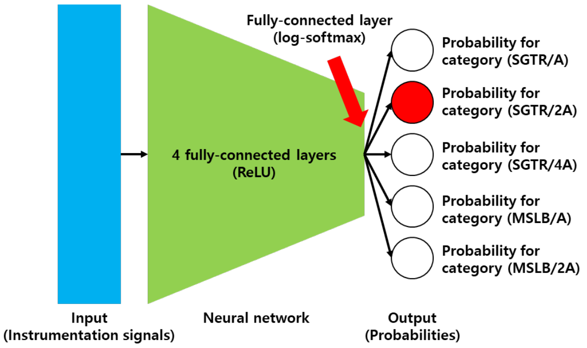
Fully-connected layer
(log-softmax)
4 fully-connected layers
(ReLU)
Probability for category (SGTR/A)
Probability for category (SGTR/2A)
Probability for category (SGTR/4A)
Probability for category (MSLB/A)
Probability for category (MSLB/2A)
Input
(Instrumentation signals)
Neural network
Output
(Probabilities)

Fig. 1. Schematic of the developed accident scenario classification model

2.2 Application of XAI Method

After the development of the model, four kinds of representative XAI methods were applied[7]. These methods are; integrated gradients(IG), gradient SHAP, DeepLift, and layer-wise relevance propagation(LRP). These methods deduce attribution scores for every input elements, which imply the relative importance of input element for deducing specific output. However, although the input is same, the order of deduced attribution scores can be different since their underlying algorithms are not same.

Fig. 2 shows the example of XAI method application results for 10,000-th data and 17-th variable(RCS subcooling margin). It can be easily found that the trends of attribution scores are highly distinct to each other, although these results are deduced for same data and variable.

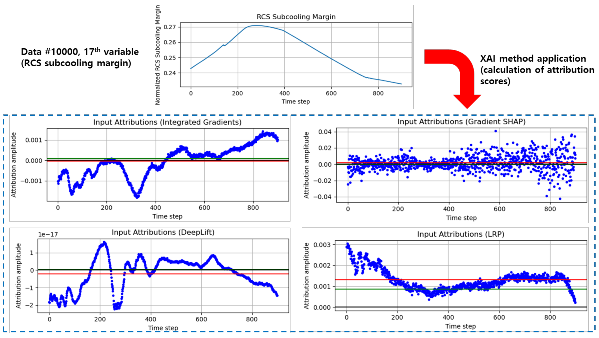

Fig. 2. Example of XAI method application results

2.3 Application of Perturbation Analysis

After the application of XAI methods, perturbation analysis is conducted. Perturbation analysis is a method that observes the changes of output, while iteratively changing the input elements to the random or pre-defined value, starting from the higher attribution score. Perturbation analysis is based on the idea that if the XAI method has found important input elements better, the probability value for correct label would be rapidly decreased, since important input element is changed into meaningless or neutral value. Processes of perturbation analysis are as follows.

1) Data selection: randomly selects data that model correctly classifies, from both training and testing datasets.
2) Calculating attribution score: calculates attribution scores for selected data for each XAI method.
3) Sorting: arrange the attribution scores derived from each XAI method in descending order.
4) Perturbation infusion: replace input element as random value, starting from the element with higher attribution score.
5) Probability calculation: apply replaced input to the sample model and calculate the probability for true label.

Procedures 4) and 5) are repeated to infuse perturbation for multiple input elements, and entire procedures are repeated to conduct perturbation analyses for multiple data.

It is extremely important to define meaningless or neutral value to appropriately conduct perturbation analysis. Generally, for the image data, this value is set to ‘0’ for greyscale color encoding or (0, 0, 0) for RGB color encoding. For the time-series data however, defining meaningless or neutral value is much harder as it highly depends on the characteristics of data. Therefore, further experiments were conducted for the various perturbing values.

In this study, 500 data that correctly classified by the model were selected randomly, and perturbations were infused up to the 1,710-th input elements(10% of total input elements) during the perturbation analysis. For the perturbing value candidates including random(sampled from uniform distribution), zero, one, and inverse(1-x), perturbation analyses were repeatedly conducted. For the perturbation infusing orders, additional to the orders deduced from the XAI methods, random perturbation order is also considered for comparison. Following figures(Fig. 3, Fig. 4, Fig. 5, and Fig. 6) represent the results of perturbation analyses, corresponding to the each perturbing value. It is easily found that the trends of output probability changes are highly different, according to the applied XAI method and perturbing value.

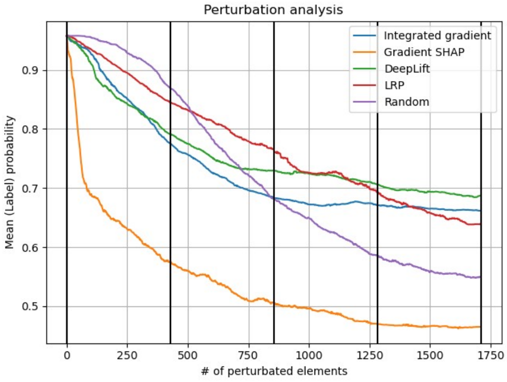

Fig. 3. Result of perturbation analysis for ‘random’ perturbing value

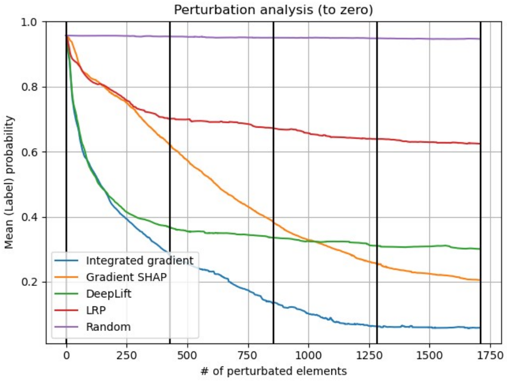

Fig. 4. Result of perturbation analysis for 'zero' perturbing value

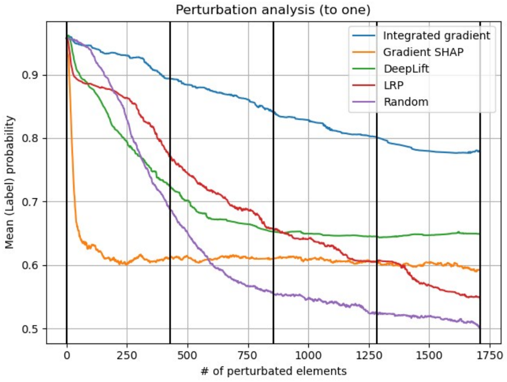

Fig. 5. Result of perturbation analysis for 'one' perturbing value

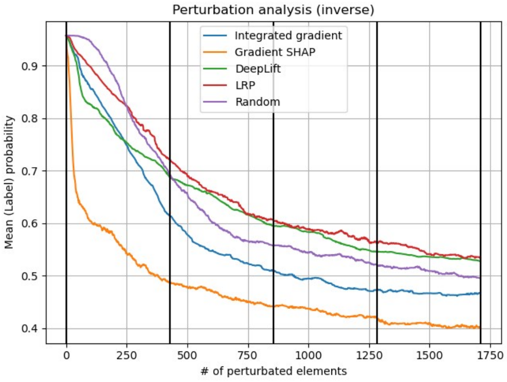

Fig. 6. Result of perturbation analysis for 'inverse' perturbing value

2.4 Method for Selecting Perturbing Value

The results from perturbation analysis are highly affected by the perturbing value settings. However, it is difficult to find meaningless or neutral value, especially for the time-series data.

If the perturbing value is optimal(i.e. perfectly value-neutral), the output for N-class classifier will be a vector that every elements have same value of 1/N(i.e. perfectly uncertain), when all input elements are changed into that perturbing value. If the perturbing value is sub-optimal but value-neutral enough to conduct perturbation analysis, the probability values for every classes will become more similar and the uncertainty of the output will become more higher while more and more input elements are changed. Based on this idea, we propose to utilize information entropy as a scale for selecting perturbing value. If the specific perturbing value is more value-neutral than the others, information entropy of the output will more rapidly increased along with the progression of perturbation analysis.

Consequently, it is able to quantitatively compare and select best XAI method through, 1) select the perturbing value that mostly increases information entropy, 2) conduct the perturbation analysis based on the selected perturbing value, 3) find the XAI method that mostly decreases output probability for correct label.

In this study, additional experiments were conducted to confirm which perturbing value is more appropriate for conducting perturbation analysis. Same to the previous sub-section, 500 data that correctly classified by the model were selected randomly, and perturbations were infused up to the entire 17,100 input elements(100% of total input elements) during the perturbation analysis. For the perturbing values, random, zero, one, and inverse were considered. For the perturbation infusing orders instead of the order deduced from the XAI method application, three kinds of orders were tested including: randomly select perturbation points(among 17,100 input elements), variable-wise perturbation(among 19 kinds of variables), and timestep-wise perturbation(45 seconds as unit timestep, among 900 seconds of time length).

Following figures(Fig. 7, Fig. 8, and Fig. 9) represent the mean entropy profiles for every perturbation infusion orders. It is easily found that the trends of mean information entropy are increased only when the perturbing value is zero for every perturbation orders. It implies that for the given model and data, perturbing value '0' is most appropriate among four kinds of perturbing value candidates, for conducting perturbation analysis. Moreover, it can be concluded that the 'integrated gradients' is deducing best explanations among four kinds of applied XAI methods, as it was shown most rapid probability decrement during the perturbation analysis with perturbing value is set to '0'(corresponds to Fig. 4). This result is also accorded with the result that the random perturbation order has shown almost no decrement during the perturbation analysis with same perturbing value settings.

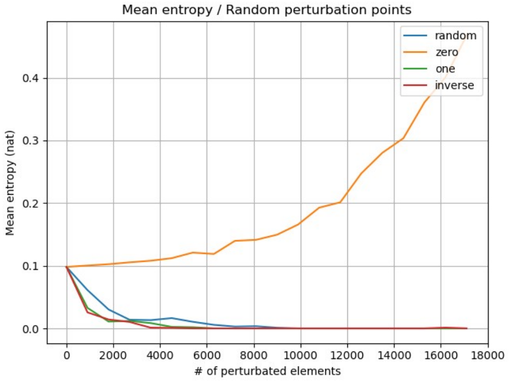

Fig. 7. Mean entropy profiles for random perturbation points

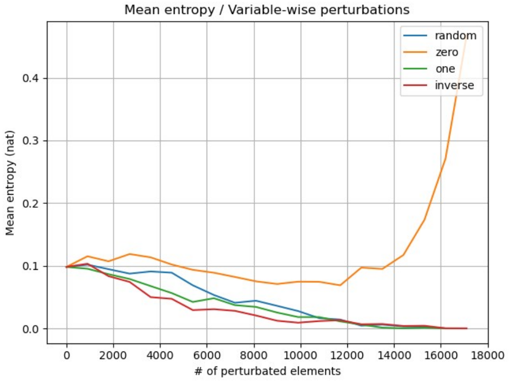

Fig. 8. Mean entropy profiles for variable-wise perturbations

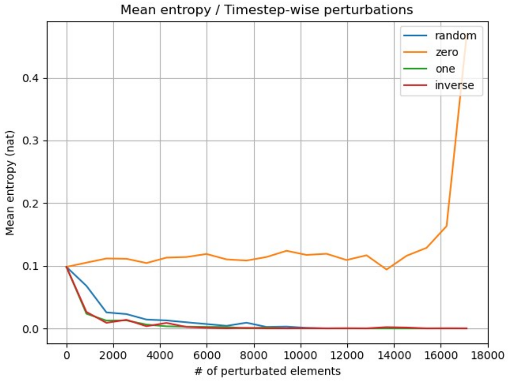

Fig. 9. Mean entropy profiles for timestep-wise perturbations

3. Conclusions

In this study, a simple accident scenario classification model based on DNN is developed, and various XAI methods are applied to deduce attribution scores. Then, perturbation analysis is conducted to quantitatively compare XAI methods on the model that uses time-series data. During the experiment, to deal with the problem that the results of perturbation analysis are heavily affected by the perturbing value, information entropy-based perturbing value selection method is proposed. Through the experiment, it is concluded that among the applied four kinds of XAI methods, integrated gradients deduced best explanations for given model and data. This experiment can be conducted similarly for the other models and XAI methods that deduce attribution score.

However, the best XAI method may changes according to the characteristics of model and data. In this point of view, the implication of this study is that the proper XAI method should be selected based on the perturbation analysis and proposed perturbing value selection method, not integrated gradients is the best XAI method for time-series data-based models applied in nuclear field.

There is a limitation that based on the perturbation analysis and the proposed perturbing value selection method, only quantitative yet relative comparison is possible. Experiment conducted in this study only reveals which XAI method is the best among the applied candidates, while it does not reveal whether the deduced explanation is valid or meaningful. Therefore, as a further study, we are planning to adopt and enhance sanity check[8] method for time-series data-based models.

ACKNOWLEDGMENT

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT: Ministry of Science, ICT) (No. RS-2022-00144405).

REFERENCES

[1] Y. H. Choi, G. M. Yoon, and J. H. Kim, Unsupervised Learning Algorithm for Signal Validation in Emergency Situations at Nuclear Power Plants, Nuclear Engineering and Technology, Vol. 54, No. 4, pp. 1230-1244, 2022.

[2] S. G. Kim, Y. H. Chae, and P. H. Seong, Development of a Generative-adversarial-network-based Signal Reconstruction Method for Nuclear Power Plants, Annals of Nuclear Energy, Vol. 142, 2020.

[3] Y. H. Chae, C. Y. Lee, S. M. Han, and P. H. Seong, Graph Neural Network based Multiple Accident Diagnosis in Nuclear Power Plants: Data Optimization to Represent the System Configuration, Nuclear Engineering and Technology, Vol 54, No. 8, pp. 2859-2870, 2022.

[4] S. H. Ryu, H. M. Kim, S. G. Kim, J. H. Jin, J. H. Cho, and J. K. Park, Probabilistic Deep Learning Model as a Tool for Supporting the Fast Simulation of a Thermal-hydraulic Code, Expert Systems with Applications, Vol. 200, 2022.

[5] D. I. Lee, M. Arigi, and J. H. Kim, Algorithm for Autonomous Power-increase Operation using Deep Reinforcement Learning and a Rule-based System, IEEE Access, 2020.

[6] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, A Survey on Neural Network Interpretability, IEEE Transactions on Emerging Topics in Computational Intelligence, Vol. 5, No. 5, 2021.

[7] S. G. Kim, and J. H. Cho, Comparative Study of Explainable Artificial Intelligence Methods in Nuclear Field, 32nd European Safety and Reliability Conference(ESREL 2022), Aug.28-Sep.1, 2022, Dublin, Ireland.

[8] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, Sanity Checks for Saliency Maps,

Proceedings of the 32nd International Conference on Neural Information Processing Systems(NIPS'18), Dec.3-8, Montreal, Canada.