Ensuring Trustable Results through Misdiagnosis Detection with Explainable AI

Ji Hyeon Shin and Seung Jun Lee*
Department of Nuclear Engineering, Ulsan National Institute of Science and Technology, 50, UNIST-gil, Ulsan 44919, Republic of Korea
**Corresponding author: sjlee420@unist.ac.kr*

***Keywords :** abnormal state diagnosis, explainable AI, relevance score

1. Introduction

When an abnormal problem occurs within a system or component in a nuclear power plant (NPP), operators should take appropriate actions to alleviate the plant state to a normal. These are performed by following tasks outlined in the operating procedures corresponding to the specific abnormality. For these, operators should first perform diagnostic tasks to identify the specific problem.

Recently, several classification models have been studied to support operators in diagnosing abnormal events in NPPs. Most of these classification models are based on artificial neural networks. However, due to the black-box nature of artificial neural networks, operators may require additional information to trust the provided diagnosis. Therefore, additional techniques need to be introduced to enhance operators' trust and applicability of artificial neural networks.

In this study, we purpose to detect misdiagnosis by classifying whether the causes for diagnosis of the neural network are appropriate, thereby ensuring trustable diagnosis. The relevance of all input features to the model's diagnosis is calculated using explanation techniques. A classifier trained on these feature relevances, as a new input, determines whether the model has performed the diagnosis based on appropriate causes.

2. Background

Artificial neural network models are called "black-box," making it challenging to ensure transparency in how these models function. To address this issue, prior research has been focused on improving model interpretability, leading to the development of various techniques such as Deep Learning Important FeaTures [1], Local Interpretable Model-agnostic Explanation [2], and SHapley Additive exPlanations [3]. In line with these advancements, studies also have been conducted to interpret models used for diagnosing the NPP states. These interpretations can reveal which input features are relevant to the model diagnosis. However, despite these, the reasons for why specific inputs lead to particular outputs often remain unexplained to operators. This is because the model learns patterns from the data without explicitly revealing what those patterns are. Therefore, directly providing model interpretation might confuse operators, as it could differ from their understanding. Consequently, it is needed to process the model interpretation into clear information that can be effectively understand to operators.

3. Methods

In this study, the interpretation of model diagnosis is processed using the following methods. Through this approach, we propose to detect instances of misdiagnosis in model diagnosis.

3.1 Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) is a technique used in neural networks to explain predictions by redistributing the output back through the layers to the input features [4]. It works by decomposing the prediction score layer by layer, attributing relevance to each neuron, and ultimately assigning relevance scores to the input features, which highlight the relevance of each feature to the final decision.

3.2 Consistency about Relevance of Model Diagnosis

An explanation technique provide clear insights into how each input feature influences the model's predictions. In this point, this study assumes the following:

(1) Diagnosis relevance scores for each abnormal event are consistent.
(2) Diagnosis relevance scores for misdiagnosis results are inconsistent.

Therefore, we can detect mis-diagnostic cases with low consistency in feature relevance scores for each abnormal event, as shown in the figure below

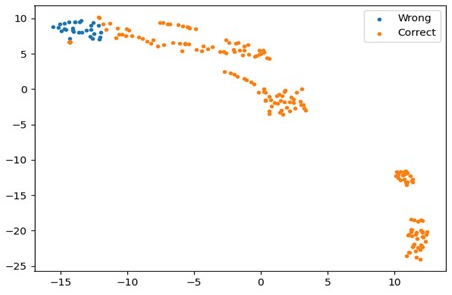

Fig. 1. Example of relevance score visualization.

3.3 Proposed Approach

This technique used the classifier for training feature relevance calculated from explanation techniques. The calculated scores of each feature can represent the model diagnostic relevance. To make a model that detects inconsistent relevance scores, new class (mis-diagnostic class) is added and trained with inconsistent scores different from the original scores. For this, the classifier are trained feature relevance scores for the second-best diagnosis as a new class with wrong relevance scores.

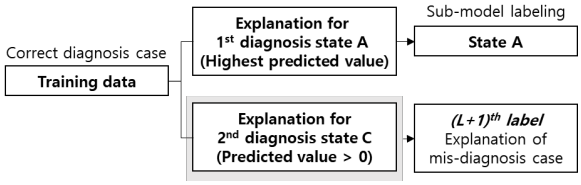

Fig. 2. 1st diagnosis state : 2nd diagnosis state = 1 : 1/2.

The training and evaluation process of the classifier, which uses the labeled relevance scores as training data, is introduced in the figure below.

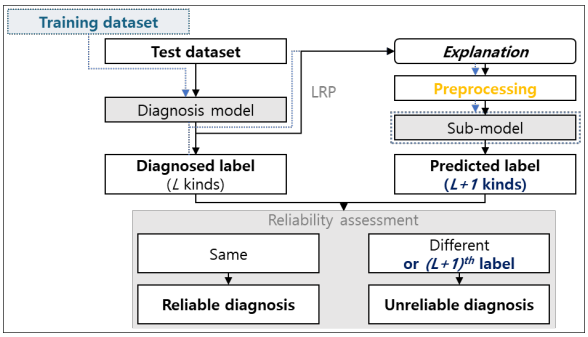

Fig. 3. The proposed approach.

4. Case Study

4.1 Abnormal State Datasets

Datasets were generated from Generic 2-loop pressurized water reactor simulator. It was sampled every second during 60 time-steps for 391 monitoring parameters. 49 datasets which has different malfunction fraction each other were generated for each of the 15 abnormal states in table I. There are a total 40,425 data by 735 datasets for model training. 1% Gaussian noise was added to half of the total datasets. The test data was generated by simulating 735 datasets in the same process and then adding 5% Gaussian noise.

Table I: Kinds of Abnormal States

Num.	Abnormal state	Min.	Max.
1	Steam generator tube leakage	4	10
2	Charging line break	10	100
3	Letdown line leakage	100	1000
4	Loss of condenser vacuum	45	50
5	Pilot-operated safety relief valve leakage	0.2	1
6	Circulating water tube leakage	65	100
7	Main steam isolation valve positioner failure	0	0.3
8	Loss of reactor coolant pump seal injection water	0	0.03
9	Main steam header steam leakage	2	3
10	Pressurizer spray valve positioner failure	70	100
11	Component cooling water service loop header leakage	10	100
12	Low-pressure feedwater heater tube break	10	100
13	High-pressure feedwater heater tube break	55	90
14	Main feedwater pump recirculation valve positioner failure	0.45	0.7
15	Turbine control valve positioner failure	0	0.25

4.2 Abnormal State Diagnosis Model

We used two-channel convolutional neural network for abnormal state diagnosis model [5]. By monitoring both current values and their changes as shown in a below figure, this model can better detect abnormalities and sudden shifts in the data. These input features were arranged considering the system locations for image data shape and used as training data. The model hyperparameter was shown in a below table.

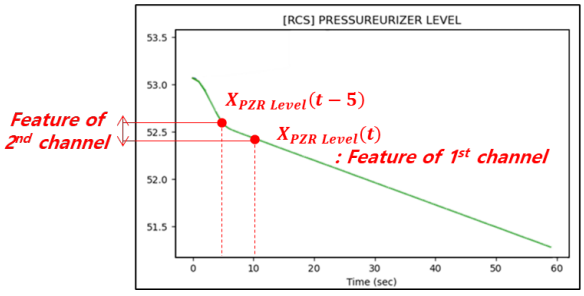

Fig. 4. Description of features in two-channel convolutional neural networks

Table II: Model Hyperparameter

Number of convolution layers	3
Number of filters in convolution layers	10
Kernel size of filters in convolution layers	(2, 2)
Activation function of convolution layers	ReLU
Activation function of dense layers	softmax
Loss function	categorical crossentropy
Optimizer	Adam

Model training was stopped earlier by monitoring validation loss with patience of 20 epochs. This model achieved an accuracy of 85.76% at test dataset.

4.3 Relevance Appropriateness Classifier

In this case study, Light Gradient-Boosting Machine [6] was used as the classifier to classify feature relevance scores for each abnormal state. The outputs of this classifier on the test datasets can be compared with the outputs of the abnormal state diagnosis model to detect misdiagnosis cases. In other words, this determination whether the diagnosis model performed its classification based on consistently relevant features. The results are shown in the table below.

Table III: Results of Case Study

Test results of diagnosis model			
	Correct diagnosis	Incorrect diagnosis	Number of incorrect case
	85.76 %	14.24 %	5,755
Relevance appropriateness results of classifier			
	Correct diagnosis	Incorrect diagnosis	
Trustworthy diagnosis	96.85 %	11.94 %	687 / 5,755
Untrustworthy diagnosis	3.15 %	88.06 %	5,068 / 5,755

This classifier detected that more than 88% of the cases where the abnormal state diagnosis model made a misdiagnosis were based on inconsistent feature relevance.

5. Conclusions

This study addresses the issue of operators' lack of trust in the results provided by the neural network due to its black-box nature. To resolve this, we introduced a classifier that determines whether the model performs its diagnosis based on appropriate causes. This classifier learns the relevance of input features calculated through explanation techniques, enabling it to detect instances of misdiagnosis by the model. Consequently, the proposed approach can prevent the abnormal event diagnosis model from providing incorrect information to the operator with high probability. Therefore, it provides a foundation for operators to trust on neural network-based diagnostic models more effectively. This study is expected to further expand the applicability of neural network technology in the operation and management of NPPs in the future.

REFERENCES

[1] Shrikumar, A., Greenside, P. and Kundaje, A., 2017, July. Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145-3153). PMlR.
[2] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
[3] Lundberg, S., 2017. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.
[4] Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning. 2019:193-209.
[5] Lee, G., Lee, S.J. and Lee, C., 2021. A convolutional neural network model for abnormality diagnosis in a nuclear power plant. Applied Soft Computing, 99, p.106874.
[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

ACKNOWLEDGEMENTS

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2022-00144042).