

Aligning Deep Diagnostic Explanations with Prior Domain Knowledge in Nuclear Power Plants

Ji Hyeon Shin *, Young Ho Chae, Seo Ryong Koo

*Advanced Instrumentation & Control Research Department, Korea Atomic Energy Research Institute, 111,
Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon, 34057*

**Corresponding author: jhshin0127@kaeri.re.kr*

***Keywords :** domain-knowledge, abnormal event diagnosis, explanation alignment, operator support system

1. Introduction

Research on applying artificial intelligence (AI) to the operator support of nuclear power plants (NPPs) has been carried out actively in recent years. The reported applications cover early diagnosis, trend prediction, response support, autonomous operation, and signal validation. In Korea, this line of work has been continued through a series of national projects, including the development of AI based startup and shutdown operation support, the development of AI based decision support for the prevention of human error, and the development of core technologies for high level autonomous operation. Among them, diagnosis models trained with simulator data have reported high classification accuracy for abnormal states [1].

Despite this volume of research, few of the resulting models have been applied to an operating plant. One reason is the difference in evaluation criteria. Research on diagnosis models is evaluated mainly by classification performance, whereas application to a plant is evaluated by design and verification criteria. Current standards assume a deterministic software V-model, in which the requirements are explicitly specified and the code can be traced back to them. These assumptions do not hold for an AI component. Park et al. [4] proposed a verification and validation strategy for AI components in NPP instrumentation and control systems, and showed that both process and functional completeness must be secured over the software development life cycle. Another reason is that the requirements demanded of an AI component are not satisfied by current practice. The International Atomic Energy Agency [5] indicates that the objective used to train a model has to be consistent with the operating objectives of the plant, that expert knowledge should be built into the algorithm rather than left outside it, and that for safety critical applications the ability to account for a decision outweighs predictive accuracy. In practice, however, model development has concentrated on improving accuracy while the black box nature of the model, which is the important cause of above reasons, is left unchanged.

The conventional response to the black box problem has been to attach a post-hoc explanation to the output of a trained model. In our previous work [1], three

explanation techniques were applied to a convolutional neural network trained for NPP abnormal event diagnosis, and the parameters they highlighted were shown to correspond to the physical symptoms of each event. A post-hoc explanation, however, only observes the rationale that the model has already learned. It cannot correct that rationale, and it does not by itself show that the rationale is valid. A model can therefore produce a correct diagnosis while relying on parameters that are unrelated to the physical cause of the event.

In this study, we propose a diagnosis model whose learning direction is specified in advance and which predicts according to that design. A prior database of alarms and symptoms for each abnormal event is converted into a variable mask and is injected as a constraint during training, so that the rationale of the model follows the design. The model structure is not decomposed and is kept end-to-end; only the path of the rationale is constrained.

2. Methodology

The proposed method is applied to an abnormal state diagnosis model and consists of two stages. In Stage A, the natural language description of the alarms and symptoms of each event is converted into a variable mask, and in Stage B the mask is used as an attribution prior for training the diagnosis model. Explanatory interactive learning [3] assumes repeated feedback from a human annotator on the explanation of the model. In the proposed method this feedback is replaced by a verified prior database, so that the supervision is reproducible and does not require operator participation.

2.1 Stage A: extraction of the variable mask

The set of variables related to event c , denoted R_c , is generated automatically from the prior database and from the description of the 2,564 plant variables, which includes the variable name, abbreviation, unit, and physical meaning. Domain abbreviations are first expanded and then matched by sentence embedding similarity, instrument tags are matched directly by code,

and variables that are physically connected to a symptom are included together, so that pressurizer level is included when a symptom is described in terms of pressurizer pressure. The correspondence between an event and its variables is not provided by a human.

2.2 Stage B: aligned training

The right for the right reasons loss proposed by Ross et al. [2] is applied to the multivariate diagnosis problem. The loss function is given in Eq. (1).

$$L = CE(\hat{y}, y) + \lambda \sum_{i \in R_y} \left(\frac{\partial \log p_y}{\partial x_i} \right)^2 \quad (1)$$

Here p_y is the predicted probability of the correct class y , R_y is the set of variables extracted in Stage A for class y , and the partial derivative of $\log p_y$ with respect to x_i is the sensitivity of the log probability to the input variable x_i . The second term suppresses the influence that variables outside R_y have on the predicted class probability. The normal state is excluded from the penalty because R_c is not defined for it, and λ is increased linearly during the early epochs so that the penalty does not interfere with the classification learning.

The alignment of the explanation is measured by the in-mask ratio given in Eq. (2). A larger value means that the rationale of the model lies on the variables of the prior database.

$$in - mask = \frac{\sum_{i \in R_c} \left| \frac{\partial \log p_y}{\partial x_i} \right|}{\sum_i \left| \frac{\partial \log p_y}{\partial x_i} \right|} \quad (2)$$

3. Experimental Setup

Abnormal operating scenarios generated by a simulator of the APR-1400 were used. One normal state and ten abnormal events were selected, and 2,564 variables were sampled. The training set consisted of 258 files and the test set of 66 files, each covering the 120 s interval after the onset of the event. The target events and the size of the extracted mask are listed in Table I. The size of R_c ranges from 10 to 93 variables, and the union of all masks contains 391 of the 2,564 variables.

A simple fully connected network of 2,564-256-64-11 was used so that the effect of the proposed training could be separated from the effect of the model structure. The model was trained for 50 epochs with the Adam optimizer at a learning rate of 1e-3 and a weight decay of 1e-4, and λ was increased linearly between the 15th and the 30th epoch. All comparisons were made with the same data split and the same hyperparameters, changing

only λ .

Table I: Target abnormal events and the size of the extracted variable mask

Event num.	Event	$ R_c $
1	SG tube leak	69
2	RCP seal damage	33
3	Letdown line leak	78
4	Charging line valve fault	64
5	ESW pipe leak	10
6	MSIV closure	56
7	SBCS valve fails open	66
8	LP heater tube rupture	93
9	Circulating water pump trip	62
10	T_{avg} input signal failure	31

4. Results

4.1 Diagnostic accuracy and explanation alignment

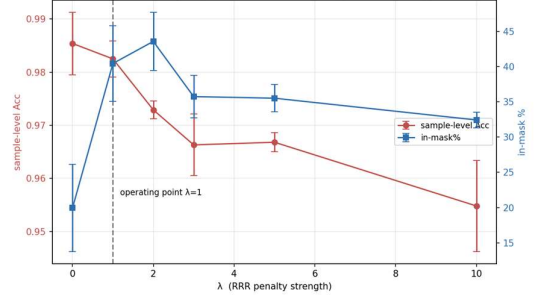


Fig. 1. Diagnostic accuracy and explanation alignment for varying λ .

Fig. 1 shows the file level accuracy and the in-mask ratio for each λ . The in-mask ratio does not increase in proportion to λ . It reaches a maximum of 43.56% at $\lambda=2$ and decreases again at larger values. The accuracy is not degraded at $\lambda=1$ or 2 with respect to the vanilla model, but it decreases to 0.9803 ± 0.0133 at $\lambda=10$. Since no loss of accuracy was detected at $\lambda=1$, this value was taken as the representative operating point. At this point the in-mask ratio increased from $19.95 \pm 3.84\%$ to $40.43 \pm 3.32\%$.

4.2 Alignment for each event

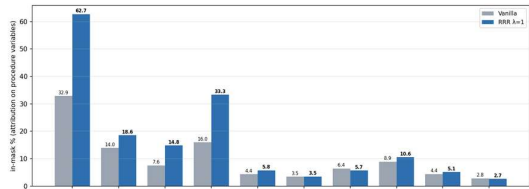


Fig. 2. Explanation alignment for each abnormal event.

Fig. 2 compares the in-mask ratio of each event for the

vanilla model and for the proposed model at $\lambda=1$. The alignment improved clearly for five events, the largest changes being from 32.92% to 62.70% for SG tube leak event, from 15.96% to 33.32% for charging line valve fault event, and from 7.59% to 14.85% for letdown line leak event. However, the ratio slightly decreased for SBCS valve fails open event and T_{avg} input signal failure event.

Table II lists, for each event, the proportion of test files in which the variable with the largest attribution belonged to the prior database. The attribution maps were produced for the vanilla model and for the model trained at $\lambda=10$. This value is used as the strong-penalty limit, because the top-1 variable is a discrete statistic measured over only 4 to 13 test files per event and is not displaced reliably at $\lambda=1$; the operating point is characterized instead by the continuous in-mask ratio of Fig. 1 and Fig. 2. The proportion increased from 8% to 100% for letdown line leak event and from 0% to 100% for LP heater tube rupture event, and it remained 0% for both models in four events. It should be noted that the proportion decreased for RCP seal damage event and for ESW pipe leak event, which shows that a strong penalty does not always move the highest ranked rationale into the prior database.

Table II: Proportion of test files whose top-1 attribution falls inside the prior database

Event	Vanilla	Proposed ($\lambda = 10$)
SG tube leak	100 %	100 %
RCP seal damage	80 %	20 %
Letdown line leak	8 %	100 %
Charging line valve fault	75 %	100 %
ESW pipe leak	33 %	17 %
MSIV closure	0 %	0 %
SBCS valve fails open	0 %	0 %
LP heater tube rupture	0 %	100 %
Circulating water pump trip	0 %	0 %
T_{avg} input signal failure	0 %	0 %

4.3 Reallocation of the rationale

Fig. 3 shows the change of the attribution share of the input variables for SG tube leak event, the event with the highest alignment. The share of the R_c set increased from 34.4% to 64.5%. The four main steam line radiation monitors, PRRI217N to PRRI220N, increased from a range of 4.04 to 4.41% to a range of 6.64 to 6.89%, CMTI032A increased from 1.50% to 4.30%, and PRRI231A from 0.50% to 2.59%, whereas most of the variables outside R_c decreased, as in the case of ESLI141 from 1.59% to 0.50%. The penalty derived from the prior database suppresses the variables outside the mask, and the suppressed share is transferred to the variables of the prior database.

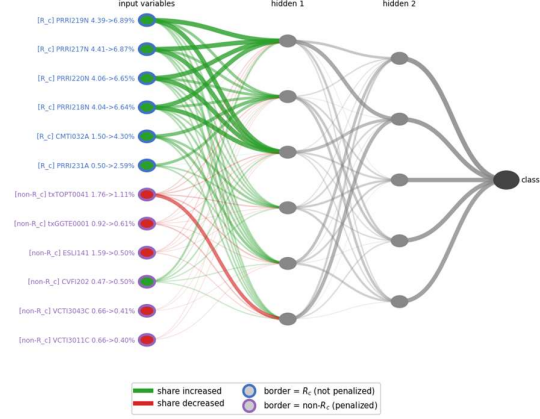


Fig. 3. Change of the attribution share for SG tube leak.

5. Conclusion

In this study, a pipeline that automatically converts the natural language description of a prior database of abnormal events into an attribution mask was proposed, together with a diagnosis model that uses the mask as a constraint during training. The explanation of the model could be moved toward the guided variables without a loss of diagnostic accuracy. At $\lambda=1$ the file accuracy was 0.9970 ± 0.0038 and the in-mask ratio increased from 19.95% to 40.43%. From the viewpoint of verification, the alignment of the explanation can therefore be treated as a design specification that is checked in advance rather than as a property that is observed afterwards.

At the same time, the results separate the events for which the design condition of the prior database works from those for which it does not. Even when the variables of the prior database carry enough information to discriminate an event, the model bypasses them if other variables carry the same information, and the same penalty does not produce alignment. Therefore, further work is required to analyze the cause for the events that do not follow the guidance, to extend the constraint to the time axis, and to confirm the approach with other model structures required by an operator support system.

ACKNOWLEDGEMENT

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry, & Energy (MOTIE) of the Republic of Korea (No. 20224B10100130).

REFERENCES

- [1] J. H. Shin, J. Bae, J. M. Kim, and S. J. Lee, An interpretable convolutional neural network for nuclear power plant abnormal events, *Applied Soft Computing*, Vol. 132, 109792, 2023.
- [2] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, Right for the right reasons: training differentiable models by constraining their explanations, *Proceedings of IJCAI*, pp.

2662-2670, 2017.

[3] S. Teso and K. Kersting, Explanatory interactive machine learning, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 239-245, 2019.

[4] J. Park, T. Kim, and S. Koo, Verification strategy for artificial intelligence components in nuclear plant instrumentation and control systems, Progress in Nuclear Energy, Vol. 164, 104842, 2023.

[5] IAEA, Considerations for deploying artificial intelligence applications in the nuclear power industry, IAEA Nuclear Energy Series No. NR-T-1.26, Vienna, 2025.