

Open-Set Recognition for Root Cause System Identification of Untrained Abnormal Condition in a Small Modular Reactor

Dong Geon Jang, Eon Sang Jeon, Young Hun Kim, Man Gyun Na*

Department of Nuclear Engr., Chosun Univ., 10 Chosundae 1-gil, Dong-gu, Jeonnam-Gwangju Special Self-Governing City, Republic of Korea

*Corresponding author: magyna@chosun.ac.kr

***Keywords :** Open set recognition, Root Cause System, System-wise autoencoder, Small modular reactor, Untrained fault detection

1. Introduction

The Small Modular Reactor (SMR) is a next generation reactor type characterized by miniaturization and modularization, and an operating concept in which a small crew monitors several modules together from a single main control room is under review. When an abnormal event occurs in this environment, the volume of alarms and measurement signals that a single operator must interpret becomes larger than in a conventional large scale nuclear power plant, and researchers have actively conducted deep learning based diagnosis studies to reduce this burden. However, a supervised classifier remains valid only within the set of scenarios it has learned. Such a classifier assigns an abnormal event that is absent from the training set to the most similar trained scenario and outputs a high confidence value together with that assignment, and this behavior can mislead the operator. A diagnosis system must therefore assume that untrained scenarios occur, and it must distinguish such inputs from known scenarios and report them to the operator. This requirement corresponds to the open-set recognition problem, which demands both the classification of known classes and the identification of untrained inputs.

The diagnosis system must also provide information to the operator when an input is judged as untrained. In a situation where the system cannot specify a scenario name, what the operator first checks is which plant system the abnormal event occurred in. For untrained abnormal conditions, this study proposes an open-set abnormal event diagnosis method that identifies the root cause system, using normal and abnormal operating data collected from an Integral Pressurized Water Reactor (iPWR) simulator. The learning layer follows an open-set recognition structure. It first judges whether an input is untrained by using the reconstruction error of one dimensional convolutional autoencoders (CNN-AE) that are trained independently for each plant system, and an Inception classifier then classifies an input judged as trained into one of the trained scenarios. The learning layer does not pass an input judged as untrained to the classifier, and it identifies the root cause system by using the reconstruction error of each plant system.

The proposed method provides the classification of trained scenarios, the identification of untrained abnormal events and the blocking of their misclassification, and the identification of the root cause system within a single framework. By delivering information about the plant system requiring action to the operator even in abnormal situations that the training data does not cover, the proposed method is expected to contribute to improving plant safety in an SMR operating environment where a small crew monitors multiple modules.

2. Methods

2.1 CNN-AE

An autoencoder is an unsupervised neural network trained solely to minimize the reconstruction error, compressing the input into a low-dimensional latent representation with an encoder and restoring it with a decoder. The CNN-AE is a variant in which the encoder and the decoder are composed of one-dimensional convolutional layers. Owing to local receptive fields and weight sharing, each filter detects waveform patterns over short time intervals regardless of their position[1].

A multistage convolutional encoder compresses the input window into a bottleneck latent feature map, and a decoder symmetric to the encoder restores the input from this feature map. The autoencoder reconstructs an input within the training distribution with a small error, whereas an untrained pattern leaves a large error. The learning layer identifies as an untrained pattern any input whose reconstruction error exceeds a threshold determined from the error statistics of the training data. The structure of the CNN-AE is illustrated in Fig. 1.

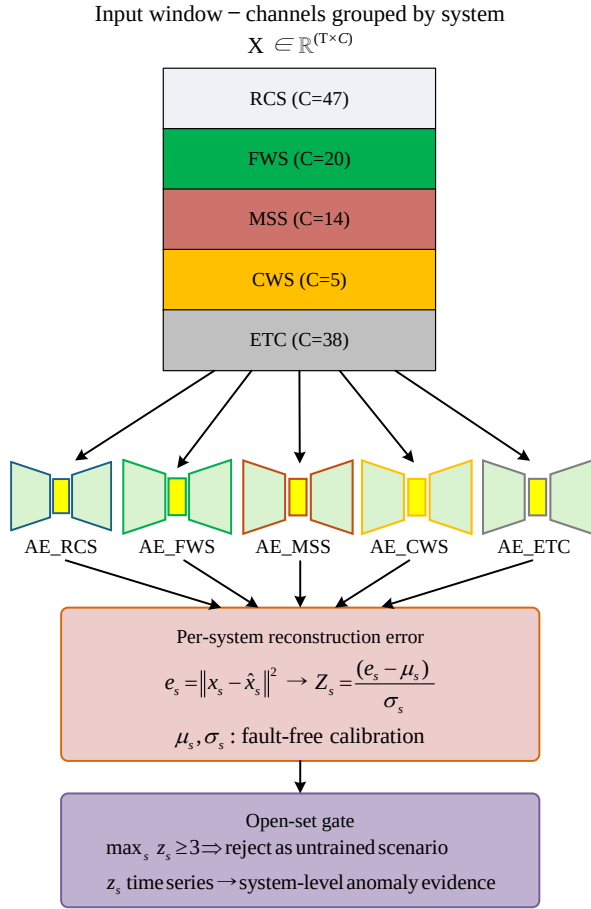


Fig. 1. System-wise CNN-AE structure

2.2 Inception network

The Inception network is a deep convolutional neural network for time series classification that uses the Inception module, which applies convolution kernels of different lengths in parallel, as its basic building block. It addresses the problem that a single kernel size cannot simultaneously capture the short abrupt changes and the slow trends that coexist in a time series.

In each Inception module, the bottleneck convolution reduces the number of channels of the input features so that the subsequent parallel convolutions are performed with fewer parameters. Convolution kernels of different lengths are then applied in parallel to the reduced features, and waveform patterns at multiple time scales are extracted simultaneously. A separate branch, which applies max pooling and a channel reduction convolution to the module input, adds features that are robust to small disturbances. The outputs of the branches are concatenated along the channel dimension and pass through batch normalization and an activation function to form the output of the module.

These modules are stacked in multiple layers, and the residual connection that adds the module input to its output keeps the training signal from vanishing as the depth increases. The output of the last module is summarized over the time axis by global average

pooling, and a fully connected layer with softmax produces the class probabilities. Global average pooling greatly reduces the number of parameters in the fully connected layer and thus suppresses overfitting under small sample conditions. The structure of the Inception network is illustrated in Fig. 2[2].

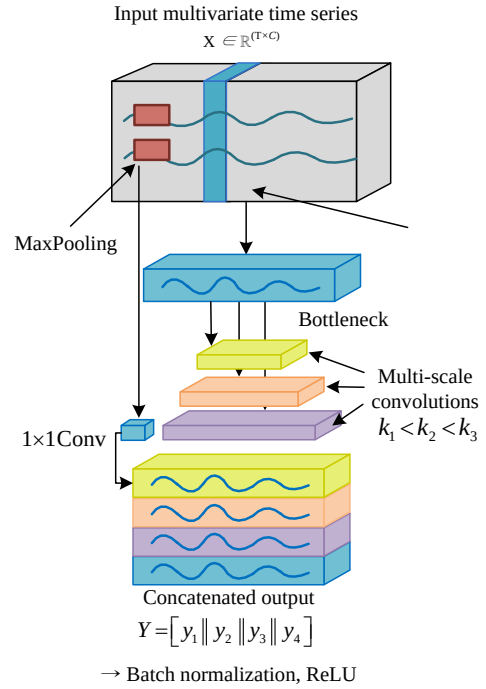


Fig. 2. Inception network structure

3. Data

3.1 Data collection and Preprocessing

The data in this study were collected using the iPWR simulator, an SMR simulator distributed by the International Atomic Energy Agency (IAEA). The simulator data were collected at 1 second intervals for 3 hours. The iPWR simulator is shown in Fig. 3.

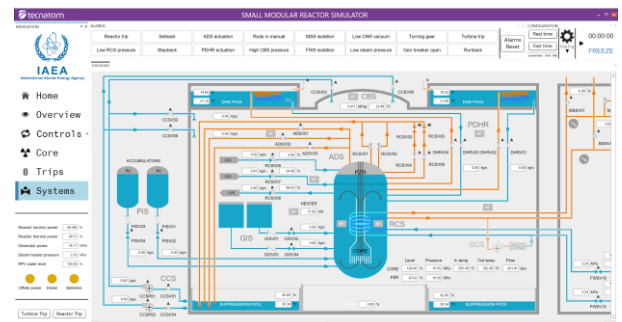


Fig. 3. iPWR simulator

The abnormal data consist of 13 scenarios across four systems. In every scenario, the abnormal event was injected 10 seconds after the start of the simulation, and runs were collected with various severity and ramp time

settings so that the data cover behaviors up to strong abnormal events leading to reactor trip. For scenarios in which severity and ramp time cannot be set, various mitigation actions were performed instead so that the data include behaviors with different response histories for the same abnormal event. For scenarios in which a reactor trip occurred, only the data before the trip were used for model training. The scenario composition and the number of runs are summarized in Table I.

Table I: Scenario composition

Systems	Scenario	Code	Runs
-	Normal	Normal	87
CWS	Condenser coolant pumps trip	CWSP01	105
FWS	FWS pump degradation	FWSP01	40
MSS	SGTR Loop1	MSS02	32
MSS	SGTR Loop2	MSS03	30
MSS	Main Steam Line Break	MSS04	163
MSS	Steam Line isolation	MSSV05	1
RCS	RCS valve leakage	RCSV04	67
RCS	Pressurizer level increase	RCSV07	50
RCS	Pressurizer spray valve failed open	RCSV09	50
RCS	Pressurizer pressure increase	RCS06	5
RCS	RCP trip	RCS	1
RCS	All RCP trip	RCSP	1

The five scenarios marked in bold were not used for training and were used only for testing. This study divided the data of the trained scenarios into training, validation, and test sets at a ratio of approximately 8:1:1 for each scenario. This study obtained the CWSP01, FWSP01, RCSV04, RCSV07, RCSV09, and RCS06 scenarios from UNIST.

This study selected 116 variables that belong to the list of variables the iPWR simulator can monitor as the input variables. This study normalized the input of the learning layer to the range between the minimum and the maximum of each variable and then composed the normalized input into sliding windows with a length of 5 s.

4. Result

This study analyzes the performance of the proposed diagnosis method in terms of the untrained judgement by the system-wise autoencoders, the classification of trained scenarios by the Inception classifier, and the identification of the root cause system for untrained scenarios. The evaluation metrics are the Accuracy on Unknown Samples (AUS) and the Accuracy on Known Samples (AKS)[3],[4]. AUS is the proportion of untrained scenario windows that are judged as untrained. AKS is the proportion of windows in the trained scenario test set that are not judged as untrained and are

also classified into the correct scenario. This study used the H-score, which is the harmonic mean of AKS and AUS, as the representative metric[5].

This study evaluated the identification performance for the root cause system with the accuracy the file level.

4.1 Detection of untrained scenarios

Among the 350,023 windows in the trained scenario test set, 340,590 windows were judged as trained and 9,433 windows were judged as untrained, which gives a false rejection rate of 2.70%. Among the 1,450,230 windows in the untrained scenario set, 1,364,815 windows were judged as untrained, which gives an AUS of 94.11%. The AUROC of the decision score is 95.44%. This study presents the results of the system-wise autoencoders in Fig. 4.

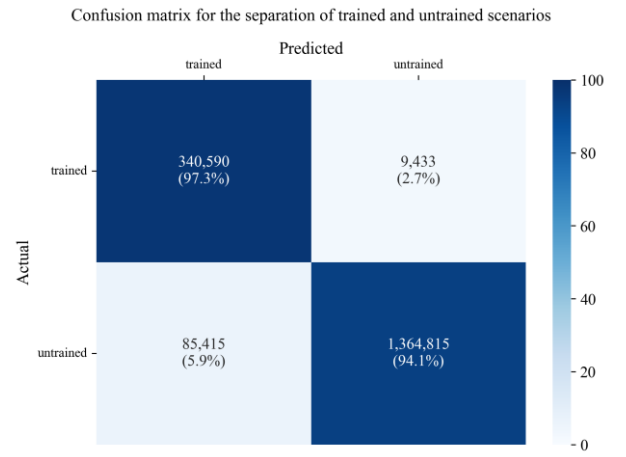


Fig. 4. Confusion matrix for detection of trained and untrained scenarios

4.2 Trained scenario classification

An Inception classifier classified the trained scenarios and achieved a balanced accuracy of 99.65% and a micro F1 of 99.75%. This study presents the confusion matrix of the Inception classifier in Fig. 5.

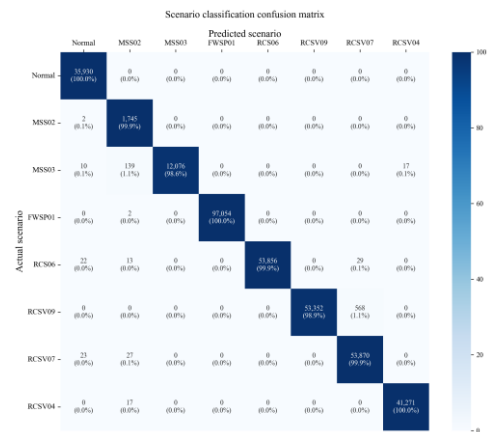


Fig. 5. Scenario classification confusion matrix.

The largest misclassification occurred for RCSV09, where 568 of the 53,920 windows, or 1.05%, were classified as RCSV07. Both scenarios are valve failures in the reactor coolant system, so their early behavior is similar. Every other misclassification is below 0.20%.

The proposed method blocked 94.11% of the untrained inputs and did not pass them to the classifier. This prevents an abnormal event that is absent from the training set from being wrongly assigned to one of the trained scenarios and reported to the operator. The remaining 5.89% were accepted as trained and passed to the classifier, and in this case the method outputs a wrong scenario name. The method judged 2.7% of the trained inputs as untrained. It provides no scenario name for these windows. Further work is in progress to reduce the untrained inputs that are wrongly accepted as trained. Open-set diagnostic performance evaluation results are presented in Table II.

Table II: Open-set diagnostic performance evaluation

Performance evaluation metrics	Performance
AKS	0.9706
AUS	0.9411
OSA	0.9468
H-score	0.9556
Balanced accuracy	0.9965
Micro F1-score	0.9975

4.3 Root cause system identification

For an input judged as untrained, this study normalized the reconstruction error of each plant system against normal operation, normalized it again by the threshold of that plant system, and selected the plant system whose score averaged over the early interval of the abnormal event is the largest as the root cause system.

A comparison of values without normalization structurally favors the plant system with the smaller standard deviation, so the comparison across plant systems does not hold. This study therefore normalized the score in two stages. It first standardizes the reconstruction error of each plant system by the mean and the standard deviation of that plant system obtained from the normal operation data. It then divides the result by the threshold of that plant system obtained from the intervals of the trained scenarios. The first stage removes the difference in signal scale across the plant systems and the second stage removes the difference in the magnitude of the fault response across the plant systems.

The accuracy is 99.63% and the mean accuracy over the plant systems is 99.80%. The accuracy for each true plant system is 1.000 for the circulating water system, 99.39% for the main steam system, and 1.000 for the reactor coolant system. Only one misattribution occurred, in which one main steam system run was

judged as the feedwater system.

The length of the early interval used for the root cause system judgement directly affects the performance. This study varied the judgement interval from 27s to 127s after injection and examined the change in performance. The results are presented in Table III.

Table III: Root cause system identification performance versus the length of the judgement interval

Judgement interval (s)	accuracy	CWS	MSS	RCS
27	0.996	1.000	0.994	1.000
37	0.996	1.000	0.994	1.000
67	0.738	1.000	0.567	1.000
127	0.461	1.000	0.116	0.500

The results are identical for the first 27s and the first 37s after injection. The performance is stable over a range of interval lengths rather than at a single value, so this study adopted 37s within this range as the judgement interval.

Extending the interval to 67s lowers the accuracy to 73.80%, and extending it to 127s lowers it to 46.13%. The decrease occurs entirely in the main steam system, and the circulating water system stays at 1.000 for every interval.

The difference between the two plant systems comes from how the abnormal event develops. When the circulating water pumps trip, the condenser vacuum stays degraded, so that plant system shows the largest deviation regardless of time. In contrast, when a main steam line breaks, the steam side variables change immediately, but the deviation of the feedwater system is a response that follows the change in the heat removal condition of the steam generator, so it grows with time. A longer judgement interval includes this response, and the responding plant system exceeds the deviation of the root cause plant system.

5. Conclusions

This study proposed an open set abnormal event diagnosis method based on system-wise autoencoders in order to address untrained abnormal events in a small modular reactor. The learning layer divides the 116 variables by plant system and trains a one-dimensional convolutional autoencoder independently for each subset, and it judges whether an input is trained by using the reconstruction error of each plant system. An Inception classifier classifies an input judged as trained into one of the trained scenarios, and the learning layer does not pass an input judged as untrained to the classifier.

For an untrained input, the method identifies the root cause system by a score obtained from the reconstruction error of each plant system through two stages of normalization.

This study evaluated the proposed method with data collected from an iPWR simulator. The method judged 94.11% of the untrained scenario windows as untrained and did not pass them to the classifier, and the false rejection rate for the trained scenario windows was 2.70%. The AUROC of the decision score is 95.44%. For the scenario classification, the balanced accuracy was 99.65% and the micro F1 was 99.75%, and every scenario recall was 98.60% or higher. AKS is 97.06% and the H-score, which is the harmonic mean of AKS and AUS, is 95.56%. The accuracy of the root cause system identification at the file level is 99.63% and the mean accuracy over the plant systems is 99.80%.

The proposed method provides the classification of trained scenarios, the judgement of untrained abnormal events with the blocking of their misclassification, and the identification of the root cause system within a single framework. By delivering information about the plant system requiring action to the operator even in abnormal situations that the training data does not cover, the proposed method is expected to contribute to improving plant safety in a small modular reactor operating environment where a small crew monitors multiple modules.

Acknowledgment

This research was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (No. GTL2403Y-000) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-02353161)

REFERENCES

- [1] J. Li, M. Lin, B. Wang, R. Tian, S. Tan, Y. Li, and J. Chen, Open Set Recognition Fault Diagnosis Framework Based on Convolutional Prototype Learning Network for Nuclear Power Plants, *Energy*, Vol.290, p. 130101, 2024.
- [2] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P. A. Muller, and F. Petitjean, InceptionTime: Finding AlexNet for Time Series Classification, *Data Mining and Knowledge Discovery*, Vol.34, p. 1936, 2020.
- [3] C. Geng, S. J. Huang, and S. Chen, Recent Advances in Open Set Recognition: A Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.43, p. 3614, 2021.
- [4] P. R. Mendes Junior, R. M. de Souza, R. de O. Werneck, et al., Nearest Neighbors Distance Ratio Open-Set Classifier, *Machine Learning*, Vol.106, p. 359, 2017.
- [5] S. Bucci, M. R. Loghmani, and T. Tommasi, On the Effectiveness of Image Rotation for Open Set Domain Adaptation, *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.