Feature Selection Using Machine Learning for Complex Abnormal Event Diagnosis

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*

1. Introduction

Operators have to recognize large and small abnormal problems occurring in nuclear power plants (NPPs) and take appropriate measures to prevent them from deteriorating the condition of NPPs. Since an NPP operates by the interrelationship of its constituent components, even if a problem that is not taken action does not immediately cause an accident in operation, it can bring about a result that leads to the shutdown of the reactor later. Therefore, an artificial intelligence model developed to support operator diagnosis task in abnormal situations must be able to perform detailed diagnosis even for complex events. However, most previous studies have been conducted with the goal of diagnosing simple abnormal events. In addition, there are physical limitations in conceiving and acquiring scenarios for all complex abnormal events for training of artificial intelligence models.

In this study, we tried to propose a model for diagnosing complex abnormal events with only scenarios for single abnormal events. The proposed model applies a method of performing characteristic feature selection by using a machine learning in advance to diagnose each abnormal event even in complex events. The proposed approach enabled diagnosis of complex abnormal events with higher performance compared to the existing model.

2. Methods

In an NPP abnormal event, a problem usually occurs in a specific component or system, and the related parameters are affected. Therefore, it is important to select major parameters among thousands of parameters in order for the artificial intelligence model for abnormal event diagnosis to effectively learn NPP data. The method for feature selection is introduced below.

2.1 Feature Selection Using Model

As a method of feature selection, there is a method of deriving a feature set showing high performance in a machine learning model. Below, two machine learning models are introduced, and important features can be selected through the feature importance of these models.

- Extremely randomized trees classifier (Extra trees classifier)
  Extra trees increase randomness by randomly splitting each candidate feature in the forest trees [1]. It is similar to the existing random forest trees classifier, but the splitting approach is different, and it can be ensemble many trees to increase the classification accuracy for the validation dataset.
- LightGBM
  LightGBM is a tree-based learning algorithm using the gradient boosting framework [2]. Since this is an algorithm that uses leaf-wise expansion, it can reduce more loss compared to the tree model using the existing level-wise expansion.

2.2 Selected Feature Number Using Model Training

It is important to provide appropriate information to improve the performance of the model. Therefore, it is necessary to limit the amount of information given in the data. To this end, the number of features with high importance may be specified in advance, but a specific mathematical criterion such as average importance may be used. In addition, there is a method of selecting features for the moment when the model has the highest performance. Recursive feature elimination with cross validation (RFECV) removes an unimportant feature (feature with the lowest feature importance) one by one backward while repeating model training [3]. Next, by calculating the performance using cross validation, the number of features that can secure the highest performance is checked.

3. Experimental Setup

Even in complex events in NPPs, the model must be able to recognize individual events. Below, we introduce the data used for training and test of the model and approach of the model.

3.1 Abnormal Event Dataset for Model Training

For model learning, we acquired data using a general pressurized water reactor-based 3KEYMASTER simulator provided by Western Services Corporation [4]. To consider the relevance of the two events later, eight abnormal events were selected as follows. For a training dataset, 25 scenarios of various intensities were obtained for each abnormal event, for a total of 200 scenarios. The intensity of the abnormal event is determined by values such as valve position, degree of leak, and size of tube rupture. Each scenario was sampled for 60 seconds for 797 parameters.

Table I: Abnormal Event Description for Dataset

Label	Abnormal event description
CHRG	Charging line break upstream of FT-121
LTDN	Letdown line leak inside containment
CDS	Loss of condenser vacuum
POSRV	Pilot operated safety relief valve (HV456A) leak
CWS	Circulating water tube leak in low-pressure condenser
RCP	Reactor coolant pump seal injection water loss by stucking of valve (HV8351A)
PZR	Pressurizer spray valve (PV455B) open by positioner failure
CCW	Component cooling water service loop header leak to aux atm

3.2 Abnormal Event Dataset for Model Test

The test dataset for single abnormal events was acquired with a total of 200 scenarios, 25 for each abnormal event with a different intensity from the training dataset. In order to confirm that the model can have high diagnostic accuracy for complex events, it is also necessary to use complex event scenarios that include low-intensity abnormal events. For example, a small degree of line leak will just result in a small parameter change, but the model should be able to detect it. To this end, as shown in the figure below, a complex event scenario was created considering the combination of abnormal events and the combination of intensities within the abnormal events. A complex event dataset for model test can consider a total of 28 combinations of abnormal events by considering the simultaneous occurrence of two out of eight abnormal events. Complex event dataset was acquired with a total of 700 scenarios, 25 for each abnormal event combination.

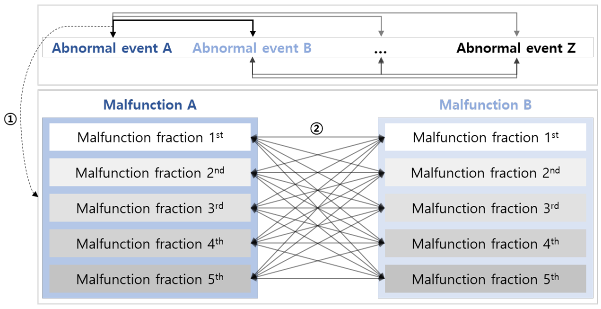

Fig. 1. Consideration of Complex Abnormal Event Scenario

3.3 Proposed Algorithm

We proposed a model for diagnosing the occurrence of each abnormal event in complex events as shown in the figure below. Obtain the feature importance in the machine learning model that can diagnose each abnormal event through the training dataset. Then, based on the importance of the acquired features, the main parameters for diagnosing the occurrence of each abnormal event are selected according to the given criteria such as mean Gini importance. Next, the training dataset is preprocessed through a set of key parameters for each abnormal event, and used to train a binary classification model that diagnoses whether the target event has occurred. sub models that perform binary classification output 1 when the target abnormal event is detected for a given evaluation scenario, and 0 when the target abnormal event is not detected. The results of each sub model are voting on to derive the final diagnosis result. In this study, a convolutional neural network model with one layer was used as sub model structure, and the hyperparameters are as follows.

- Filter number of convolution layer : 32
- Kernel size of convolution layer : 3
- Activation function of convolution layer : ReLU
- Activation function of dense layer : softmax
- Loss function : binary crossentropy
- Optimizer : Adam [5]
- 100 epochs using early stopping monitored validation loss with 10 patience

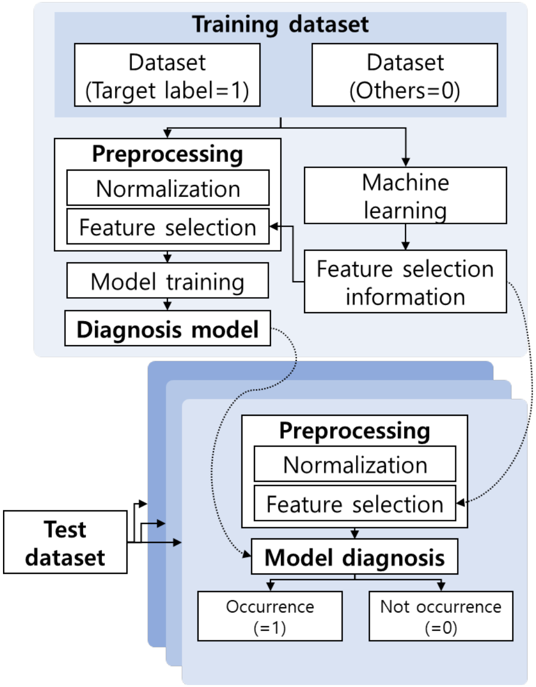

Fig. 2 Proposed Approach for Complex Abnormal Events

4. Results

In this study, Extra trees classifier and LightGBM were used as machine learning models for feature selection. In addition, a method of selecting parameters when the feature importance is above mean Gini importance and a method of automatically selecting parameters through RFECV were used. As shown in the figure below, when less important features were removed one by one, the final number of features was determined at the point where the cross-validation score was highest. The number of parameters finally selected is shown in the table below, and in the RFECV method, the number of key parameters for most abnormal events was derived as the highest model performance at the minimum setting step.

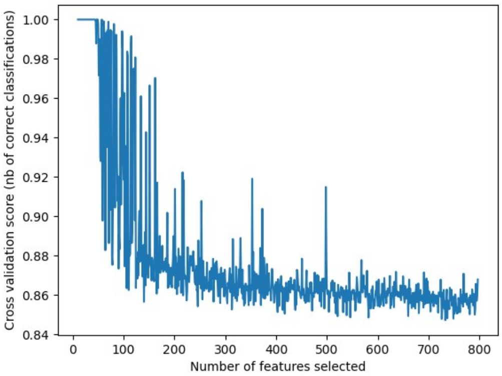

Fig. 3. Cross Validation Score with RFECV for RCP label

Table II: Selected Feature Number at Each Model

Label	Extra trees		LightGBM	
	Mean	RFECV	Mean	RFECV
CHRG	77	10	38	10
LTDN	136	10	93	10
CDS	93	10	48	10
POSRV	212	10	45	10
CWS	111	10	79	91
RCP	186	10	49	10
PZR	242	13	50	10
CCW	103	504	88	11

The table below shows the diagnostic accuracy of the test dataset for the models to which each method for feature selection is applied. As a result, the proposed approach showed approximately 99% diagnostic accuracy for single abnormal events. In addition, it showed more than 85% diagnostic accuracy even for complex abnormal events.

Table III: Test Accuracy with Each Model

Feature selection		Abnormal event of test dataset	
Model type	Selected type	Single	Complex
Extra trees	Mean	99.02 %	85.49 %
Extra trees	RFECV	99.31 %	94.92 %
LightGBM	Mean	99.28 %	86.69 %
LightGBM	RFECV	98.68 %	85.80 %

In particular, the table above showed the highest accuracy of 94.92% for complex abnormal events when the feature selection method using the extra trees classifier as a machine learning model with RFECV was applied. The table below compares the results of the proposed model and the general CNN model to which our algorithm is not applied.

Table IV: Performance Improvement from Based Model

	Based model (A)	Proposed model (B)	Improvement (B-A)
Single	99.22 %	99.31 %	0.09 %p
Complex	62.56 %	94.92 %	32.36 %p
CHRG	65.42 %	95.67 %	30.25 %p
LTDN	67.14 %	93.59 %	26.45%p
CDS	65.89 %	96.01 %	30.12 %p
POSRV	57.33 %	91.67 %	34.33 %p
CWS	81.62 %	99.11 %	17.50 %p
RCP	73.45 %	99.02 %	25.57 %p
PZR	14.47 %	89.19 %	74.72 %p
CCW	75.14 %	95.11 %	19.97 %p

Compared to the existing model, the proposed approach not only maintained the accuracy for single abnormal events, but also improved the accuracy to 32.36 % points for complex abnormal events. Among them, when the CHRG event and the PZR event occurred simultaneously, the existing model could not diagnose the PZR event at all. However, the proposed model detected both events with a diagnostic accuracy of 98.27%. In addition, the proposed model improved the accuracy by 91.07% points compared to the existing model when the RCP event and the PZR event occurred simultaneously. We confirmed that in the proposed approach, the extra trees classifier selected 7 out of 10 key parameters for diagnosing PZR events as parameters for 'Pressurizer pressure' and 'Pressurizer level'.

5. Conclusions

An operator support system in diagnosing abnormal events in NPPs should be able to secure diagnostic performance for particularly complex events. In this study, two machine learning models were used to select features for detecting each abnormal event. In addition, each sub model trained only whether the target abnormal event occurs through selected parameters. When parameters were selected through RFECV using the extra-trees classifier as a machine learning model, it showed the highest composite event diagnosis performance at 94.92 %. In further studies, we have to conduct a comparison analysis and sensitivity analysis using other feature selection methods and classification models. In addition, it is required to verify by extending it to a larger number of abnormal events to apply the proposed approach to actual NPPs.

ACKNOWLEDGEMENT

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2022-00144042).

REFERENCES

[1] Geurts, P., Ernst, D. and Wehenkel, L., 2006. Extremely randomized trees. *Machine learning*, *63*, pp.3-42.
[2] 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.
[3] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine learning, 46, pp.389-422.
[4] 3KEYMASTER Simulator 2013, Western Service Corporation, Frederick, MD, USA.
[5] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).