Enhancing the Reliability of Abnormal State Diagnosis in Nuclear Power Plants Using Explainable Boosting Machine

Ho Jun Lee, Ji Hun Park, Ji Woo Hong, Man Gyun Na
Department of Nuclear Engr., Chosun Univ., 10 Chosundae 1-gil, Dong-gu, Gwangju, Republic of Korea
**Corresponding author: magyna@chosun.ac.kr*

1. Introduction

Nuclear Power Plants (NPPs) can experience abnormal state due to a variety of causes. If an abnormal state is not caught promptly, it may result in economic loss or in severe cases of dangerous accidents such as core damage and radiation leakage. When an abnormal state occurs, the operators must quickly diagnose the condition and take appropriate actions. During this process, there are various monitoring variables that the operator must identify. Various monitoring variables that need to be identified in the process of quickly diagnosing and taking actions can lead to human error by operators.

Recently, there have been a lot of researches on anomaly detection and diagnosis using Artificial Intelligence (AI) to help operators make decisions. However, existing AI models cannot solve the explainability-accuracy trade-off at once. This means that the accuracy is in conflict with explainability. The problem is that even if an AI makes an accurate diagnosis, it cannot be trusted if it cannot explain why it made that diagnosis. Therefore, in this study, we used an Explainable Boosting Machine (EBM) to address accuracy and explainability at once. Post-processing a high-accuracy black box model with Shapley Additive exPlanations (SHAP) can make the model explainable. However, this is not a direct way to explain the model.

In this paper, we verified the performance and explainability of the model for diagnosing abnormal states using EBM. We also compared the performance of the black-box models of Light Gradient Boosting Machine (LightGBM) and Deep Neural Network (DNN). In addition, we applied SHAP to each black box model and compared the EBM with them. We conducted this study to see if we can provide faster information for real-time diagnostics.

2. Method and Data

This section describes the methodology and data used in this study.

2.1 Explainable Boosting Machine

EBM is one of the most recent eXplainable AI (XAI) models to solve the problematic explainability-accuracy trade-off of AI models. XAI is classified into two type: ante-hoc and post-hoc. Ante-hoc is to extract explainable features or rules together during the model training process. Post-hoc takes a trained black-box model as input and queries it to obtain the underlying relationships the model has learned. Post-hoc XAI algorithms include SHAP and Local Interpretable Model-agnostic Explanation (LIME). These methods require a black-box model to be created before they can be applied. On the other hand, EBM may learn training data while extracting descriptive features. Therefore, EBM is a method that belongs to ante-hoc. EBM with these features has high explainability and performs as well as black box models.

EBM is an extension of Generalized Additive Model (GAM), a tree based recursive gradient boosting generalized additive model. The traditional GAM model can be represented as shown in Eq. (1):

$$g(E[y]) = \beta_0 + \sum f_r(x_r) \tag{1}$$

where g is the link function, $E[y]$ is the dependent variable, x_r is the explanatory variable, and f_r is the feature function. The link function is used to model the relationship between the dependent variable and the explanatory variables. The feature function is used to describe the relationship between the dependent variable and the explanatory variables. These two functions are used to model the non-linear relationship between the dependent and explanatory variables. This GAM is good at representing the non-linear relationship between feature and label values [1]. However, GAM has the disadvantage of not being able to represent pairwise interactions between features. EBM has made improvements to these GAM. The f_r feature function is trained using bagging and recursive gradient boosting on a traditional GAM. During the boosting step, we additionally employ round-robin cycles to ensure that only one feature is learned at a time. Round-robin means learning in order, with no prioritization between features. In this way, we cycle through each feature to mitigate the effects of co-linearity. It also trains the best for each feature to tell you how much each feature contributed to the model classification. And EBM automatically detects and includes pairwise interactions in the form of the Generalized Additive Model plus interactions) GA2M algorithm, an improved model of GAM. This provides greater accuracy while maintaining a clear and distinct nature. The GA2M algorithm can be represented by Eq. (2).

$$g(E[y]) = \beta_0 + \sum f_r(x_r) + \sum f_{cr}(x_c, x_r) \quad (2)$$

GA2M is a model in which a pairwise interaction term is added to express the interaction between features in GAM [2]. GA2M contains all pairwise interaction terms. EBM includes only the top N pairwise interaction terms that are added to the model. In this way, EBM is a fast implementation of the GA2M algorithm [3].

2.2 Data information

In this study, data was collected using a Compact Nuclear Simulator (CNS). CNS is a simulator based on a three-loop pressurized water reactor manufactured by Westinghouse.

Seven abnormal scenarios and one normal scenario were collected through CNS. Table I provides information about the collected scenarios.

Table I: Scenario information

Number	Scenario
0	Normal
1	PRZ water level channel failure 'high'
2	PRZ water level channel failure 'low'
3	PRZ pressure channel failure 'high'
4	PRZ pressure channel failure 'low'
5	PRZ spray valve failure 'opening'
6	PRZ PORV 'opening'
7	Loss of coolant accident
PRZ: pressurizer PORV: power operated relief valve	

The collected data is CNS data consisting of about 2200 variables. However, variables that are unnecessary for classification may rather complicate the model and hinder a good performance. Therefore, only important variables were extracted from about 2200 variables. The variables selection method was to extract the relevant variables for each scenario and the variables related to the abnormal operation procedure. For example, for the PRZ PORV 'opening' scenario, we used PORV valve, PRZ pressure, temperature, water level, back-up heater, etc. In this way, a total of 23 variables were selected. Table II shows the information on the extracted variables.

Table II: Selected variables

Number	Scenario
ZPRTL	PRT water level
ZINST66	PRZ spray flow
ZINST65	PRZ pressure (wide range)
ZINST63	PRZ level
ZINST62	PRZ temp
ZINST48	PRT pressure
ZINST39	Charging flow
ZINST38	Letdown flow
ZINST26	Containment pressure
ZINST25	Containment temp
ZINST23	Containment relative humidity
ZINST22	Containment radiation
WSPRAY	PRZ spray flow from RCS loop
WPRZSV	PRZ safety valve flow
QPRZP	Proportional heaters power
QPRZH	Proportional heater fractional power
QPRZB	Back-up heaters power
PVCT	VCT pressure
PPRZ	PRZ pressure
PPRT	PRT pressure
BPRZSP	PRZ spray valve position
BPORV	Power operated relief valve position
BLV459	Letdown isolation valve position
BHV6	HV6 valve position
WPORV	PRZ PORV flow rate to PRT
VCT: Volume Control Tank PRT: Pressurizer Relief Tank RCS: Reactor Coolant System	

3. Results

In this study, in order to compare and evaluate the performance of EBM, the performance of the models that applied SHAP to each classification model generated by LightGBM and DNN was also evaluated. LightGBM is a tree-based machine learning model that is as accurate as the traditional GBM and 20 times faster [4]. This model is useful for processing large amounts of data. DNN is an artificial neural network consisting of an input layer, a hidden layer, and an output layer. DNN can process a variety of data, including images, speech, and text. SHAP is a framework for describing the output of a black-box model using Shapley values [5].

Performance evaluation was carried out in three ways: 1) accuracy, 2) macro f1-score, and 3) XAI execution time of each model. The macro f1-score used as an evaluation index can resolve data imbalances that cannot be confirmed with accuracy. Macro F1-score is a method of extracting and averaging f1-scores for each class. F1-score is one of the evaluation indicators composed of the recall and precision. The recall and precision are composed of true positive (TP), false negative (FN), and false positive (FP). Accuracy is also used with true negative (TN) including TP, FN and FP. Recall, precision, f1-score, macro f1-score, and accuracy can be expressed by Eq. (3)-(7).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$f1_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

$$\text{Macro } f1_{\text{score}} = \frac{1}{n}\sum_{k=0}^{n} f1_{\text{score}}(k) \quad (6)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

We checked the execution time of XAI for each model. The standard for execution time is the time required to analyze one test data. One test data corresponds to 1 second of CNS data. Table III shows the accuracy and macro f1-score for each model. Table IV shows the running time of XAI.

Table II: Performance evaluation by each model

	Evaluation metrics	
Model	Accuracy	Macro f1-score
EBM	1.00	1.00
LightGBM	1.00	1.00
DNN	1.00	1.00

Table IV: XAI runtime for each model

Model	XAI runtime
EBM	0.006sec
LightGBM	0.005sec
DNN	0.647sec

In addition, EBM can explain which features the model considers important and which classes each feature is highly involved in. Fig. 1 shows which features the model considers important by selecting the top 15 features. Figs. 2 and 3 explain which classes the features are involved in. Figs. 2 and 3 shows the top 2 features from the features in Fig. 1.

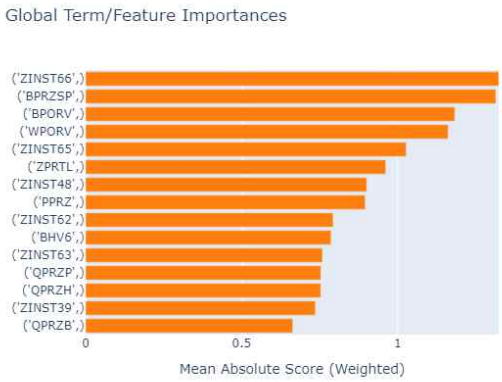

Fig. 1. The top 15 variables that the trained model considers important.

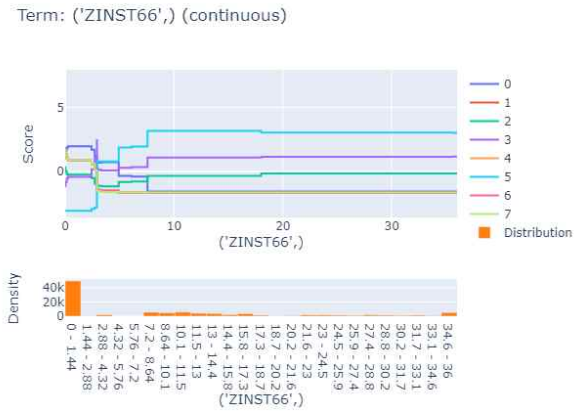

Fig. 2. The variable that the trained model considers most important.

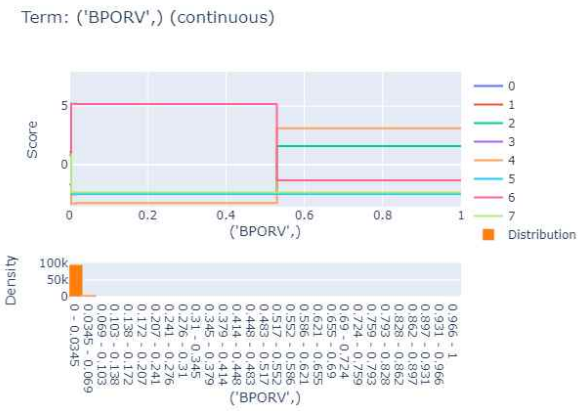

Fig. 3. The Variables that the trained model considers second most important.

The score in Figs. 2 and 3 means the contribution of a feature to a class. Density indicates where the values of each variable are distributed in the train data. Figs. 4 and 5 illustrate how much each feature contributed to

each class when predicting the data for scenarios 1 and 7.

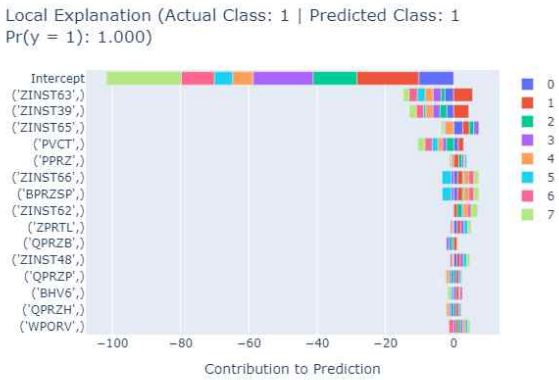

Fig. 4. Explanation of scenario 1 predictions.

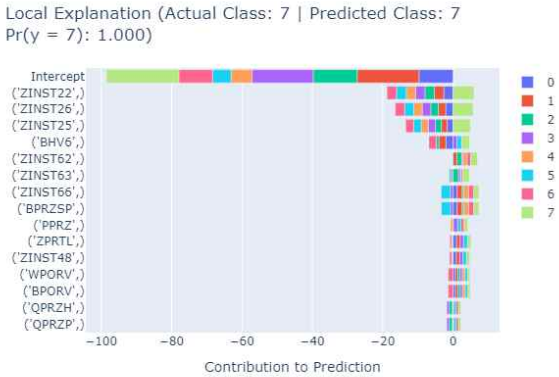

Fig. 5. Explanation of scenario 7 predictions.

EBM can see how much each feature contributed to each class when predicting the scenario. In Figs. 4 and 5, we can see that each class prioritized the most important features.

4. Conclusions

In this paper, we compared the performance of EBM, LightGBM, and DNN models. In addition, we compared the XAI execution time of the EBM and the models with SHAP added to the LightGBM and DNN models to compare the difference when diagnosing abnormal conditions in real time. EBM showed the same accuracy as the black box model when compared to other black box models. We also found that the XAI running time of the EBM was 0.641 sec faster than the XAI running time of the model applying SHAP to the DNN model. When SHAP was applied to LightGBM, the XAI execution speed was 0.001sec faster than EBM, but it was a very small difference. EBM also showed high accuracy and high explainability when compared to each model. We were also able to see how EBM diagnosed the abnormal state and what variables were used to diagnose the abnormal state. This improves the intuitiveness of the model and solves the trade-off between explainability and accuracy, which is a problem of existing AI models. Therefore, the proposed EBM model can help NPP operators to trust AI diagnosis results in the process of diagnosing abnormal states. If this model is applied, it is expected that operators can be able to improve the safety of NPPs through quicker diagnosis.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (Grant No. NRF-2018M2B2B1065651).

REFERENCES

[1] T. Hastie and R. Tibshirani, Generalized Additive Models: Some Applications, Journal of the American Statistical Association, Vol. 82, No. 398, pp. 371-386, 1987.
[2] Y. Lio, R. Caruana, J. Gehrke and G. Hooker, Accurate Intelligible Models with Pairwise Interactions, 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 623-631, 2013.
[3] H. Nori, S. Jenkins, P. Koch and R. Caruana, InterpretML: A Unified Framework for Machine Learning Interpretability, Vol. abs/1909.09223, 2019
[4] G. Ke et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in Advances in Neural Information Processing Systems (NIPS 17), Vol. 30, pp. 3146-3154, 2017.
[5] S.M. Lundberg and S.-I Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems (NIPS '17), Vol. abs/1705.07874, pp. 4765-4774, 2017.