Feasibility Study of XAI Application to Severe Accident Uncertainty Analysis – Model Development

Semin Joo [a], Tae-woo Kim [b], Miro Seo [b], Jeong Ik Lee [a*]
[a]*Dept. Nuclear & Quantum Eng., KAIST, Yuseong-gu, Daejeon, Republic of Korea*
[b]*Korea Hydro & Nuclear Power Co., Central Research Institute, Yuseong-gu, Daejon, Korea*
[*]*Corresponding author: jeongiklee@kaist.ac.kr*

***Keywords :** Explainable AI, severe accident, uncertain parameter, data augmentation

1. Introduction

Severe accident analysis contains large uncertainty because the accident progression is driven by coupled thermal-hydraulic, chemical, and structural phenomena. Experimental data are limited, and plant-scale behavior must be assessed mainly through system analysis codes with modeling and numerical constraints. As a result, uncertainty studies commonly assume probability distributions for uncertain input parameters, generate a limited set of samples using Latin Hypercube Sampling (LHS) or Monte Carlo method, run a severe accident code for each case, and then evaluate the relationship between a figure of merit (FOM) and each input parameter [1][2][3].

However, two practical gaps remain. First, the dataset size is often restricted to around 100 code runs because it is selected using Wilks' formula [4]. Wilks' formula determines the minimum number of simulations required to estimate statistical tolerance limits, not to guarantee reliable sensitivity or importance ranking across tens of uncertain parameters. When the number of uncertain parameters is large, correlation-based screening with ~100 samples can be unstable and can miss important effects.

Second, correlation-based screening has a structural limitation: the FOM response is frequently non-linear and can show threshold/regime behavior. Also, the inputs can interact because they represent dependent phenomena such as oxidation, relocation path, and structural integrity. In such conditions, correlation coefficients can under represent the true influence of inputs and cannot directly quantify interaction driven effects.

To address these gaps, the present study combines a data augmentation step with explainable AI (XAI). To relax the sample-size limitation for surrogate modeling and sensitivity interpretation, a data augmentation approach is applied to the MAAP cases using Dirichlet-weighted mixing. The main contribution is the XAI-based analysis: a regression surrogate is trained to predict the FOM, and Shapley Additive exPlanations (SHAP) are used to quantify non-linear main effects and interactions.

2. MBLOCA Analysis Result

In the previous study [1], a total of 49 uncertain input variables defined in the MAAP5 code were selected to account for model uncertainties affecting reactor pressure vessel (RPV) failure under MBLOCA accident conditions. Based on engineering judgment, 49 variables that could directly influence RPV failure behavior were selected, and the range and probability distributions were defined.

Latin Hypercube Sampling (LHS) was used to generate 100 samples that represent the defined distributions of the 49 variables. These samples were used to evaluate the impact of uncertain parameters on RPV failure behavior.

The RPV failure behavior was analyzed for an OPR1000 under MBLOCA without operator mitigation actions. The scenario assumed a 6-inch break in the cold leg with no operation of the safety injection system. MAAP5 simulations were performed for a total of 100 uncertainty samples, and RPV failure occurred in all cases.

Table I summarizes the timing of major events. The RPV failure time showed a relatively large variation around 1 hour. This indicates that even under the same accident scenario, the timing of RPV failure can vary significantly depending on the uncertain parameters.

Table I. Major event timing

Event	Time (sec)
Initiation of MBLOCA	0.0
Reactor scram	12.821
SAMG entrance	6696.0~6845.3
Core relocation	9282.7~10673.9
RPV failure	13433.37~34928.44

3. Methodology

3.1 Data Augmentation

Although RPV failure occurred in all cases, the total of 100 samples is insufficient for training a machine learning model and making a reliable explanation on that model. Explainable AI (XAI) is a tool that reveals how a machine learning model processes the training data and

identifies which variables are considered important for prediction. Hence, the reliability of XAI depends on whether the model has undergone sufficient training. Therefore, data augmentation was employed to increase the amount of training data without performing more number of simulations with MAAP code.

Data augmentation refers to a technique that artificially increases the quantity and diversity of training data by applying various transformations to existing samples. In this study, a Dirichlet-based mix-up method was adopted. Original samples were linearly combined using weights drawn from a Dirichlet distribution. Since this approach performs interpolation within the range defined by the minimum and maximum values of the original data, unrealistic outliers are not generated. Among the 49 variables, categorical variables in binary form (0/1) were excluded from the weighted combination. To ensure similarity to the original MAAP data, augmented samples with distributions that deviated significantly from the original data were discarded. By generating augmented samples equal to twice the size of the original dataset, a total of 300 samples were obtained.

$$x_{aug} = \sum_{i=1}^{k} w_i x_i \qquad \text{Eq. 1}$$

$$y_{aug} = \sum_{i=1}^{k} w_i y_i \qquad \text{Eq. 2}$$

Fig. 1 illustrates how the distributions change before and after mix-up using two representative variables. FGBYPA is defined as “a flag to divert gas flows in the core to the bypass channel when an entire axial row in the core is completely blocked”. Since FGBYPA is a categorical variable, it was excluded from the mix-up process, thereby preserving its original distribution.

In contrast, XGAP0 is “the initial size of the gap between the debris and the inner surface of penetrations in the lower head” and is a continuous variable. As it was subjected to the mix-up process, the number of samples with intermediate values increased. Nevertheless, the overall triangular distribution shape was preserved.

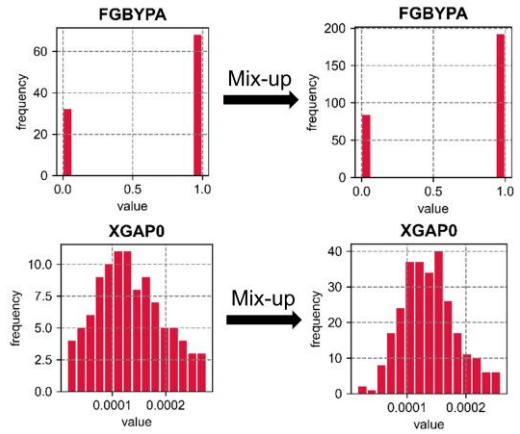

Fig. 1. Example of uncertain parameter (FGBYPA, XGAP0) distribution before and after mix-up

3.2 Machine Learning

In this study, XGBoost was adopted as the regression model architecture for predicting the RPV failure time. XGBoost constructs a strong predictive model by combining multiple weak learners. It is known to be effective in capturing rule-based patterns in structured numerical data with clearly defined rows and columns.

The input to the XGBoost model consists of 49 uncertain parameter values, and the output is the RPV failure time obtained when these parameter values are applied to the MAAP input. Since the regression performance of the XGBoost model depends on the hyperparameter settings, hyperparameter optimization was performed using a grid search method. The hyperparameter combination that yielded the best regression performance is presented in Table II.

Table II. The best hyperparameters of the XGBoost model

Hyperparameters	
n_estimators	3000
max_depth	3
min_child_weight	1
learning_rate	0.3
subsample	1
colsample_bytree	0.8
reg_lambda	10
reg_alpha	0

3.3 XAI

In this study, SHAP (SHapley Additive exPlanations), one of the explainable artificial intelligence (XAI) techniques, was applied to interpret the predictions of the machine learning model. SHAP quantitatively evaluates the contribution of each input variable to the model prediction and is based on the Shapley value concept from cooperative game theory [6]. The Shapley value provides a fair allocation of the contribution of each participant when multiple participants jointly produce an outcome. When applied to machine learning, it quantifies how much each input variable contributes to increasing or decreasing the prediction. A positive SHAP value indicates that the variable contributes to increasing the predicted value, while a negative SHAP value indicates that the variable contributes to decreasing the predicted value. This enables interpretation of the model prediction based on the actual contribution of each variable rather than simple correlation.

In this study, an XGBoost regression model was trained using 49 uncertain parameters as inputs and the RPV failure time as the output. The TreeExplainer algorithm was then used to compute the SHAP value for each variable and each sample. This enabled evaluation of both the magnitude and direction of the influence of each uncertain parameter on the RPV failure time.

4. Results and Discussions

4.1 Regression Performance

Table III compares the test performance of the XGBoost surrogate model trained on the MAAP data before and after augmentation. When the Dirichlet-based mix-up augmentation was applied, MAE, RMSE, and MAPE all increased, indicating that the prediction performance did not improve.

Table III. Regression performance before and after data augmentation

	MAE (sec.)	RMSE (sec.)	MAPE (%)
Before augmentation	324.7051	410.6226	2.145051
After augmentation	424.7341	624.3595	2.807433

The augmentation method adopted in this study inherently assumes that if the input variables are linearly interpolated, the output variable is also linearly interpolated. However, the relationship between the inputs (uncertain parameters) and the output (RPV failure time) exhibits strong nonlinearity and complex interactions among variables, which cannot be preserved under a simple interpolation assumption. The observed degradation in regression performance after augmentation supports the presence of such nonlinear relationships. This result highlights that the problem cannot be adequately interpreted using simple statistical correlation analysis alone.

4.2 Comparison of Statistical Correlation vs. SHAP

Fig. 2 presents the top 10 variables ranked by SHAP-based importance. The color of each point represents the magnitude of the variable value, and the x-axis represents the SHAP value. Among the top 10 variables, FGBYPA was identified as the most dominant factor, followed by FUPOOL, XGAPLH, FZORUP, and LMCOL1.

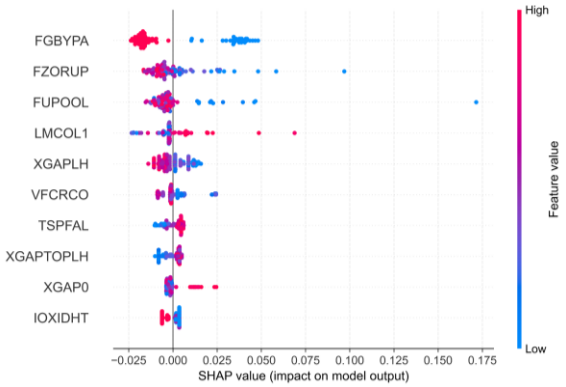

Fig. 2. SHAP Summary of Top 10 uncertainty parameters

In the previous study [1], the relationships between uncertain parameters and the FOM were analyzed using various statistical correlation metrics, including PCC (Pearson Correlation Coefficient), PRCC (Partial Rank Correlation Coefficient), SRC (Standardized Regression Coefficient), and SRCC (Spearman Rank Correlation Coefficient). Among these, PRCC and SRC were selected for comparison because they are conceptually more comparable to SHAP in that they quantify the relative influence of input variables on the output. The top 10 variables were identified based on PRCC and SRC rankings, and the overlap between these statistically derived variables and those identified using SHAP was examined.

Fig. 3 presents a Venn diagram of the top 10 variables identified by SHAP, SRC, and PRCC analyses. SRC represents the contribution of each variable in a linear regression model [5]. PRCC evaluates the pure relationship between a specific variable and the outcome after removing the linear rank effects of other variables. Both correlation-based metrics are valid only when the relationships between variables are monotonically increasing or decreasing. The variables commonly identified across all three metrics were FGBYPA, FUPOOL, and VFCRCO. These variables were consistently evaluated as dominant factors influencing the RPV failure time.

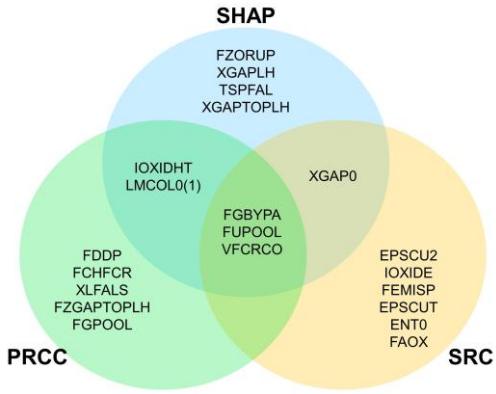

Fig. 3. Venn diagram comparing the Top 10 parameters identified by three sensitivity analysis methods: SHAP, PRCC, and SRC.

Fig. 4 presents the SHAP values of FGBYPA, which ranked first across all sensitivity analysis metrics, along with the distribution of its values. FGBYPA is defined as a flag to divert gas flows in the core to the bypass channel when an entire axial row in the core is completely blocked. When the value is 1, gas flow is diverted, resulting in reduced steam cooling capability in the core.

A clear pattern is observed in Fig. 4. When FGBYPA=0, the SHAP values are largely positive, whereas when FGBYPA=1, the SHAP values are largely negative. Compared to other uncertain parameters, the SHAP values of FGBYPA exhibit a highly clustered distribution. This indicates that the presence or absence of gas flow diversion has a dominant influence on the

RPV failure time, while the interaction effects with other variables are relatively limited.

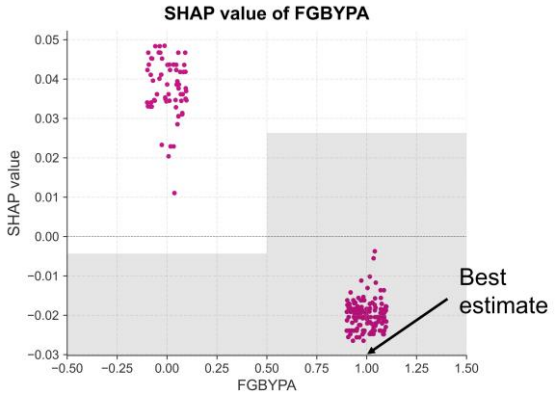

Fig. 4. SHAP value of FGBYPA with parameter distribution

4.3 SHAP Interaction Results

The total SHAP contribution of a specific variable i, denoted as ϕ_i, can be decomposed into its main effect (ϕ_{ii}), which represents the independent contribution of the variable, and the interaction effects (ϕ_{ij}) arising from its interactions with another variable j (Eq. 3). This enables analysis beyond identifying which variable is important, and reveals how the prediction changes when two variables vary together. Such analysis is necessary because certain prediction behaviors cannot be explained solely by individual variable importance or correlation coefficients.

$$\phi_i = \phi_{ii}(\text{main}) + \sum_{i \neq j} \phi_{ij} \ (\text{interaction}) \qquad \text{Eq. 3}$$

Table IV summarizes the top 10 parameter pairs with the highest SHAP interaction effect. Among them, the top 3 parameter pairs were analyzed in detail. Figs. 5-7 present SHAP dependence plots illustrating the interactions between the uncertain parameter pairs. Each point represents an individual simulation case, where the horizontal axis indicates the value of one parameter, the vertical axis indicates its SHAP value. The color of each point represents the value of the interacting parameter.

Table IV. Parameter pairs with the top 10 highest interaction effect

Rank	Parameter 1	Parameter 2	Interaction effect (ϕ_{ij})
1	FUPOOL	LMCOL1	0.002902
2	FGBYPA	XGAPLH	0.002216
3	FZORUP	VFCRCO	0.002211
4	FUPOOL	LMCOL0	0.001627
5	FGBYPA	FZORUP	0.001613
6	XGAPLH	VFENT	0.001221
7	FUPOOL	ENT0	0.000947
8	XDJETO	XGAP0	0.000874
9	FGBYPA	IOXIDHT	0.000869
10	FWHL	FUPOOL	0.000830

Fig. 5 shows the SHAP dependence plot for FUPOOL and LMCOL1. A clear increase in the SHAP value of FUPOOL is observed when FUPOOL is small and LMCOL1 is large. This indicates that when both variables satisfy specific conditions simultaneously, a significant delay in the RPV failure time occurs.

Fig. 6 presents the interaction effect between FGBYPA and XGAPLH. When FGBYPA is 0, the SHAP value decreases more clearly as XGAPLH increases.

Fig. 7 shows the interaction effect between FZORUP and VFCRCO. Large positive SHAP values are observed when both FZORUP and VFCRCO are small, corresponding to conditions where the RPV failure time is significantly delayed.

These results demonstrate that not only the individual effects of variables but also their interaction effects play an important role in determining the RPV failure time. The importance of major variables was generally consistent with the statistical analysis results. However, the SHAP analysis revealed nonlinear behavior and large contributions under specific conditions that could not be explained by simple correlation analysis alone. In particular, certain combinations of variables resulted in significant delays or accelerations in the failure time due to interaction effects. This indicates that the influence of individual variables may be overestimated or underestimated when interaction effects are not considered. Therefore, evaluating the impact of uncertain parameters requires consideration of both individual variable effects and their interactions.

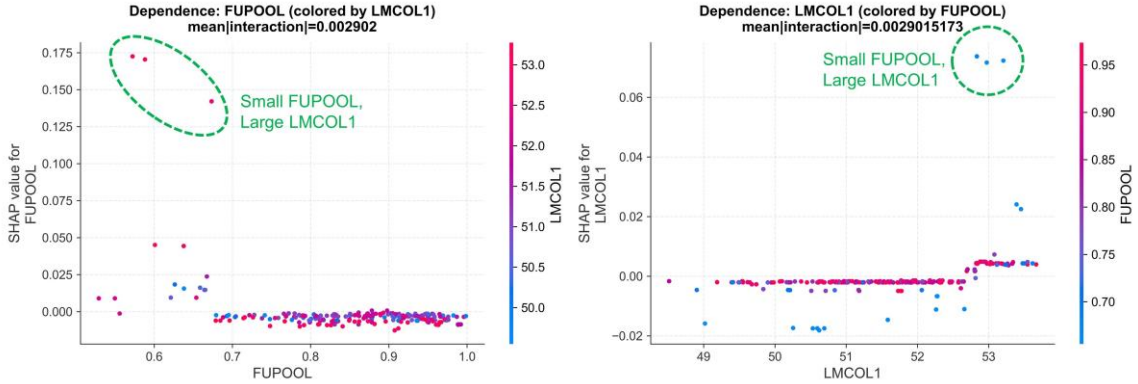

Fig. 5. Interaction effect between FUPOOL and LMCOL1 (Ranked first)

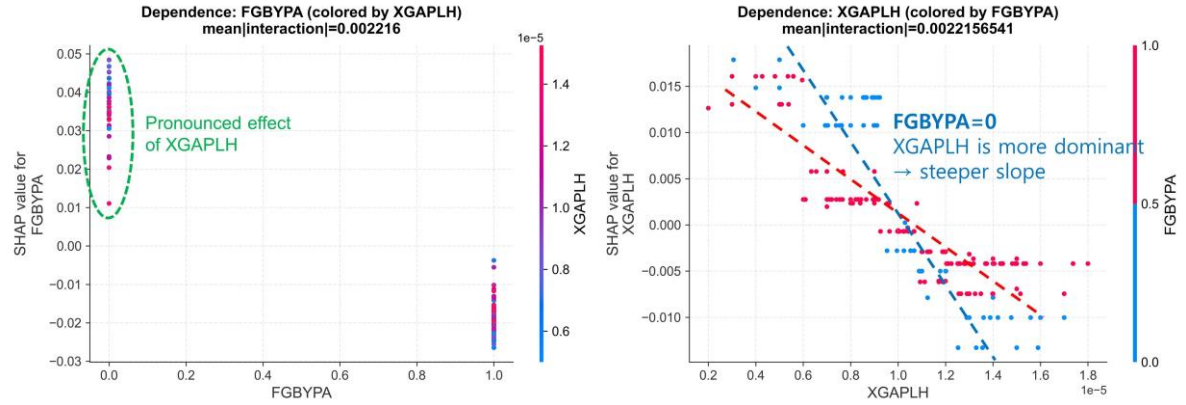

Fig. 6. Interaction effect between FGBYPA and XGAPLH (Ranked second)

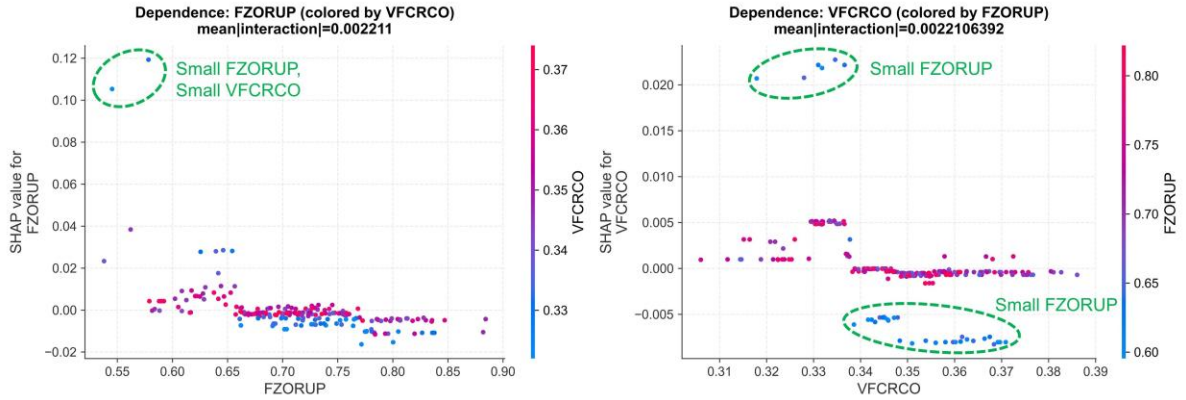

Fig. 7. Interaction effect between FZORUP and VFCRCO (Ranked third)

5. Conclusions

This study presented an explainable machine learning framework to analyze the impact of uncertain parameters on RPV failure time under an OPR1000 MBLOCA scenario. A Dirichlet-based data augmentation method was applied to expand the training dataset while preserving the physical bounds and statistical characteristics of the original uncertain parameters.

Although the augmented data did not improve regression accuracy, this result highlights that the relationship between uncertain parameters and RPV failure time is governed by strong nonlinearities and interaction effects that cannot be fully captured by linear interpolation.

The SHAP-based analysis quantified both the main effects and interaction effects of uncertain parameters, providing insight beyond conventional correlation-based sensitivity analysis. In particular, SHAP enabled identification of interaction-driven contributions that are not detectable using traditional statistical screening methods. This confirms that XAI is a novel approach for understanding complex input-output relationships.

The proposed framework, which combines data augmentation with XAI, provides a practical solution to overcome the inherent sample-size limitations of severe accident simulations. This approach enhances the interpretability and reliability of uncertainty analysis and can serve as a scalable methodology for future severe accident studies involving high-dimensional uncertain parameter spaces.

Future work will focus on developing advanced augmentation methods that better reflect nonlinear physical relationships and extending the proposed framework to additional accident scenarios and figures of merit.

ACKNOWLEDGMENT

This work was supported by the K-Cloud project of KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2024-Tech-08).

REFERENCES

[1] T.-W. Kim, S. Shin, M. Seo, “Uncertainty Study of RPV Failure and Operator Actions in an MBLOCA Scenario of the OPR1000,” Transactions of the Korean Nuclear Society Spring Meeting, May 2025.

[2] K.-I. Ahn, S.-Y. Park, “Best-practice severe accident uncertainty and sensitivity analysis for a short-term SBO sequence of a reference PWR using MAAP5”, Annals of Nuclear Energy, 2022

[3] S. Brumm et al., “Uncertainty quantification for severe-accident reactor modelling: Results and conclusions of the MUSA reactor applications work package”, Annals of Nuclear Energy, 2025.
[4] S. S. Wilks, “Determination of Sample Sizes for Setting Tolerance Limits,” The Annals of Mathematical Statistics, 1941.
[5] A. Saltelli et al., Global Sensitivity Analysis: The Primer, Wiley, 2008.
[6] S. M. Lundberg, S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” NeurIPS, 2017.