Feasibility Study of Development in Machine Learning Prediction Models For Cavity Parameters in OPR1000

Yujin Kim [a], Joon Young Bae [a], Cheolwoong Kim [a], JinHo Song [a], Joon Eon Yang [a], Taewoo Kim[b], Mi Ro Seo[b], Yoonhee Lee [c] and Sung Joong Kim [a,d*]
[a] *Department of Nuclear Engineering, Hanyang University,*
222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea
[b] *Korea Hydro and Nuclear Power Central Research Institute,*
70, Yuseong-daero 1312beongil, Yuseong -gu, Daejeon, Republic of Korea
[c] *Department of Quantum System Engineering, Jeonbuk National University,*
567, Baekje-daero, Deokjin-gu, Jeonju-si, Jeollabuk-do 54896, Republic of Korea
[d] *Institute of Nano Science and Technology, Hanyang University,*
222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea
[*] *Corresponding author: sungjkim@hanyang.ac.kr*

Keywords : Severe Accident, Machine Learning, Hydrogen Concentration, Ablation Depth, Gradient Boosting

1. Introduction

The safe operation of a nuclear power plant depends on the main control room (MCR), where operators continuously monitor and manage the power generation system. The MCR oversees all processes, including plant startup, steady-state operation, power output adjustments, and shutdown. In emergency situations, it functions as the central command center for accident response.

To diagnose and implement effective emergency measures, operators continuously monitor key information from the MCR, including reactor cooling conditions, containment integrity, and the status of emergency safety systems. However, in certain critical situations, MCR data may become unavailable, and the reliability of control element measurements cannot always be guaranteed. Therefore, the ability to accurately predict essential parameters—even under worst-case conditions—would be invaluable for mitigating the severity of an accident [1].

This study focuses on predicting cavity parameters that are closely related to containment integrity. The cavity is a main structure that serves as the final barrier to mitigate severe accidents and prevent radioactive leaks. However, predicting cavity parameters during an accident remains extremely challenging. To address this issue, the study investigates high-performance machine learning models to identify critical cavity parameters relevant to severe accident management and to determine optimized models capable of accurately predicting these parameters. The predictions are based on observable thermal-hydraulic data available from the MCR. By conducting a comparative performance analysis of various models, the study aims to recommend the most effective model for practical application.

2. Methods

2.1 Data Construction

The datasets for training predictive models were generated using MAAP (Modular Accident Analysis Program) version 5.0.6, a severe accident analysis code for nuclear power plants. The selected scenario for the study is the ELAP+LOUHS (Extended Loss of Alternate Power + Loss of Ultimate Heat Sink) accident scenario, which was postulated based on the stress test report for the OPR1000 reactor model of Hanul Units 3 & 4, a standard Korean nuclear power plant.[2] The ELAP+LOUHS accident assumes reactor vessel failure. Following reactor vessel failure, various phenomena occur in the cavity, such as Molten Core-Concrete Interaction (MCCI). This situation limits the actuation of Engineered Safety Features.

With limited mitigation strategies available, three parameters are defined within specific ranges to implement various case scenarios. Due to the loss of alternating current power, the battery powering the turbine-driven auxiliary feedwater pump (TDAFW) is the only available power source. The battery capacity varies, with depletion times ranging from 4 to 11 hours. External mobile equipment is assumed to be available for actuating the containment spray system (CSS), which depends on the inventory level of the Refueling Water Storage Tank (RWST). Two variables are set in the study: the remaining RWST percentage, ranging from 10% to 100%, and the CSS actuation timing, starting from 2 hours up to 30 hours after the initiation of severe accident management guidelines (SAMG).

By appropriately combining the three parameters, a total of 2,320 detailed accident scenarios were generated, each exhibiting time-series characteristics. Using these generated scenarios, data preprocessing was conducted. The input variables consist of thermal-hydraulic parameters observable from the main control room. Hydrogen concentration and ablation depth were selected as target variables corresponding to cavity parameters. These two indicators are considered key factors in determining the external release scenario of radioactive materials during a severe accident. Specifically, hydrogen concentration represents the

amount of hydrogen produced by MCCI and serves as a critical indicator for assessing explosion risk and the potential loss of containment building integrity. Ablation depth was chosen because it directly reflects the potential for physical failure of the final barrier, which could lead to molten material penetrating the concrete floor. **Table 1** lists the variables used to train the predictive model [3].

Table 1: Model Variables List

Input Variable		Thermal Hydraulics
	1	SIT Pressure
	2	RWST Level
	3	Hot Leg Temperature
	4	Cold Leg Temperature
	5	RCS Pressure
	6	Cavity Pressure
	7	RPV level
	8	CET
	9	SG1 Pressure
	10	SG2 Pressure
	11	SG1 DC Level
	12	SG2 DC Level
Target Variable	1	Hydrogen Mole Fraction in Cavity
	2	Concrete Floor Ablation Depth

2.2 Feature Engineering for Enhancing Time-Series Characteristics

In the 2,320 accident scenarios composed of time series data, the target variables, hydrogen concentration and ablation depth, remain at zero during the early stages of the accident. Hydrogen concentration begins to increase after core exposure due to zirconium oxidation. Subsequently, hydrogen is continuously generated from the MCCI reaction following reactor vessel failure. In contrast, ablation depth increases only after the reactor vessel failure. The ablation process occurs later in the accident and progresses very gradually; therefore, changes in ablation depth data are not readily noticeable.

Imbalanced data and delayed changes in values following an accident make training a machine learning model challenging. To address these issues and enable the model to effectively learn temporal patterns in time-series data, feature engineering was incorporated into the framework. Feature engineering involves creating new features to enhance the performance of machine learning models. The selected techniques included lag features, rolling statistics, and difference features. These methods capture the passage of time and the magnitude of change, aiding in the prediction of target variables that fluctuate rapidly after a specific event.

For feature engineering, six variables highly correlated with the target variable were selected. These new features provide the model with a comprehensive understanding of accident progression. Specifically, lag features were generated to serve as the model's memory, utilizing values from eight different past time points (ranging from 1 to 1,440 steps prior). Lag features enable the model to capture the historical state leading up to the current moment. Secondly, rolling statistics were computed by calculating the moving average and standard deviation over seven different window sizes. The rolling statistics approach provides context by capturing recent trends and stability. Finally, difference features were created to detect the rate of change by computing the difference from 10 time steps prior.

2.3 Classification-Regression Hybrid Modeling

The predictive model aims to simultaneously forecast hydrogen concentration and concrete ablation depth within the reactor cavity, using observable thermal-hydraulic variables from the Main Control Room (MCR) as inputs. For model training, a total of 2,320 severe accident scenarios generated with the MAAP code were utilized. Each scenario consists of time-series data recorded at 1-minute intervals over a 72-hour period. The entire dataset was randomly sampled and divided into training, validation, and testing subsets. The number of scenarios and data points assigned to each subset is shown in **Table 2** below.

Table 2: Overview of the Dataset

Type	Scenario	Time Steps per Scenario	Total Data Points
Train	1624	4321	7,017,304
Valid	464	4321	2,004,944
Test	232	4321	1,002,472
Total	2320	4321	10,024,720

Due to the characteristics of the target variable, the dataset contained a large number of zero values. A large number of zero values can cause problems in machine learning. Instead of learning from the relatively few important events (non-zero values), the model becomes biased toward predicting the majority class of normal states (zero values). As a result, the predictive model misses important accident events. To address the zero-value problem of the data, a hybrid approach combining classification and regression models was applied.

Hybrid modeling operates in two distinct stages. First, a classification model is trained to predict whether the target variable at a specific time step is zero or non-zero. Subsequently, a regression model is applied only to the data points predicted as non-zero by the classifier to estimate the actual value. The overall workflow of the hybrid modeling approach is illustrated in **Fig. 1**.

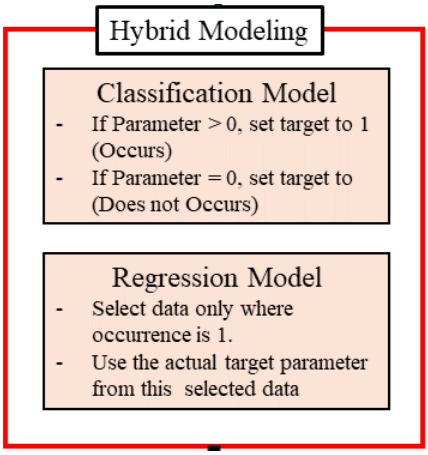

Fig. 1. Hybrid Modeling Process

2.4 Predictive Model Selection

For the accurate prediction, the Gradient Boosting Algorithm (GBM) was adopted as the base model, known for its ability to learn complex patterns and its high predictive performance. **Fig. 2**. shows the decision tree of Gradient Boosting Algorithm.

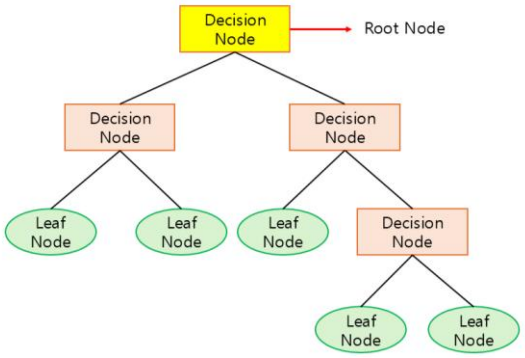

Fig. 2. Decision Tree Algorithms

GBM is a method that sequentially combines multiple simple models to create a powerful predictive model. Initially, a single decision tree addresses the entire problem, making predictions from the root node to the leaf nodes, as shown in **Fig. 2.** During this process, the errors between the predictions and the actual outcomes are calculated. Subsequently, a second tree is trained to learn the residual errors left by the first tree. The error values predicted by the second tree are then added to the predictions of the first tree to improve their accuracy. The decision tree cycle of prediction, error calculation, error learning, and prediction correction is repeated to generate predictions that progressively approach the true values.

Among gradient boosting algorithms, XGBoost, LightGBM, and CatBoost were selected. These three models are widely used GBM techniques within the machine learning community. They were chosen because of their high stability and reliability, which are suitable for enhancing the credibility of predictions. The characteristics of each machine learning model are summarized in **Table 3** [4].

Table 3: Comparison of XGBoost, LightGBM, and CatBoost

Feature	XGBoost	LightGBM	CatBoost
Tree-Growth	Level-wise	Leaf-wise	Symmetric
Categorical Features	Manual Handling Required	Automatic Handling	Optimized Automatic Handling
Advantage	Stability, Regularization	Speed, Efficiency	Accuracy, Ease of Use
Training Speed	Fast	Fastest	Relatively Slower
Memory Usage	Relatively High	Lowest	Relatively High

2.5 Model Performance Metrics

It is essential to use quantitative performance metrics to objectively evaluate and compare a model's predictive performance. Since the model developed in this study employs a hybrid approach—combining classification (to distinguish between 'zero' and 'non-zero' values) and regression (to predict the actual values)—both types of metrics were used to evaluate each component separately.

The classification model is evaluated based on its accuracy in detecting the occurrence of a severe accident event (indicated by a non-zero value).

- Accuracy: The proportion of total predictions that are correct. A higher value indicates better performance; however, accuracy can be misleading in imbalanced datasets where zero values dominate.
- Precision: The proportion of predicted 'events' that were actually correct. A higher value indicates more reliable predictions.
- Recall: The proportion of actual events that the model correctly identifies. A higher value indicates that the model is better at detecting important events without missing them.
- F1-Score: The harmonic mean of Precision and Recall. It is a key metric for comprehensively evaluating a model's performance on imbalanced data, with a higher score indicating better performance.

The regression model is evaluated based on how accurately it predicts the actual magnitudes of non-zero values.

- MAE (Mean Absolute Error): The average of the absolute differences between predicted and actual values. It provides an intuitive measure of error magnitude, with lower values indicating better performance.
- RMSE (Root Mean Square Error): The square root of the average of the squared errors. Unlike MAE, it penalizes larger errors more heavily, making it useful for assessing performance when outliers are significant. A lower value indicates better performance.
- R^2 (Coefficient of Determination): Indicates the proportion of variance in the data that the model can explain. A value closer to 1 signifies greater explanatory power.
- PICP (Prediction Interval Coverage Probability): The proportion of actual data points that fall within a predicted interval (e.g., the 90% prediction interval). For an interval set at 90%, a resulting PICP value close to 0.9 indicates that the model's uncertainty estimation is well-calibrated and trustworthy. Evaluating PICP is therefore crucial for validating the model's ability to provide reliable worst-case guarantees.

3. Results

3.1 Hydrogen Concentration Prediction Results

The performance of the XGBoost, LightGBM, and CatBoost models was measured. **Table 4** displays the classification results for hydrogen concentration, and **Table 5** displays the regression results.

Table 4: Hydrogen Concentration Classification Model Performance

Algorithm	Type	Accuracy	Precision	Recall	F1-Score
XG Boost	Test	0.9258	1	0.9005	0.9476
	Valid	0.9285	1	0.9034	0.9492
	Train	0.9266	1	0.901	0.9479
Light GBM	Test	0.9752	1	0.9668	0.9831
	Valid	0.9768	1	0.09687	0.9841
	Train	0.9758	1	0.9673	0.9834
Cat Boost	Test	0.9401	1	0.9196	0.9581
	Valid	0.9434	1	0.9236	0.9603
	Train	0.9434	1	0.9237	0.9604

Table 5: Hydrogen Concentration Regression Model Performance

Algorithm	Type	MAE	RMSE	R^2	PICP
XG Boost	Test	0.0005	0.0014	0.9835	0.9311
	Valid	0.0005	0.0015	0.9799	0.9151
	Train	0.0004	0.0009	0.9941	0.9128
Light GBM	Test	0.0009	0.0029	0.9302	0.9388
	Valid	0.0008	0.0026	0.9406	0.9294
	Train	0.0008	0.0027	0.9392	0.9224
Cat Boost	Test	0.001	0.0024	0.9528	0.9261
	Valid	0.001	0.0023	0.9545	0.9124
	Train	0.0009	0.0021	0.9636	0.9078

An analysis of the predictive performance on the test dataset revealed that different models exhibited distinct strengths in event detection and value prediction. In terms of classification performance, LightGBM achieved the best results with an F1-score of 0.9831, indicating it was the most effective at detecting the onset of hydrogen generation. Notably, LightGBM and CatBoost attained a perfect Precision score of 1.0 on the test set ensuring high reliability in all hydrogen generation predictions. For regression performance, XGBoost proved to be the best, demonstrating the highest predictive accuracy with an MAE of 0.0005, RMSE of 0.0014, and an R^2 of 0.9835. Its leading performance in the RMSE metric, which is sensitive to large errors, suggests it provides the most stable predictions.

However, for safety-critical applications like severe accident management, average performance is insufficient; a model must also reliably quantify its predictive uncertainty. To evaluate the capability, the Prediction Interval Coverage Probability (PICP) was assessed. All models achieved excellent PICP scores of over 92% for a 90% confidence interval on the test set. PICP result confirms that the models not only predict accurately on average but also provide a well-calibrated and trustworthy range of uncertainty.

Fig. 3 and 4 show the time-series prediction results for hydrogen concentration. **Fig. 3.** presents the most accurate scenario (average RMSE = 0.00354), whereas **Fig. 4.** displays the least accurate one (average RMSE = 0.008331). A comparison was made between the ground truth data and the time-series forecasts generated by different machine learning models for these two scenarios.

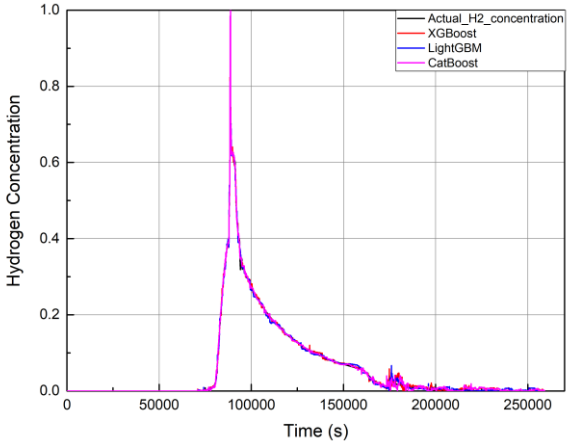

Fig. 3. Hydrogen Concentration: Best-Predicted Scenario

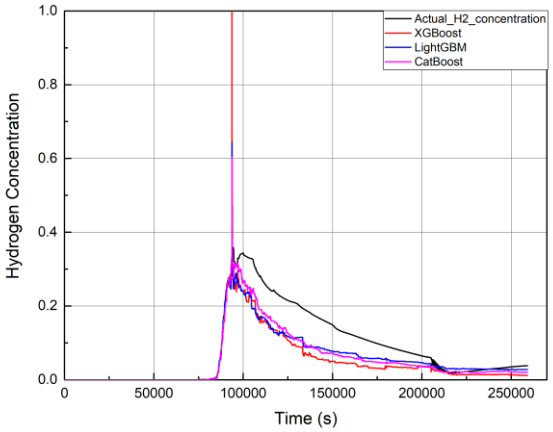

Fig. 4. Hydrogen Concentration: Worst-Predicted Scenario

To visually complement the quantitative results, a graphical analysis of the time-series predictions was conducted for two distinct scenarios. **Fig. 3.** illustrates the most accurate prediction scenario, in which all models closely tracked the actual hydrogen concentration profile. They accurately captured the timing and magnitude of the peak concentration, as well as the subsequent decay trend, demonstrating high reliability across all models in this case. In contrast, **Fig.4.** represents the least accurate scenario, highlighting a crucial distinction between visual interpretation and quantitative metrics.

A visual inspection of the peak reveals clear differences in the models' overestimation errors. XGBoost and CatBoost exhibit the largest errors, each predicting a peak of 1.0, which is nearly three times the actual value of approximately 0.35. In comparison, LightGBM's peak prediction of around 0.7 is noticeably closer to the true value. Although still inaccurate, its prediction error is roughly half that of the other two models, indicating a less extreme response to the surge. However, the quantitative RMSE results for the specific scenario tell a different story. The actual RMSE scores were lowest for CatBoost (0.0075), followed by

LightGBM (0.0080) and XGBoost (0.0096). The reason for the discrepancy is that although CatBoost had a significant error at the peak, it consistently maintained lower errors than the other models during the prolonged subsequent decay phase, resulting in the smallest overall cumulative error across the entire duration.

Synthesizing the results led to the optimal combination for the proposed hybrid model. For the initial classification stage, LightGBM is the preferred choice due to its superior F1-Score. For the subsequent regression stage, XGBoost was selected as the optimal model because of its superior generalization performance. Although analyzing outlier cases, such as the one illustrated in **Fig.4.** —where CatBoost exhibited lower error—is important for understanding model limitations, the ultimate goal is to choose a model that performs reliably across the broadest range of unseen scenarios. XGBoost's significantly lower average RMSE (0.0014) across the entire test dataset demonstrates its superior ability to generalize. This indicates that XGBoost has most effectively learned the underlying patterns of hydrogen behavior, making it the model most likely to provide dependable predictions for future, unforeseen events. The superior generalization performance of XGBoost is attributed to its capacity to capture complex nonlinear patterns effectively and its inherent regularization mechanisms that prevent overfitting.

Therefore, the optimal strategy is to construct a hybrid model that employs LightGBM for rapid event detection, followed by XGBoost for the most robust severity prediction. The hybrid modeling approach provides operators with the most timely and accurate information on average. The model can be used to preemptively assess the risk of hydrogen explosions and to help determine effective accident mitigation measures.

3.2 Ablation Depth Prediction Results

The same three models were also used to evaluate predictive performance for ablation depth. **Table 6** shows the classification results, while **Table 7** presents the regression results.

Table 6: Ablation Depth Classification Model Performance

Algorithm	Type	Accuracy	Precision	Recall	F1-Score
XG Boost	Test	0.9467	1	0.9212	0.959
	Valid	0.9279	0.9955	0.8961	0.9432
	Train	0.9426	1	0.9146	0.9554
Light GBM	Test	0.996	1	0.994	0.997
	Valid	0.9925	0.9955	0.9932	0.9944
	Train	0.9961	1	0.9942	0.9971
Cat Boost	Test	0.9667	1	0.9508	0.9748
	Valid	0.9562	0.9952	0.939	0.9663
	Train	0.9605	1	0.9412	0.9697

Table 7: Ablation Depth Regression Model Performance

Algorithm	Type	MAE	RMSE	R^2	PICP
XG Boost	Test	0.0061	0.0329	0.9083	0.9646
	Valid	0.0073	0.0456	0.7473	0.9448
	Train	0.0022	0.0064	0.995	0.9485
Light GBM	Test	0.0158	0.0881	0.3415	0.9642
	Valid	0.0116	0.0713	0.3827	0.9484
	Train	0.0096	0.0552	0.6279	0.9477
Cat Boost	Test	0.0103	0.0459	0.8216	0.9121
	Valid	0.0111	0.0519	0.6736	0.8901
	Train	0.0073	0.0219	0.9414	0.6293

An analysis of the predictive performance for ablation depth revealed that, while all models demonstrated strong detection capabilities, there were notable differences in their regression accuracy. Regarding classification performance, LightGBM was the most effective at detecting the onset of ablation, achieving an almost perfect F1-score of 0.997. XGBoost and CatBoost also exhibited excellent detection abilities, with F1 scores exceeding 0.95.

For regression performance, XGBoost was the clear leader across all metrics, achieving a MAE of 0.0061, RMSE of 0.0329, and an R^2 of 0.9083. These results indicate that XGBoost most accurately predicted the actual changes in ablation depth, explaining nearly 91% of the variance in the data. In contrast, LightGBM exhibited a significantly lower R^2 of 0.3415. Although LightGBM excels at event detection, it is limited in its ability to predict the actual depth. In conclusion, LightGBM was identified as the most suitable model for detecting the onset of ablation, while XGBoost was the superior model for precisely predicting the depth after the event began.

Furthermore, XGBoost demonstrated superior performance in uncertainty quantification. It achieved a PICP of 96.46% for a 90% confidence interval on the test set. Such a high PICP score demonstrates that its prediction intervals are well-calibrated and reliably conservative. This score was significantly higher than that of CatBoost (91.21%) and LightGBM (96.42%, though with much lower R^2). Therefore, XGBoost is not only the most accurate model on average but also the most trustworthy in providing a safe range of potential outcomes for ablation depth.

Fig. 5 and 6 show the time-series prediction results for ablation depth. **Fig.5.** presents the most accurate scenario (average RMSE = 0.000491), whereas **Fig.6.** displays the least accurate one (average RMSE = 0.461827). Similar to the hydrogen concentration predictions, for these two scenarios, we compared the forecasts from each machine learning model against the actual data.

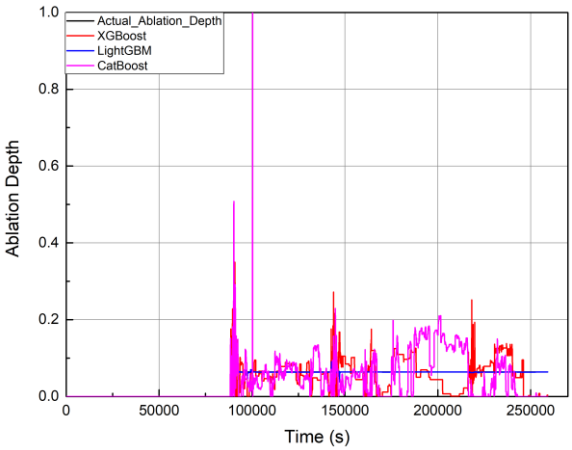

Fig. 5. Ablation Depth: Best-Predicted Scenario

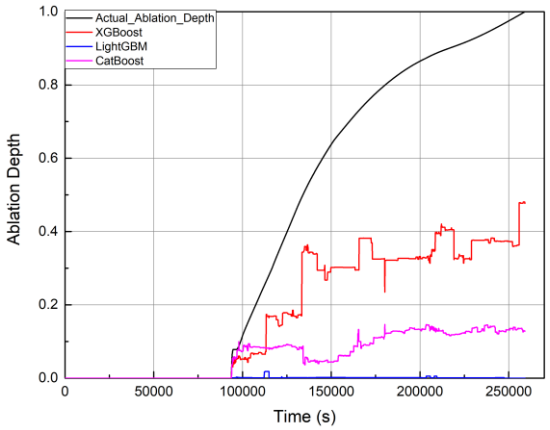

Fig. 6. Ablation Depth: Worst-Predicted Scenario

A visual analysis of the time-series predictions highlights these characteristics. In **Fig. 5**, where ablation was minimal, all models accurately predicted near-zero depth. However, in **Fig. 6**, the limitations of all models became apparent. In this case, all models severely underestimated the actual ablation depth. Despite this underestimation, XGBoost (red line) performed best among the models, as it was the only one to capture the initial increasing trend of the ablation. CatBoost also showed a slight response, while LightGBM almost completely failed to predict the progression. This visual assessment is directly corroborated by the quantitative data for the specific challenging scenario, where XGBoost had a significantly lower RMSE (0.3320) compared to CatBoost (0.4881) and LightGBM (0.5654). Therefore, XGBoost is not only the best model on average but also the most reliable in the worst-case scenarios.

The analysis concludes that LightGBM is optimal for the initial detection of an event, while XGBoost is the most stable and accurate model for predicting the subsequent ablation depth. LightGBM's inferior performance in the regression stage may be attributed to the slow, gradual nature of the ablation process, which can be challenging for its leaf-wise tree growth method. In contrast, XGBoost's ability to learn complex, long-term patterns makes it superior for this task.

Therefore, the optimal strategy is to construct a two-part hybrid model that uses LightGBM for rapid event detection, followed by XGBoost for precise prediction of ablation depth. Such a model provides operators with the most reliable information for assessing the risk of final barrier failure and evaluating the effectiveness of mitigation strategies.

3.3 Uncertainty and Interpretability Analysis

The current model's reliability was assessed using average error metrics. However, to trust a model's predictions in a real severe accident scenario, it is necessary to quantify predictive uncertainty and ensure the model's decision-making process is comprehensible.

Predictive uncertainty can be demonstrated by providing prediction intervals. Prediction intervals offer a guarantee that the predictions will not fall outside of a safe range, even in the most dangerous situations. In **Fig. 7 and 8**, we present 90% confidence prediction intervals for hydrogen concentration and ablation depth using the Quantile Loss method.

Subsequently, SHapley Additive exPlanations (SHAP) analysis is used to identify which thermal-hydraulic signals have the greatest influence on the predictions. By analyzing the impact of each input feature, we can improve the model's interpretability and reliability. In **Fig. 9 and 10**, the SHAP analysis results confirm the variables primarily used for each machine learning model's prediction.

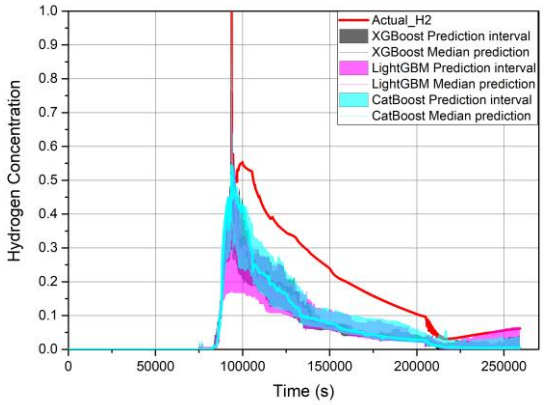

Fig. 7. Uncertainty Analysis of Hydrogen Concentration

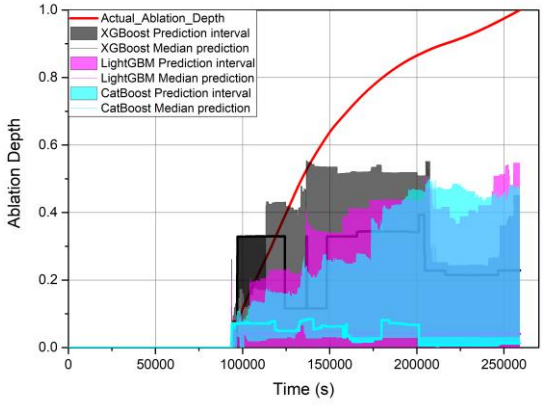

Fig. 8. Uncertainty Analysis of Ablation Depth

To verify the reliability of our model's predictions in a worst-case scenario, we conducted an analysis of predictive uncertainty.

Looking at the hydrogen concentration prediction graph, all three models accurately capture the rapid increase in hydrogen concentration at approximately 80,000 seconds. They also follow the overall downward trend after the peak. Notably, the XGBoost model's predictions align more closely with the actual values. However, the actual values (the red line) consistently lie above the 90% prediction intervals of all three models after the peak. This indicates that the models are persistently underestimating the hydrogen concentration and fail to adequately capture the full extent of the

uncertainty. While the models are valid in predicting the overall trend, they fall short of providing a safety margin that includes the worst-case values.

The ablation depth follows a steady increasing trend over time. All three models accurately capture the onset of erosion. While all models predict an upward trend in ablation depth, their accuracy in predicting the magnitude of the depth varies significantly. The XGBoost model's curve more closely follows the actual ablation depth. Its prediction interval largely encompasses the actual values, suggesting it reliably captures the uncertainty for this parameter. In contrast, the LightGBM and CatBoost models deviate significantly, consistently underestimating the actual values. Their prediction intervals rarely include the actual values, indicating low reliability for this specific task.

In conclusion, The machine learning models have proven effective at predicting the onset of key physical phenomena in a severe accident. But they show limitations in consistently providing reliable severity predictions and quantifying uncertainty across all scenarios. Machine learning prediction models research's significance lies in its ability to accurately identify the timing of an event, which can be crucial for an effective and timely response to mitigate the accident.

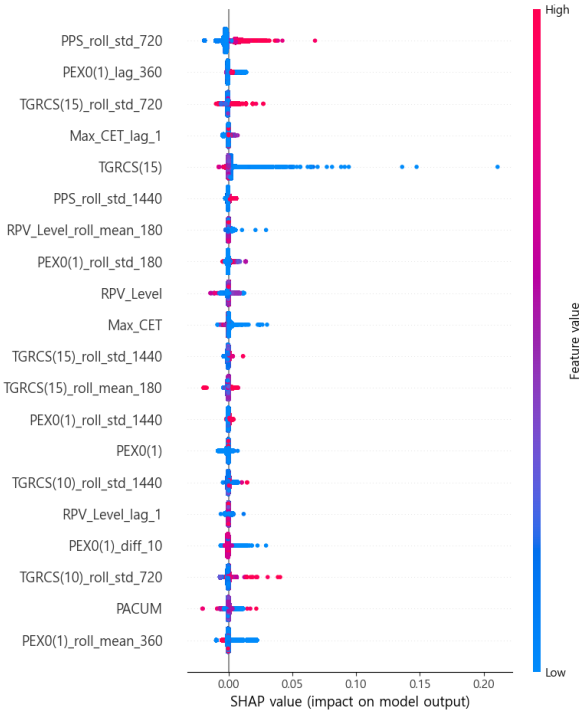

Fig. 9. Hydrogen Concentration SHAP Values

The hydrogen concentration SHAP analysis revealed that the model correctly identified variables related to system instability and loss of cooling as the most important predictors for hydrogen production. This indicates that the model learned physically meaningful relationships.

For instance, the model accurately associated high RCS pressure variability (PPS_roll_std_720) with an increased likelihood of hydrogen production. High variability in the RCS over a long period (indicated by red dots) suggests system control instability and deepening core damage as indicated by its positive SHAP values.

It also learned that a low hot tube temperature (TGRCS(15)) serves as a precursor to core overheating and subsequent hydrogen production by indicating a lack of core cooling. The hot tube temperature was low (indicated by blue dots), the model tended to overestimate the hydrogen concentration.

Furthermore, the analysis confirmed the importance of feature engineering that incorporates temporal context. The most influential variables were those representing trends, variability, and past states from historical data, such as roll_std, roll_mean, and _lag. Feature Engineering method demonstrates that extracting this kind of information is essential for predicting complex, time-dependent severe accidents.

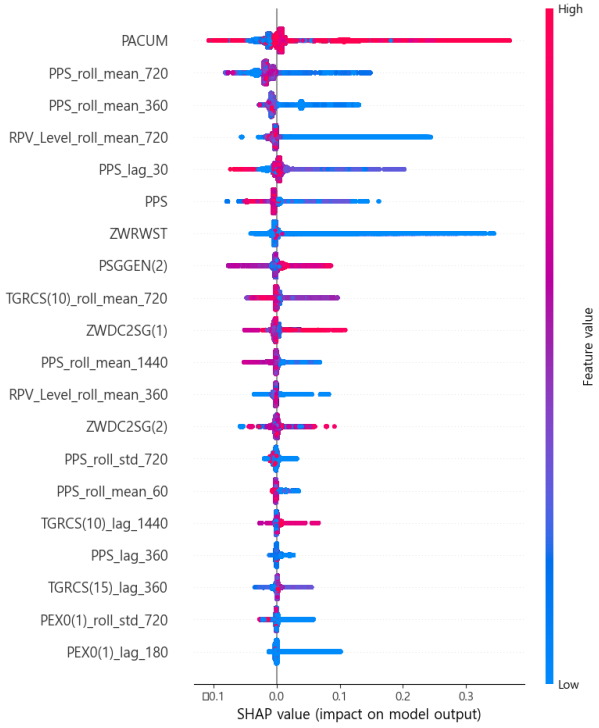

Fig. 10. Ablation Depth SHAP Values

The ablation depth SHAP analysis revealed that the ablation depth prediction model reflects distinct physical characteristics compared to the hydrogen concentration model. The model relies on long-term instability and the occurrence of specific events rather than short-term value changes as its key predictors.

SIT Pressure (PACUM) is the most influential variable in the ablation prediction model. The wide distribution of SHAP values on the graph indicates that even small changes in this variable can significantly impact the predicted ablation depth. The model learned that high SIT pressure (represented by red dots) is a critical signal for an increase in ablation risk. Conversely, low pressure (represented by blue dots) is the most

powerful factor in mitigating the risk. This demonstrates that the model correctly identified SIT Pressure as a crucial prerequisite for the occurrence of ablation.

Overall, the model does not rely on single variables but comprehensively considers multiple variables from key systems like the Reactor Coolant System (PPS) and Reactor Pressure Vessel (RPV). This demonstrates that the model monitors overall system stability to make its predictions.

In conclusion, the SHAP analysis confirmed that the model developed in this study learned physically meaningful relationships from the data and made reasonable inferences based on system stability and accident progression stages, which are considered important by actual operators.

4. Conclusions

The goal of this feasibility study was to predict critical cavity parameters, such as hydrogen concentration and concrete ablation depth, which are difficult to measure directly. We successfully developed a high-performance machine learning model that utilizes observable thermal-hydraulic data from the Main Control Room (MCR).

The analysis revealed that the optimal model choice depends on the specific task: event detection (classification) or severity prediction (regression). For the initial classification task, LightGBM consistently demonstrated the best performance for both target variables, making it the ideal choice for rapid event detection. For the subsequent regression task, the primary criterion for model selection was superior generalization performance—the ability to provide reliable predictions across the widest possible range of scenarios. XGBoost was identified as the model with the best generalization capability for both variables, although it demonstrated capability in different ways.

For hydrogen concentration, superior generalization was evidenced by its significantly lower average RMSE across the entire dataset. A strong overall average for XGBoost indicates that it has most effectively learned the underlying patterns, making it the most dependable choice for future, unforeseen situations. In contrast, for ablation depth, its superiority was unambiguous: it achieved the best average performance metrics and also outperformed competitors in the most challenging scenarios.

The uncertainty analysis revealed that the models tended to underestimate the actual values for hydrogen concentration and ablation depth, indicating they do not always provide a sufficient safety margin for the worst-case scenario. However, The models are effective at predicting the onset of key physical phenomena in a severe accident. The ability of the models to accurately identify the timing of an event is crucial for an effective and timely response to mitigate the accident.

Furthermore, the SHAP analysis provided critical insights into the model's decision-making process. The model is confirmed that it learned physically meaningful relationships from the data. With respect to hydrogen concentration, the model correctly identified variables related to system stability and core overheating as key predictors. In contrast, for ablation depth, the model's reliance on long-term instability and cooling system status variables, accurately reflected the physical precursors of ablation. This enhanced interpretability demonstrates its potential as a reliable tool for human operators by providing plausible explanations for its predictions.

Therefore, the study proposes a robust two-part hybrid model that employs LightGBM for rapid event detection and XGBoost for precise severity prediction as the optimal strategy. The primary significance of the research lies in demonstrating the feasibility of predicting key severe accident indicators using only observable MCR variables. The predictive capability of the developed model can significantly contribute to accident mitigation by supporting operator decision-making in complex emergency situations.

The objective of current work was to validate the effectiveness of predicting cavity parameters using machine learning techniques. The model was developed using simulation data from the MAAP code, on the assumption that the MAAP results were true values with no uncertainty. This, however, inherently limits the model's ability to predict real-world data and introduces epistemic uncertainty, as real-world power plant data contain significant uncertainties not present in the simulation. Consequently, the research focused on demonstrating the applicability of machine learning using simulation data. Overcoming potential discrepancies with real-world power plant data and enhancing the model are left for future research.

Acknowledgement

This work was supported by KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2024-Tech-08). This work was supported by the Innovative Small Modular Reactor Development Agency grant funded by the Korea Government (MSIT) (No. RS-2023-00259516).

REFERENCES

[1] J. H. Song, S. J. Kim, A machine learning informed prediction of severe accident progressions in nuclear power plants, Nuclear Engineering and Technology, Volume 56, Issue 6, 2024.

[2] 한울 3,4 호기 스트레스 테스트 수행보고서 (공개본), 한수원, 2018.

[3] J. Y. Bae et al., “Autoregressive Multivariate Time Series Forecasting of Total Loss of Component Cooling Water Accident Sequences Calculated by Modular Accident Analysis Program, Inspired by Video Prediction Methods Using Deep Learning”, Transactions of the Korean Nuclear Society Spring Meeting, Jeju, Korea, May 8–10, 2024.

[4] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 5831–5866, Mar. 2021