Enhancing Nuclear Power Plant Safety and Reliability: Integrating Explainable AI for Operator Performance Optimization

Merouane Najar*, He Wang
Fundamental Science on Nuclear Safety and Simulation Technology Laboratory, Harbin Engineering University, Harbin, 150001, China

Abstract: Minimizing operator errors during reactor accidents is pivotal for enhancing safety and reliability in nuclear power plants, thereby sustaining the use of clean energy. To address this challenge, the integration of artificial intelligence (AI) with nuclear safety systems is proposed to optimize operator performance, enabling swift and accurate responses to abnormal operational situations. However, the complexity of such models which considered as "black boxes," poses a challenge for operators in comprehending their decision-making processes, leading to a lack of trust. Overcoming this limitation necessitates the application of Explainable AI (XAI) to provide transparency in model results, thus establishing trust in AI. This study focuses on simulating Loss of Coolant Accident (LOCA) scenarios for various pipe break fractions, utilizing supervised machine learning to classify LOCA break types (small, medium, or large). The classification is then integrated with regression models to predict variations in safety margins. Followed by the application of metric parameters to evaluate their performance. Finally, XAI is utilized to explain the model results, facilitating an understanding of the reasons behind their decisions.

Keywords: Nuclear safety, Reliability, Operator performance, XAI, Trustworthiness, Supervised ML, Transparency.

1. INTRODUCTION

Despite advances in the design and safety systems of nuclear power plants (NPPs) aimed at reducing the need for operator intervention during reactor accidents, enhancing operator reliability remains crucial to prevent or mitigate the consequences of accidents. However, human behavior is complex and difficult to predict, particularly during reactor accidents where an accumulation of various factors influences the operator's response. Some factors, such as time pressure, alarm hierarchy, and the need to collect data from different sources, are observable, while others remain hidden. This complexity poses a challenge in ensuring whether the operator will make errors or not. Consequently, there is a pressing need for an alternative approach to reduce the load on the operator and support their decision-making.

The integration of AI with nuclear safety systems addresses this issue by rapidly collecting and analyzing data from several detectors, providing accurate suggestions and recommendations. This aids operators in making the correct actions, thereby decreasing the likelihood of human failure events (HFE). By the beginning of 2020, AI technology had become less complex and more accessible, as reflected by numerous works applying AI in the nuclear safety area. These researchers are divided into three main directions:

Earlier Fault Detection: Applying models such as Long Short-Term Memory (LSTM) [1] or Convolutional Neural Network (CNN)-LSTM [2], which utilize past sequential data to forecast the progress of future accidents, provide the operator with more time to make appropriate decisions to limit further consequences.

Fault Classification: Building models such as Graph Neural Network (GNN) [3], Recurrent Neural Network (RNN) [4] can characterize the type and location of a failure, enabling the operator to react quickly and correctly, especially in scenarios involving multiple failures.

Risk Assessment and Safety Margin Characterization: Training models to predict changes in safety margins [5], providing a clear picture to the operator about the severity of an accident and which resources need to be allocated. From 2020 until 2022, the number of published researches using Machine Learning (ML) in NPPs increased dramatically; however, the focus of these efforts was on training models that can make predictions with high accuracy, which led to increased complexity of the built models. Consequently, the trained models became as "Black boxes," making it difficult to understand the reasoning behind their outputs. This poses a challenge for the real application of AI in NPPs, where operators need to comprehend the inner workings of the model to trust its results. Therefore, there is a necessity for a new generation of models to be "Glass boxes," where the operator can easily pinpoint the impact of each variable and use their experience and knowledge to validate the model's prediction, consequently enhancing trust in the model's results. In 2022 and 2023, the Idaho National Laboratory (INL) published two reports [6, 7]

highlighting the importance of XAI and promoting the use of explainability as a promising method for advancing the integration of AI in NPPs. However, the application of XAI in the nuclear field remains in its early stages, hence, this work aims to enrich this area by providing a model combining performance and interpretability to establish trustworthiness.
In the first part of the study, two methodologies (integration and combination) are used to build a model with high performance in predicting the type of LOCA failure and evaluating the safety margins variation. And the second part intends to explain the reasons behind the model outcomes.

2. Methodology

2.1 ANN model development

The structure of the Artificial Neural Network (ANN) model comprises multiple layers, including input, hidden, and output layers. The hidden layers serve as the central component of the model, housing neurons interconnected through various combinations of connections. The development of this model involves importing the requisite dataset and then splitting it into inputs and outputs, subsequently divided into training, validation, and testing segments, followed by scaling and training the model. This work utilizes two methodologies to predict the type of failure and Peak Cladding Temperature (PCT), which allows for the evaluation of safety margin variations. The both approaches are discussed as follows:

- **ANN Integrated Model:** This methodology employs an integration technique of two distinct models. Firstly, a classification model is used to forecast the type of accident. The results from this model are then combined with inputs for a regression model, which predicts the key parameter, PCT, as shown in Figure 1.
- **ANN Combined Model:** In this technique, the two models receive the same inputs, and the outputs from both models are generated concurrently. This approach helps compress the overall model structure and reduces simulation time, as illustrated in Figure 2.

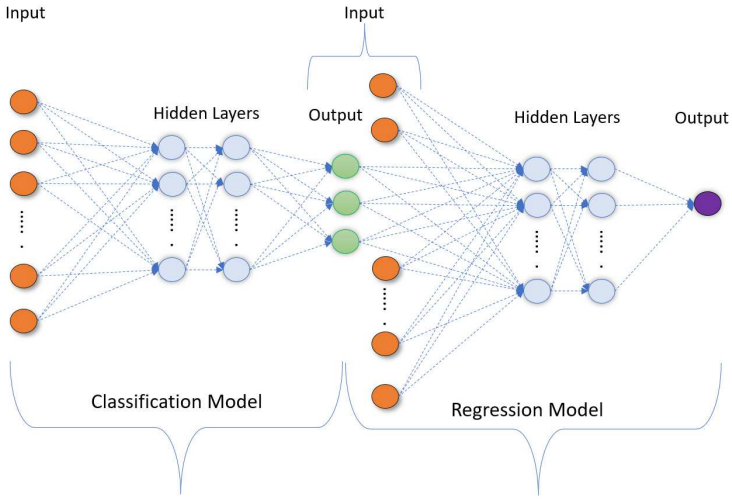

Figure.1 ANN integrated model

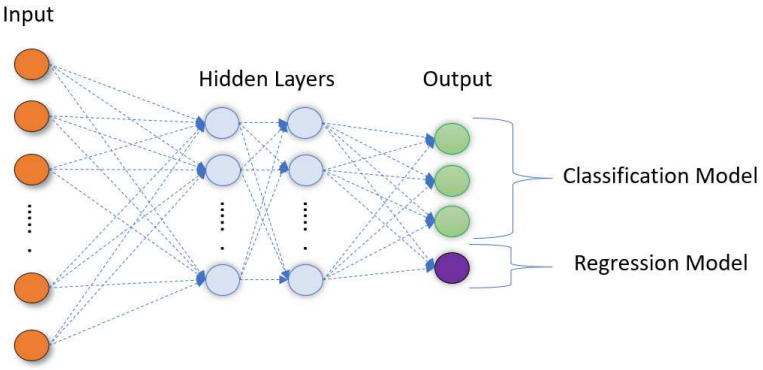

Figure.2 ANN combined model

2.2 ANN model evaluation

After training the model (integrated, combined), the next step is to evaluate their performance using the appropriate metric parameters for both regression and classification algorithms:

- **Regression model evaluation**: The regression model is used to predict the PCT. Table 1 displays the statistical parameters used to evaluate the results of this algorithm. Where, **y** is actual values and Y_{pre} is predicted values.

Table 2. Metric parameters for regression model [8]

Metric parameter	Propriety	Equation	
Coefficient of determination	is a statistical measure of how well the regression predictions approximate the actual data points.	$R^2 = 1 - \frac{\sum(y - y_{pre})^2}{\sum(y - \bar{y})^2}$	(1)
Mean absolute error	A commonly used measure of error for estimation problems.	$MAE = \frac{1}{N}\sum\lvert Y - Y_{pre}\rvert$	(2)
Root mean square error	is considered an excellent general-purpose error metric for numerical predictions.	$RMSE = \frac{1}{N}\sqrt{\sum Y - Y_{pre}}$	(3)
Mean square error	is a simple square of the difference between the measured and observed values.	$MSE = \frac{1}{N}\sum(Y - Y_{pre})^2$	(4)

- **Classification model evaluation:** To assess the classification algorithm, four evaluation matrices are employed, with their appropriateness discussed in Table 2. Where, TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives).

Table 2. Metric parameters for classification model [9, 10]

Metric parameter	Propriety	Equation	
Accuracy	This metric assesses the ratio of correct predictions to incorrect ones made by the model.	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$	(5)
Precision	This parameter reflects how cautiously the model makes positive predictions.	$PREC = \frac{TP}{TP + FP}$	(6)
Recall	Used to assess how well the model generalizes in identifying positive cases.	$REC = \frac{TP}{TP + FN}$	(7)
F1-score	This score indicates the equilibrium between precision and recall in the model's performance.	$F1 = 2 \times \frac{PREC \times REC}{PREC + REC}$	(8)

2.3 ANN model explainability

The evaluation of model performance alone is insufficient to rely on their outcomes, due to hidden reasons and complex relations upon which the model bases its predictions; therefore, it is considered a "Black box." Therefore, the application of XAI through the use of the Shapley Additive Explanations (SHAP) package is employed, offering several capabilities including global and local interpretability. These features enable the revelation of the inner workings of the model and demonstrate how it arrives at its results, thereby transforming it into a "Glass box model." Consequently, this transparency enhances trust in the model's predictions, as illustrated in Figure 3.

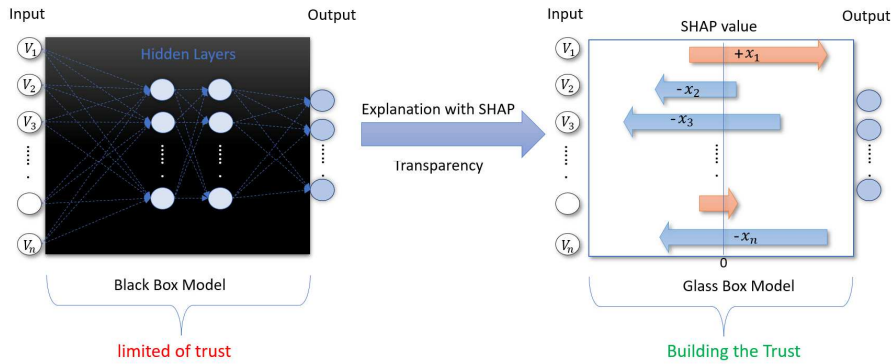

Figure.3 XAI for ANN model

3. Result and discussion

3.1 ANN models comparison results

The Personal Computer Transient Analyzer (PCTRAN) software's open data is used [11] is employed to train a model that simulates a LOCA in a PWR3P, with failure variations ranging from 1% to 100%. Ten parameters are selected as inputs for the model: Reactor Power (QMWT, MW), Steam Generator Pressure (PSGA, Bar), Pressurizer Level (LVPZ), Reactor Coolant System Pressure (P, Bar), Reactor Coolant Loop Flow (WRCA, t/h), Hot Leg Temperature (THA, ℃), Cold Leg Temperature (TCA, ℃), Average RCS Temperature (TAVG, ℃), Steam Generator A Heat Removal Power (QMGA), and RCS Liquid Volume (VOL). The target variables are Peak Clad Temperature (TPCT, ℃) and the type of failure (small, large, medium). The total dataset generated comprises 44,639 instances, with 70% allocated for training, 20% for testing, and 10% for validation.

Two methodologies, combination and integration, are applied, utilizing two supervised machine learning techniques: regression and classification. The corresponding metric parameters for both the classification and regression models are employed, with the results displayed in Tables 1 and 2. Both the combined and integrated ANN models demonstrated excellent $\boldsymbol{R^2}$ score, exceeding 99%, and good accuracy. However, the integrated model showed an accuracy that was 3.15% higher than that of the combined model. Furthermore, the accuracy of the integrated model stabilized quickly, whereas the combined model required 600 epochs to begin stabilizing, as seen in Figure 4. On the other hand, the MAE for both models decreased rapidly, reaching their lowest values within 250 epochs, as illustrated in Figure 4.

Table.1. Regression metric parameters comparison for integrated and combined model

ANN Model	Determination coefficient		MAE		MSE		RMSE	
	Train	Test	Train	Test	Train	Test	Train	Test
Integrated	99.79	99.76	5.37	5.51	182.17	207.87	13.49	14.41
Combined	99.56	99.56	7.91	8.22	308.84	961.28	17.57	19.78

Table.2. Classification metric parameters comparison for integrated and combined model

ANN Model	Accuracy		Precision		Recall		F1	
	Train	Test	Train	Test	Train	Test	Train	Test
Integrated	96.98	96.96	96.51	96.52	96.25	96.27	96.35	96.36
Combined	93.83	93.90	94.00	94.07	93.54	93.57	93.65	93.69

Figure 5 depicts the losses for both models, where the integrated method shows the lowest and earliest stabilization of losses in the classification model compared to the combined technique. Meanwhile, for the regression model, both the integrated and combined methods demonstrate similar behaviour and stabilize approximately simultaneously.

The confusion matrix is used to further compare the two techniques for the classification model, as shown in Figure 6. The results indicate that both methods struggle to distinguish between medium LOCA (class 1) and large LOCA (class 2). Specifically, the integrated method incorrectly predicted class (1) as class (2) 163 times, while the combined method made 206 misclassifications. However, both models (combined and integrated) can easily differentiate between class (2) and small LOCA (class 0).

Figure 7 illustrates the regression predictions of PCT for both the integrated and combined models, with each method demonstrating good fit. However, the combined model shows slightly higher residuals, with some samples significantly deviating from the best fit line.

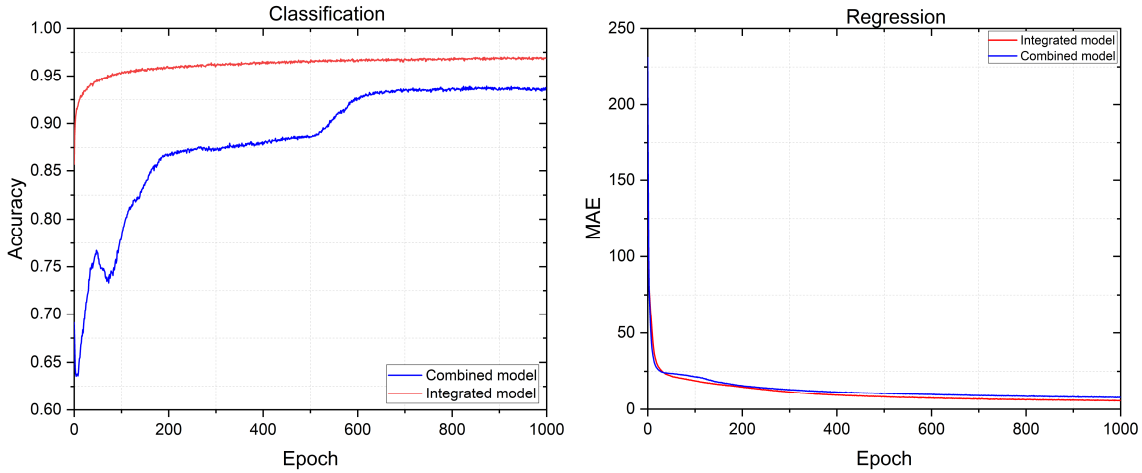

Figure 4. Combined and integrated models' comparison for accuracy and MAE

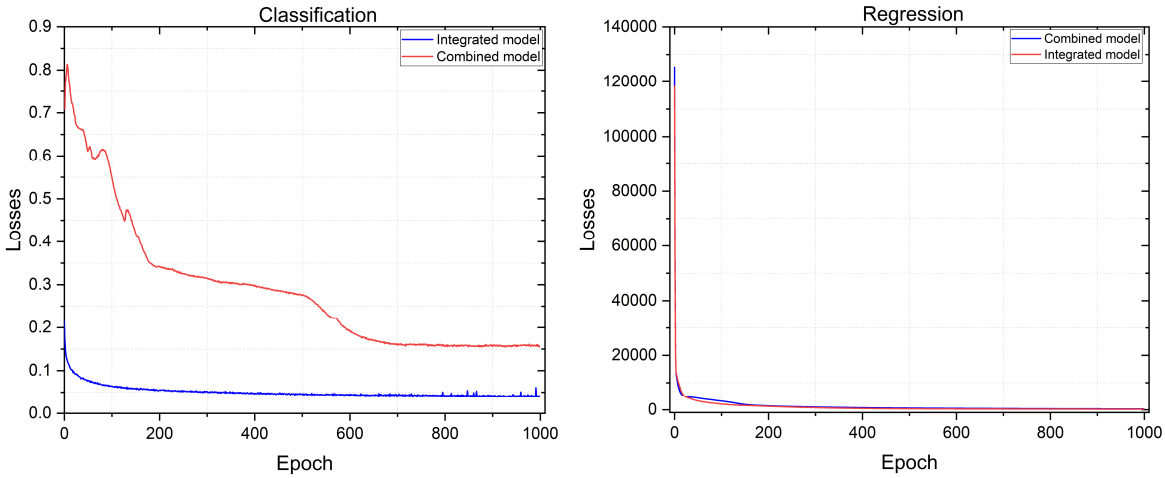

Figure 5. Combined and integrated models' comparison for losses

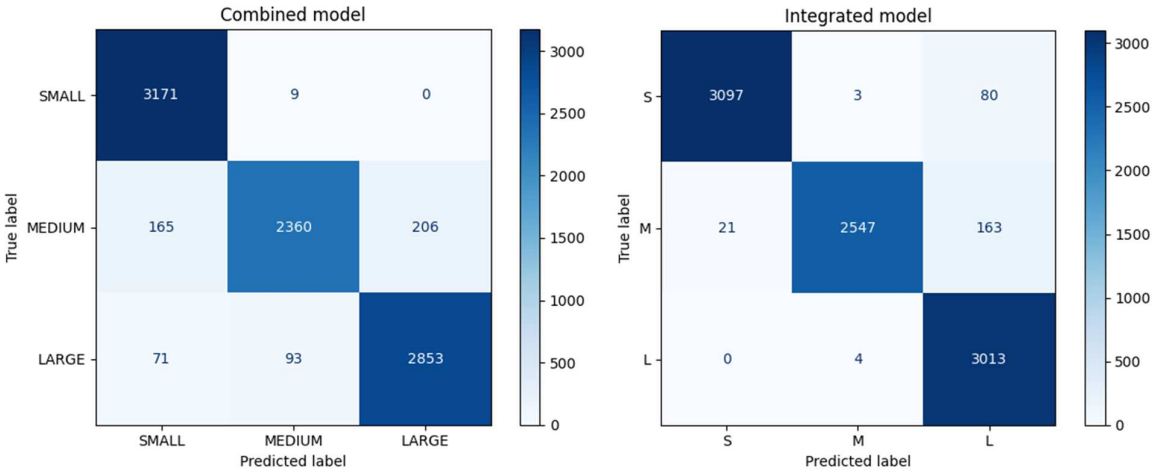

Figure.6 Combined and integrated models' comparison for confusion matrix

Finally, both methods exhibited excellent predictive performance for the regression model, while for classification, the integrated technique showed marginally better accuracy than the combined model. However, the combined model required fewer parameters and consumed less time than the integrated model, thereby demonstrating better overall performance and proving to be the more appropriate choice.

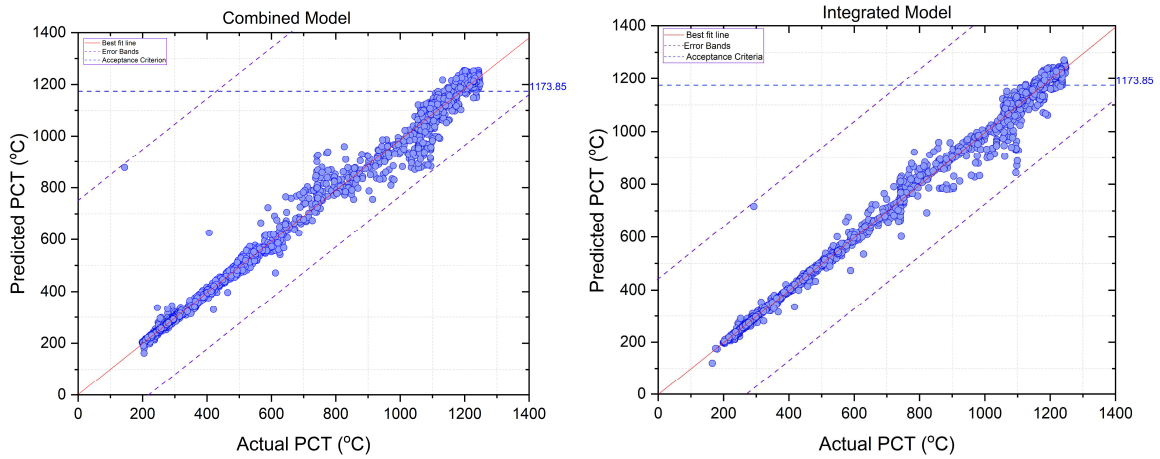

Figure 7. Combined and integrated models' comparison for PCT prediction

3.2 ANN models explainability results

In the previous part the combined technique demonstrated the best performance, however the inner working and the reason behind model prediction remain hidden, leading to lack of the trust. Therefore, in this section the XAI will apply through using SHAP package to reveal their decision underlying and offering the transparency of the model. The Figure 8 illustrated summary plot for the combined model including regression and classification explanations.

The results shows that PSGA is the most significant parameter for both regression and classification models, and this aligning with engineering meaning, where the loos of coolant in the primary loop resulting of the reducing the capability of removing the heat by secondary side leading to increasing significantly the steam generator pressure (PSGA). However, the order of top remainder contributed parameters for each model is different. The most influential variables for regression ANN are TAVA, THA and TCA, consequently, the model relaying on the temperature's values of hot and cold legs and average RCS to predict PCT, and this is reasonable. In the other hand the classification models employ the volume, pressure and average temperature of the RCS to characterize the type of the classes where these variables (VOL, P, TVAG) have closes conurbation values for class prediction.

In the previous part, the combined technique demonstrated the best performance; however, the inner workings and the reasons behind the model predictions remained hidden, leading to a lack of trust. Therefore, in this section, the XAI will be applied using the SHAP package to reveal the underlying decision-making processes and offer transparency in the model's predictions.

Figure 8 illustrates a summary plot for the combined model, including explanations for both regression and classification components. The results show that Steam Generator Pressure (PSGA) is the most significant parameter for both regression and classification models. This aligns with the engineering understanding that the loss of coolant in the primary loop results in reduced heat removal capacity on the secondary side, leading to a significant increase in PSGA.

However, the order of the remaining top-contributing parameters differs for each model. The most influential variables for the regression ANN are TAVG, THA, and TCA, indicating that the model relies on the temperatures of the hot and cold legs and the average RCS temperature to predict PCT. This reliance is logical, as the variations in these temperatures directly impact the reactor core temperature.

On the other hand, the classification models utilize the volume, pressure, and average temperature of the RCS to characterize the type of classes, where these variables (VOL, P, and TVAG) have different natures and contribute closely to class prediction. This diversity leads to challenges in distinguishing between classes, reflected by a decrease in accuracy, unlike the regression model which strongly bases predictions on a single variable nature (temperature). The dominant contribution of these temperature variables results in excellent accuracy for the regression model.

To provide deeper interpretability and a clearer understanding of the model behaviour, the SHAP waterfall plot is utilized (Figure 9), which illustrates the local explainability of the models (regression and classification)

for one sample case. The results show that PSGA is the significant variable, consistent with the global explainability; however, its influence differs for each model, contributing positively in the classification model and negatively for the regression model. This validates the previous conclusion that the regression model relies on temperature variation to predict PCT.

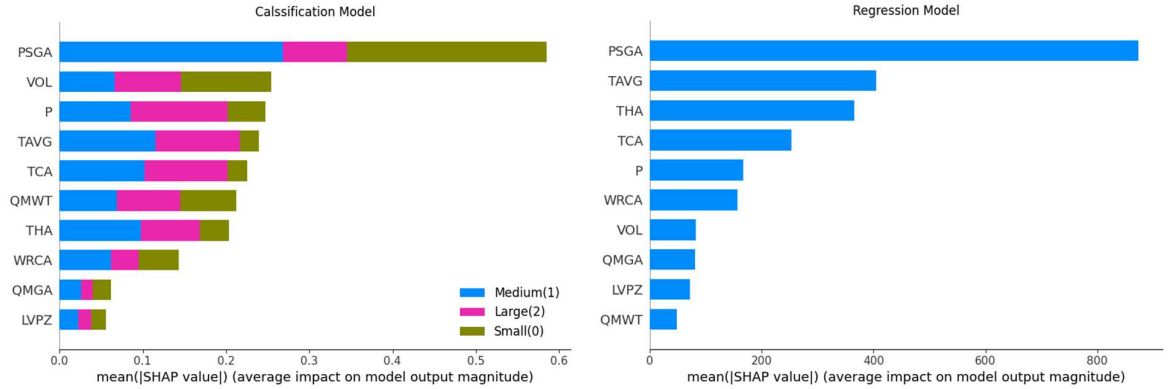

Figure 8. Classification and regression models' comparison for global interpretability

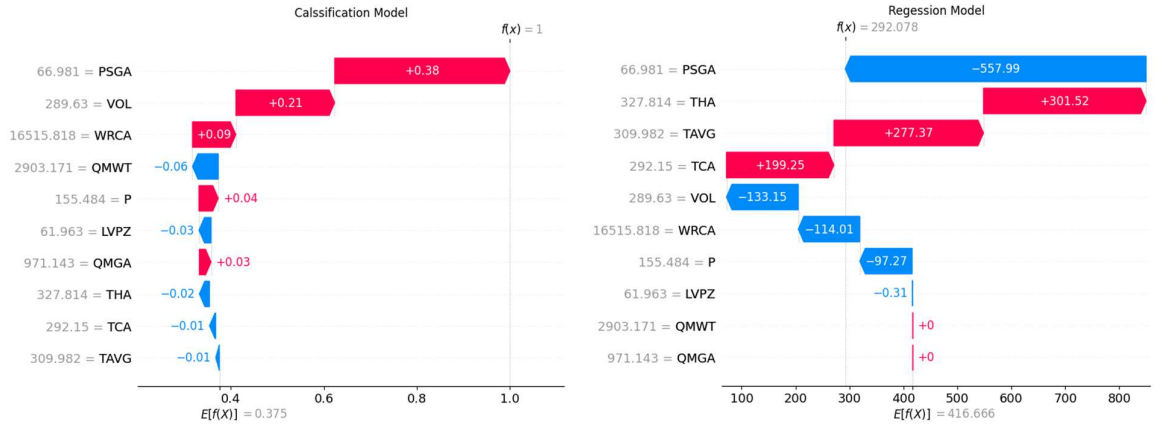

Figure 9. Classification and regression models' comparison for local interpretability

4. CONCLUSION

This work aims to enhance operator performance during LOCA by utilizing machine learning to forecast accident types and PCT values. This allows for an evaluation of safety margin variations and the severity of the accident, enabling operators to make swift and appropriate decisions to mitigate the consequences. The main conclusions are discussed as follows:

- The combined model demonstrated the best performance, characterized by greater simplicity and good accuracy, and was less time-consuming in predicting the type of failure along with PCT.
- The use of XAI explained the behavior of the combined model and the rationale behind its predictions.

Therefore, the combined model demonstrated both high performance and transparency, thus establishing operator trust in the model's predictions, leading to enhanced safety and reliability in nuclear power plants.

Acknowledgements

This work is supported by the Ling Chuang Research Project of China National Nuclear Corporation.

References

[1] J. Song, and K. Ha, "A simulation and machine learning informed diagnosis of the severe accidents," *Nuclear Engineering and Design*, vol. 395, pp. 111881, 2022/08/15/, 2022.

[2] T. C. H. Nguyen, and A. Diab, "Using machine learning to forecast and assess the uncertainty in the response of a typical PWR undergoing a steam generator tube rupture accident," *Nuclear Engineering and Technology*, vol. 55, no. 9, pp. 3423-3440, 2023/09/01/, 2023.

[3] Y. H. Chae, C. Lee, S. M. Han, and P. H. Seong, "Graph neural network based multiple accident diagnosis in nuclear power plants: Data optimization to represent the system configuration," *Nuclear Engineering and Technology*, vol. 54, no. 8, pp. 2859-2870, 2022.

[4] J. Choi, and S. J. Lee, "RNN-based integrated system for real-time sensor fault detection and fault-informed accident diagnosis in nuclear power plant accidents," *Nuclear Engineering and Technology*, vol. 55, no. 3, pp. 814-826, 2023/03/01/, 2023.

[5] M. Najar, and H. Wang, "Comparative Machine Learning Study for Estimating Peak Cladding Temperature in AP1000 Under LOFW." p. V005T05A022.

[6] V. Agarwal, C. M. Walker, K. Araseethota Manjunatha, T. J. Mortenson, N. J. Lybeck, and A. V. Gribok, *Technical Basis for Advanced Artificial Intelligence and Machine Learning Adoption in Nuclear Power Plants*, Idaho National Lab.(INL), Idaho Falls, ID (United States), 2022.

[7] C. M. Walker, V. Agarwal, L. Lin, A. C. Hall, R. A. Hill, R. L. Boring PhD, T. J. Mortenson, and N. J. Lybeck, *Explainable Artificial Intelligence Technology for Predictive Maintenance*, Idaho National Laboratory (INL), Idaho Falls, ID (United States), 2023.

[8] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *Peerj computer science*, vol. 7, pp. e623, 2021.

[9] C. Goutte, and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation." pp. 345-359.

[10] M. Sokolova, and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427-437, 2009.

[11] B. Qi, X. Xiao, J. Liang, L.-c. C. Po, L. Zhang, and J. Tong, "An open time-series simulated dataset covering various accidents for nuclear power plants," *Scientific Data*, vol. 9, no. 1, pp. 766, 2022.