Enhanced High-Resolution Prediction Surrogate Modeling with an Interpretable Attention Mechanism for Severe Accident Analysis

Wonung Jeong[1], Sangam Khanal[2], Semin Joo[3], Seokho Song[3], Jeongik Lee[3], and Joongoo Jeon[4*]
[1]*Department of Quantum System Engineering, Jeonbuk National University, Jeonju, Republic of Korea*
[2]*Graduate School of Integrated Energy-AI, Jeonbuk National University, Jeonju, Republic of Korea*
[3]*Department of Nuclear and Quantum Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea*
[4]*Division of Advanced Nuclear Engineering, Pohang University of Science and Technology, Pohang, Republic of Korea*
[*]*Corresponding author: jgjeon41@postech.ac.kr*

Keywords: severe accident, machine learning, nuclear safety

1. Introduction

Severe accident (SA) analysis in nuclear power plants has traditionally relied on system codes such as MELCOR and MAAP [1]. While these codes are generally fast and provide a comprehensive framework for analyzing plant behavior under SA [2], they face challenges in modern applications. Specifically, in current Probabilistic Safety Assessments (PSA), Fault Trees (FT) and Event Trees (ET) dictate the analysis; however, rather than considering all possible accident scenarios, experts are forced to perform limited PSAs based on selective considerations. Therefore, further computational acceleration is required to overcome these analytical constraints and enable more comprehensive evaluations [3]. Furthermore, the traditional structure of these legacy system codes makes it difficult to construct and integrate them into modern reinforcement learning frameworks required for training AI agents.

To address these challenges, recent studies have increasingly explored machine learning techniques to develop surrogate models capable of predicting accident progression with improved computational efficiency [4, 5]. Previous research has predominantly utilized conventional deep learning models, such as Artificial Neural Networks (ANNs) or Recurrent Neural Networks (RNNs), to map accident sequences and predict plant parameters.

However, these existing approaches possess notable limitations. They often struggle to capture long-term dependencies in complex time-series data; specifically, the vanishing gradient problem occurs over long sequences, which degrades accuracy when making autoregressive predictions. Furthermore, due to their inherent "black-box" nature, they lack the interpretability required to establish trust in safety-critical nuclear applications.

In this work, we aim to overcome the vanishing gradient problem of conventional RNN models by developing a surrogate model capable of properly learning the complexity of long time-series and making high-resolution predictions, while analyzing the multi-physical behavior of power plants through parallel time-series processing and attention mechanisms. We propose a Transformer-based surrogate model capable of predicting SA progression with high time resolution. This model effectively captures complex sequence dependencies, and visualizing the attention scores enhances its interpretability, demonstrating its potential as a reliable, AI-based SA analysis tool [6]. In summary, this study seeks to overcome the computational and analytical limitations of SA analysis and facilitate future AI agent training through a Transformer-based surrogate model, while validating the feasibility of trustworthy AI deployment in the nuclear domain via attention-based interpretability [7, 8].

2. Methods and Results

2.1 Dataset Overview

In this study, a Time Series Transformer model was trained on normalized MAAP simulation data for a total loss of component cooling water (TLOCCW) SA scenario in the OPR1000 plant. The dataset was adopted from a previous study [9], which selected the TLOCCW scenario based on standard Level 2 Probabilistic Safety Assessment (PSA) methodologies for the reference plant. In general Level 2 PSA frameworks, Plant Damage States (PDS) are utilized to cluster core damage accidents, significantly reducing the vast number of accident sequences to be analyzed while quantifying the frequency of each PDS. Through this probabilistic evaluation, the previous study identified TLOCCW as a primary target scenario because it exhibits a high risk contribution—determined by the product of its PDS frequency and accident progression fraction—making it one of the most dominant severe accident sequences.

This dataset was explicitly designed to reflect information realistically available to operators in the Main Control Room. It combines continuous thermal-hydraulic time-series variables with binary features representing component failure states and SA

management guideline (SAMG) signals. The overall composition of the input features used for model training is summarized in Table I.

#	(Target) thermal-hydraulic variable
1	Primary system pressure (PPS)
2	Cold leg temperature (Cold leg T)
3	Hot leg temperature (Hot leg T)
4	Reactor vessel water level (ZWV)
5	Steam generator pressure (SG P)
6	Steam generator water level (SG WL)
7	Maximum Core Exit Temperature (Max CET)
8	Containment pressure (CTMTP)
9	Pressurizer pressure (PZRP)
10	Pressurizer water level (PZRWL)
#	Component failure
1	Reactor coolant pump (RCP) seal LOCA
2	Letdown heat exchanger (HX)
3	High-pressure injection (HPI) pump
4	Low-pressure injection (LPI) pump
5	Containment spray system (CSS) pump
6	Motor-driven auxiliary feedwater (MDAFW) pump
7	Charging pump (CHP)
8	Refueling Water Storage Tank (RWST)
#	SAMG mitigation
1	Steam generator external injection
2	Reactor cooling system depressurization
3	Reactor cooling system external injection

Table I: Composition input feature

2.2 Prediction Setup

The model receives thermal-hydraulic variables, safety system signals, and SAMG variables as inputs. It is trained in an autoregressive (Fig 1), and only the thermal-hydraulic variables of the next time step are predicted, while the other variables serve solely as conditioning inputs [16].

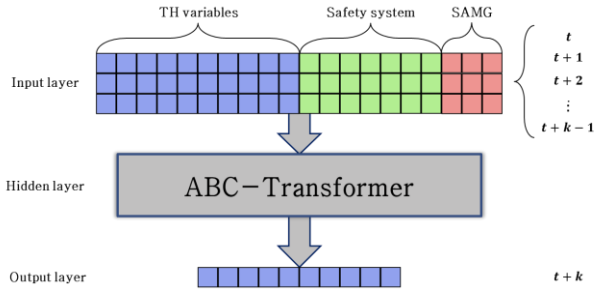

Fig. 1. Structure of the ABC-Transformer

2.3 ABC-Transformer

The proposed ABC-Transformer (Accident Binary–Continuous Transformer) embeds binary system-status signals and continuous thermal-hydraulic variables into a unified representation for severe-accident predicting (Fig 2). The binary inputs capture binary safety-system and SAMG actions, while the continuous inputs describe the plant's physical states [7]. The model employs only the decoder module, as all conditioning information is provided at each time step. The decoder integrates the embedded features and autoregressively predicts the next-step thermal-hydraulic variables.

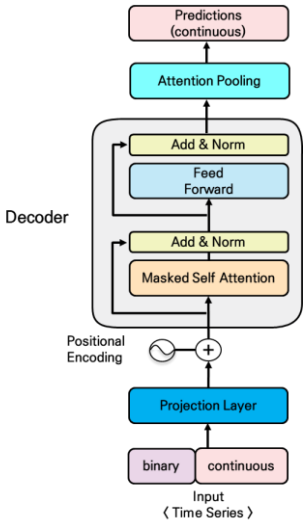

Figure 2 ABC-Transformer

An attention pooling layer is added to summarize the decoder outputs with learned attention weights, enabling the model to focus on the most influential time steps and improving long-range temporal dependency modeling [10].

2.4 Interpretability

The matrix operations within the proposed pooling layer, explicitly highlighting the explicit flow for generating final high-resolution predictions. The model is designed to apply a learned pooling query to the output hidden states (H^L), computing pooling weights that assign higher importance to the most critical time steps within the rolling window (Fig. 3).

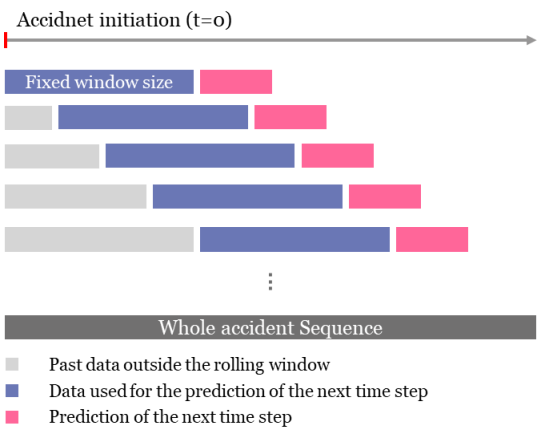

Fig. 3. Rolling window approach for time-series prediction

Crucially, these calculated pooling weights are then matrix-multiplied back with the original H^L to form an aggregated context vector. This context vector is subsequently fed through a dedicated Multi-Layer Perceptron (MLP) head to produce the final, accurate high-resolution predicted values. Notably, as shown in the pooling weight vector in Fig. 4, the highest weights are not assigned to the most recent time steps but rather to earlier points where abrupt changes occur in the thermal-hydraulic and system variables. This behavior indicates that the ABC-Transformer effectively identifies and attends to critical periods of rapid change in the severe-accident sequence, rather than relying solely on short-term temporal proximity.

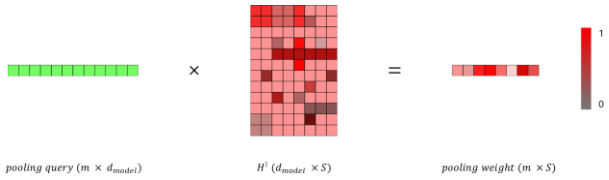

Fig. 4. pooling layer matrix operations

Specific analytical results are presented in Fig. 5. The upper panel presents the pooling weights computed over a fixed attention window of 30 steps (corresponding to a 15-minute resolution, translating to a strong model emphasis on the period about 3 to 4 hours before the current time step, or 12 to 13 steps prior). The lower panels show the attention distributions aligned with the true and predicted thermal-hydraulic variables.

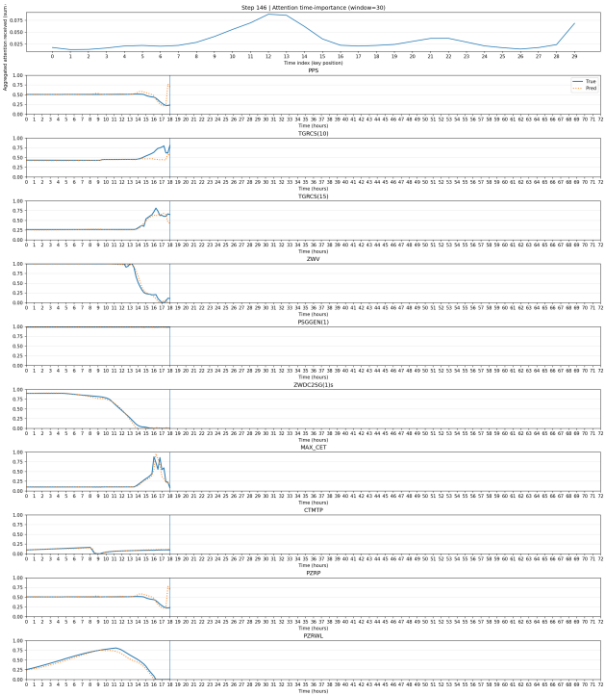

Fig. 5. Attention Weights with True and Predicted TH Values

This learned attention pattern directly aligns with the actual accident scenario, where key variables undergo critical abrupt changes during that specific timeframe. It is important to emphasize that these attention weights are dynamically computed for each unique accident episode and time step; they are not static importance factors but rather reflect the model's adaptive focus. Furthermore, since this interpretable capability is inherent to the model's architecture and did not require separate training or post-hoc analysis, the consistency between the model's focused attention and the physical accident timeline across varying moments provides strong evidence that the model has learned robust, generalized representations of severe accident dynamics.

2.5 Results

Table II summarizes the prediction performance (MAE and RMSE) of the models, evaluated using a teacher-forcing approach (Eq. 1) for 15 and 60 minute intervals.

For the 15-minute horizon, the proposed Transformer significantly outperforms all baselines, achieving the lowest MAE (0.0031) and RMSE (0.0058), a substantial improvement over the best baseline, convolution neural network – gated recurrent unit (CNN-GRU). For the 60-minute interval, while the Transformer's MAE (0.0069) is comparable to the CNN-GRU (0.0063), its RMSE is noticeably lower (0.0116 vs. 0.0163). Because RMSE heavily penalizes large errors, this reduction highlights the Transformer's robustness against major prediction deviations. Overall, these results confirm the model's superior capability in accurately capturing the complex temporal dynamics of SAs [11].

Table II: Performance Comparison Between the Previous Study Base Model and the Proposed ABC-Transformer

Model	Δt = 60 min		Δt = 15 min	
	MAE	**RMSE**	**MAE**	**RMSE**
CNN	0.0067	0.0170	0.0050	0.0140
LSTM	0.0095	0.0349	0.0053	0.0186
CNN-GRU	**0.0063**	0.0163	0.0049	0.0131
Transformer	0.0069	**0.0116**	**0.0031**	**0.0058**

$$\hat{x}_{t+1} = f([x_{t-L+1}, \dots, x_t])$$
$$\hat{x}_{t+2} = f([x_{t-L+2}, \dots, x_t, x_{t+1}]) \qquad (Eq.1)$$
$$\hat{x}_{t+k} = f([x_{t-L+k}, \dots, x_t, x_{t+1}, \dots, x_{t+k-1}])$$

t : *The current time step*
L : *The fixed length of the Rolling window size*
k : *The $k-th$ future time step to predict*
x_i : *The ground truth data at time step i*
$\hat{x}_{t+k}$: *The predicted value by the model at future tim* $+ k$

3. Conclusions

The proposed Transformer-based surrogate model outperforms conventional CNN and long short-term memory (LSTM) approaches in regression tasks. By effectively handling long temporal sequences, it enables

the high-resolution predictions crucial for severe-accident modeling. Furthermore, its attention mechanisms and pooling layer reveal which temporal segments drive predictions, overcoming traditional black-box limitations. This interpretability provides essential transparency and traceability, significantly enhancing the model's trustworthiness and practical applicability in safety-critical nuclear engineering environments.

Acknowledgement

This work was supported by the K-Cloud project of KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2024-Tech-08) and This work was supported by the Nuclear Safety Research Program through the Korea Foundation of Nuclear Safety (KoFONS) using the financial resource granted by the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (no. RS-2024-00403364).

REFERENCES

[1] R. E. Henry, C. Y. Paik, and M. G. Plys, “MAAP4 – Modular Accident Analysis Program for LWR Power Plants,” Research Project (1994): 3131–3102.
[2] International Nuclear Safety Advisory Group, “Basic Safety Principles for Nuclear Power Plants: 75-INSAG-3 Rev. 1,”(International Atomic Energy Agency, 1999).
[3] M. Saghafi and M. B. Ghofrani, “Accident Management Support Tools in Nuclear Power Plants: A Post-Fukushima Review,” Progress in Nuclear Energy 92 (2016): 1–14.
[4] H. Gu, G. Liu, J. Li, H. Xie, and H. Wen, “A Framework Based on Deep Learning for Predicting Multiple Safety-Critical Parameter Trends in Nuclear Power Plants,” Sustainability 15, no. 7 (2023): 6310.
[5] S. W. Oh, J. H. Park, H. S. Jo, and M. G. Na, “Development of an AI-Based Remaining Trip Time Prediction System for Nuclear Power Plants,” Nuclear Engineering and Technology 56, no. 8 (2024): 3167–3179.
[6] S. Yang, X. Yu, and Y. Zhou, “LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset As An Example,” in 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI), (IEEE, 2020): 98–101.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” Advances in Neural Information Processing Systems 30 (2017): 5998–6008.
[8] B. Lim, S. Ö. Arik, N. Loeff, and T. Pfister, “Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting,” Advances in Neural Information Processing Systems 33 (2019): 172–184.
[9] S. H. Song, S. Joo, Y. Lee, M. R. Seo, J. I. Lee*, "Evaluation of Multi-Input Single-Output ANN Models for Thermal-Hydraulic Predictions in Nuclear Severe Accidents: Branched vs. Non-Branched Structures," Transactions of the Korean Nuclear Society Spring Meeting, Jeju, Korea, May 22–23, 2025.
[10] H. Gholamalinezhad and H. Khosravi, “Pooling Methods in Deep Neural Networks: A Review,” Faculty of Electrical & Robotics Engineering, Shahrood University of Technology, Iran.
[11] S. Joo, Y. Lee, S. Song, K. Song, M. Seo, S. Kim, and J. Lee, “Application of deep neural network to an accelerated prediction of a severe accident in nuclear power plants,” International Journal of Energy Research, vol. 2025, no. 1, 2025.