Feasibility Study of Applying an Explainable AI (XAI) Model for an Accelerated Prediction of Severe Accident Progression

Semin Joo, Seok Ho Song, Yeonha Lee, Jeong Ik Lee*
Dept. Nuclear & Quantum Eng., KAIST, 291, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
**Corresponding author:* jeongiklee@kaist.ac.kr

Keywords : severe accident, explainable AI, machine learning, nuclear safety

1. Introduction

Since the past nuclear accidents, the importance of systematic management of severe accidents has gained the spotlight. Accordingly, the need to support severe accident management using the concept of Accident Management Support Tools (AMSTs) has emerged [1]. AMSTs aid in the assessment and mitigation of the consequences of severe accidents by predicting the progression of the accident and presenting possible options for the operator's actions.

Recent advancements in deep learning have opened new avenues for developing the AMST, owing to its fast computation speed and its excellent ability to comprehend nonlinear relationships of the given data. For instance, the leak flow in a loss-of-coolant-accident (LOCA) was predicted using a deep fuzzy neural network [2]. In another study, the convolutional neural network (CNN) was utilized to diagnose the faults under various power levels [3]. These studies underscore the capability of deep learning to enhance both the predictive accuracy and computation speed of the AMSTs.

However, deep learning is inherently bound to the 'model accuracy and explainability trade-off' problem (see Fig. 1). It refers to the AI models' tendency to lose explainability for the price of gaining prediction accuracy. Thus, the application of deep learning in nuclear safety introduces a new challenge: the need for explainability. Explainable Artificial Intelligence (XAI) addresses this by ensuring that the predictions made by AI models are transparent and understandable to human operators. This is crucial in the nuclear domain, where the rationale behind every decision must be clear to ensure trust and reliability.

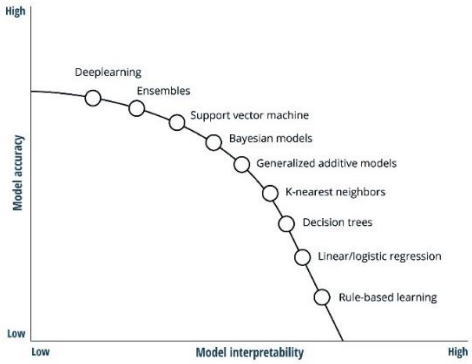

Fig. 1. Model accuracy and explainability trade-off.

It is in this background that this paper aims to explore the integration of XAI into AMSTs. In this study, a model that can predict the progression of a loss-of-component-cooling-water (LOCCW) accident is developed based on an attention mechanism, which is one of the XAI techniques. The prediction accuracy and the explainability of the proposed model will be assessed in comparison with the black box models.

2. Description of the accident dataset

In this section, the accident scenario of interest will be discussed. Based on the scenario, the datasets for training the models are produced with Modular Accident Analysis Program (MAAP) 5.03 code.

2.1 Selection of the Accident Scenario

The focus of the study is on the LOCCW accident in the OPR1000 reactor. In the event of a total-LOCCW (TLOCCW) accident, simultaneous failure occurs in seven safety-related components, as listed in Table I. However, this study also explores a subset of TLOCCW accidents to test the versatility of the machine learning model. It is presumed that the failure of safety components obeys a uniform random distribution, except the RCP seal LOCA, which is assumed to follow a lognormal distribution with around 89.2% of such failures happening within the first hour.

Table I**:** List of safety components that fail at TLOCCW accident.

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)

Together with component failures, accident mitigation strategies are also considered. Three mitigation strategies from the severe accident management guidelines (SAMGs) were adopted: water injection to the steam generator secondary side (M1), depressurization of the reactor coolant system (RCS) (M2), and water injection to the RCS (M3). These strategies are also assumed to be activated randomly in time throughout the 72-hour accident.

2.2 Dataset production

The MAAP code was utilized to simulate the outlined accident scenarios. This code forecasts the evolution of a severe accident situation for 72 hours, following the PSA mission time. A selection of ten thermal-hydraulic (TH) variables, which are observed in the main control room (MCR), were identified as key TH variables for analysis, as outlined in Table II. These variables serve as crucial markers for assessing the integrity of the reactor core and determining the conditions for activating the mitigation strategies. This results in the creation of a dataset for each accident scenario, containing ten time series with a length of 72 hours. In total, the MAAP code generated 12,121 accident scenarios. Subsequently, the data was normalized to ensure all values ranged from zero to one.

Table II: List of target TH variables

1	Primary system pressure (PPS)
2	Cold leg temperature (CLT)
3	Hot leg temperature (HLT)
4	Reactor vessel water level (RV WL)
5	Steam generator pressure (SG P)
6	Steam generator water level (SG WL)
7	Maximum core exit temperature (Max CET)
8	Containment pressure (CTMT P)
9	Pressurizer water level (PZR WL)
10	Pressurizer pressure (PZR P)

3. Model development

Following the production of accident datasets, the machine learning model is constructed. In this section, the architecture of the XAI model and its training and testing processes are discussed.

3.1 Description of the XAI model architecture

The 'attention mechanism' has been a popular technique in computer vision since its highly cited publication in 2017 [4]. The attention mechanism allows the neural networks to focus on specific parts of the input data that are most relevant to the task at hand. It is the 'attention weights' that indicate which part of the input data will be weighted more to make an efficient, accurate prediction. For instance, in tasks like machine translation, the attention weights reveal which words in the source language are most relevant to each word in the target language, providing insights into the translation process. Due to this characteristic, attention-based models are often classified as an XAI model [5]. From this background, this study adopts the attention mechanism to investigate its explainability and performance when applied to severe accident prediction.

This study utilizes the model proposed in the work of Y. Qin et al. [6] in particular – the dual-stage attention-based recurrent neural network (DA-RNN). The architecture of the DA-RNN model is depicted in Fig. 2. In the first stage, the input attention mechanism adaptively extracts the input features at each step by referring to the previous encoder's hidden state. At the subsequent stage, the temporal attention mechanism selects the relevant encoder's hidden states across all time steps. Thus, the feature importance of an input feature i at time step t is reflected in the attention weight α_t^i. The encoder and decoder's hidden states were learned through the long short-term memory (LSTM) units. For hyperparameter adjustment, various numbers of LSTM units were tested in this study: 8, 16, 32, 64, 128.

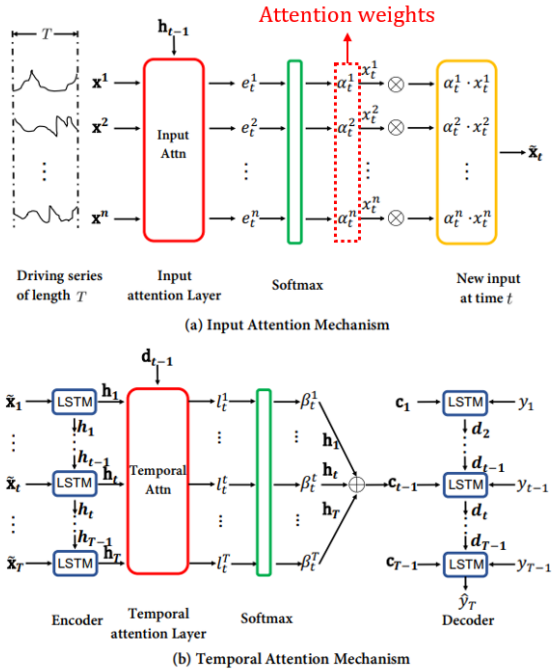

Fig. 2. Architecture of the attention-based XAI model [6].

The model takes the plant's state at the previous five-time steps as an input. One time step consists of the ten TH variables (listed in Table II), whether the seven components fail (1) or not (0) (listed in Table I), and whether the three SAMG strategies are activated (1) or not (0). Based on this input, the model forecasts the TH variable at the next time step. Each model is dedicated to predicting a single target TH variable, so there are ten separate models to be developed (see Fig. 3).

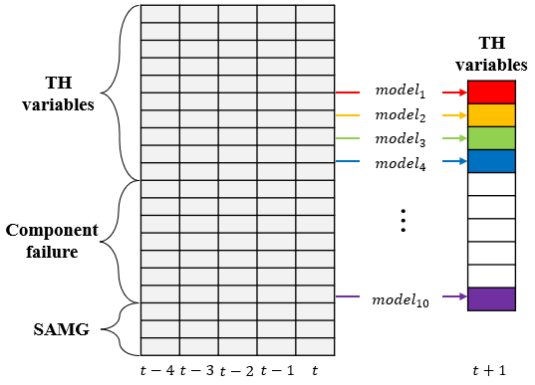

Fig. 3. Structure of the model input and output

3.2 Training and Test Methods

The datasets are divided into training sets (80%), validation sets (10%), and test sets (10%). The training set is fed randomly into the XAI models to train the models. The training process is stopped when the validation loss does not decrease for more than 50 epochs.

When the training process is completed, the performance of the trained model is evaluated using the test sets. The prediction accuracy of each model was evaluated by calculating the mean absolute error (MAE) and root mean squared error (RMSE) over the test dataset.

4. Results and Discussions

4.1 Model Accuracy

First, the prediction accuracy of the model was evaluated through MAE and RMSE values. Fig. 4. shows the boxplots of MAE and RMSE as a function of the number of nodes in the LSTM, for all target TH variables. Usually, the prediction accuracy of a model increases as the number of nodes in the LSTM increases. The MAE and RMSE tended to decrease as the number of nodes increased from 8 to 64, but the degree of improvement was not significant. Thus, it was deduced that the number of nodes in the LSTM unit did not have a marked influence on the attention-based model's performance.

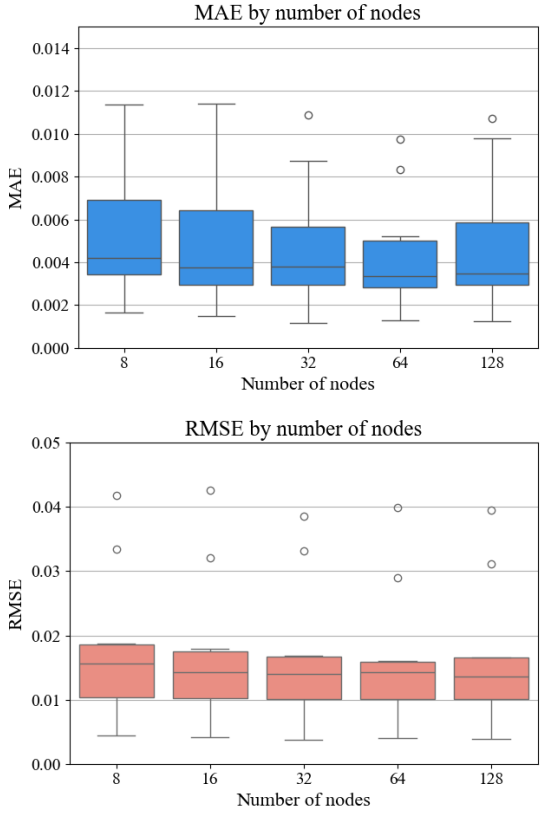

Fig. 4. Comparison of MAE (top) and RMSE (bottom) by the number of nodes in the LSTM unit. The circles represent the outliers.

However, the model's performance was significantly dependent on the type of the target TH variable. Fig. 5. shows the boxplots of MAE and RMSE as a function of target TH variables for various numbers of nodes. It was found that the type of target TH variable has a significant effect on the models' performance, regardless of the number of nodes. The MAE and RMSE values of RV WL and Max CET prediction were notably higher than the other TH variables, implying the relative difficulty in predicting the RV WL and Max CET in LOCCW accident scenarios.

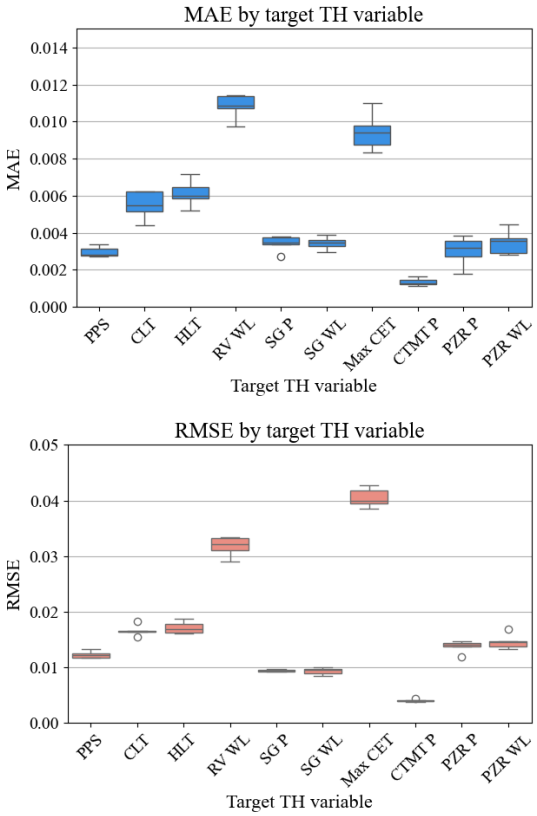

Fig. 5. Comparison of MAE (top) and RMSE (bottom) by the model's target TH variable. The circles represent the outliers.

The main purpose of this study is to develop a predictor model with explanatory power, but the prediction accuracy of the model is an important goal that cannot be ignored. Hence, it is necessary to compare the performance with the black box models in terms of prediction accuracy.

In the previous studies [7, 8], various predictor models have been developed that are based on the multi-layer perceptron (MLP), convolutional neural network (CNN), and long short-term memory (LSTM) architectures. The LSTM is an architecture specialized for long-time series data processing, thus the LSTM-based model showed excellent performance. Therefore, the performances of the attention-based model developed in this study and the LSTM model are compared. Table III compares the RMSE values of the LSTM model and the attention-based model that showed the best performance. The

attention-based model that had the lowest MAE and RMSE values on average was the one with 128 nodes in the LSTM units (N=128).

Table III: RMSE of LSTM models and attention-based models and their relative differences.

Target variable	RMSE		
	LSTM	Attention (N=128)	Relative difference [%]
Averaged	1.09E-02	1.63E-02	49.99
PPS	7.10E-03	1.25E-02	76.61
CLT	1.09E-02	1.54E-02	40.67
HLT	1.16E-02	1.60E-02	38.22
RV WL	2.11E-02	2.90E-02	37.66
SG P	6.88E-03	9.20E-03	33.77
SG WL	5.44E-03	8.52E-03	56.55
Max CET	2.73E-02	3.99E-02	46.38
CTMT P	2.48E-03	4.01E-03	61.63
PZR P	8.36E-03	1.40E-02	67.42
PZR WL	7.62E-03	1.45E-02	90.56

It is observed that the LSTM model had smaller RMSE values on average. The increase in the RMSE values of the attention model compared to the LSTM model implies that the prediction accuracy did not improve by employing the attention mechanism.

4.2 Model explainability

To prove that the proposed model is truly explainable, it should be shown that the explanations provided by the model accord with the phenomenological explanations. To do so, the similarity between the attention weights of the model and the feature importance of the MAAP data should be investigated. Here, feature importance represents the importance of an input parameter X for predicting a target variable Y.

In this study, the feature importance of the training data (that is, the MAAP dataset) is described using an index called 'mutual information (MI)'. As the TH variables of a nuclear power plant are in nonlinear relationships, it would be inappropriate to use the commonly used correlation coefficients such as Pearson and Spearman correlation coefficients. On the other hand, MI is a fundamental concept in information theory that comprehends the nonlinear relationship within the data. MI is defined as the expected value of the pointwise mutual information between two random variables X and Y (Eq. (2)). Hence, it represents the amount of information obtained about one random variable by observing the other.

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\ \log\frac{p(x,y)}{p(x)p(y)} \qquad Eq.(1)$$

The MI values between each target TH variable and the input TH variables were calculated. In Fig. 6, they are presented as a bar graph together with the attention weights. To directly compare with the attention weights, the MI values have been normalized so that their sum equals unity. The attention weights for each target TH variable are represented as a boxplot, as the weights vary for models with different numbers of nodes in the LSTM unit.

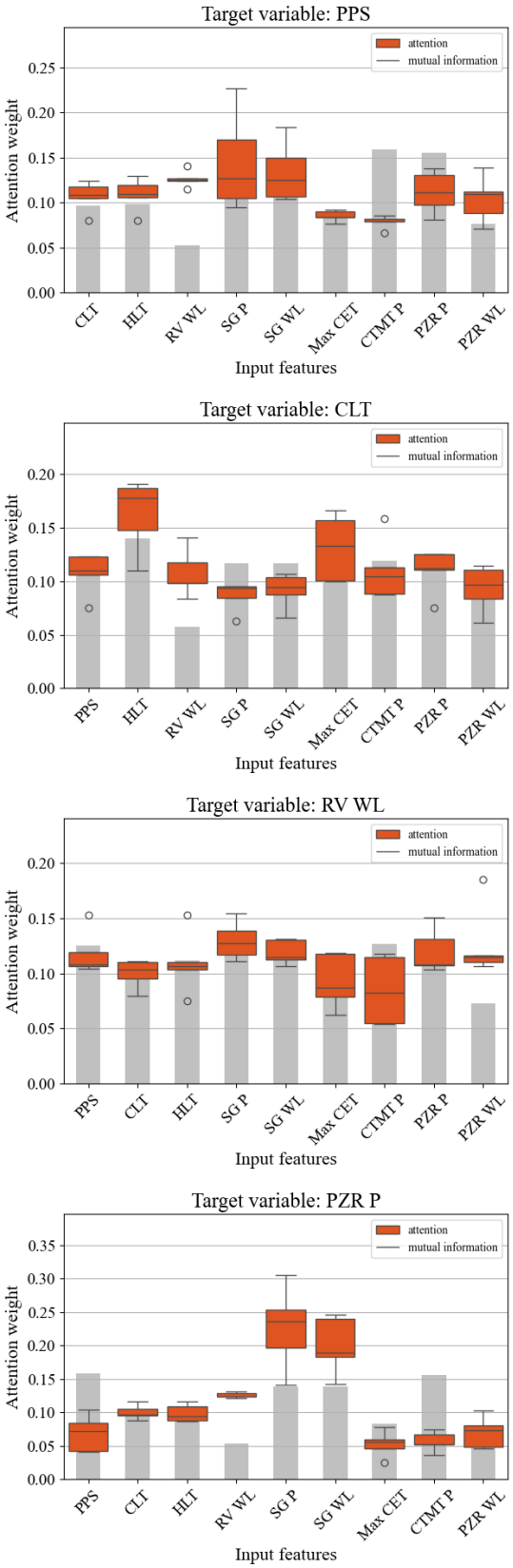

Fig. 6. Distribution of attention weights and mutual information for target TH variables: PPS, CLT, RV WL, and PZR P. The circles represent the outliers.

Taking CLT as an example, the PPS, HLT, and Max CET seems to have relatively high importance in its prediction. In the physical sense, this is because the cold leg, core exit, and hot leg all constitute the primary flow together. Furthermore, the temperatures of the primary coolant are thermo-physically correlated with its pressure (PPS). Thus, it seems obvious that the mutual information between the HLT and CLT, Max CET, PPS are high, and the attention weights appears to reflect this phenomenon well.

For most target TH variables, the attention weight of RV WL was found to be more than twice the value of mutual information. In other words, RV WL was used to predict specific target variables with more importance than necessary. As RV WL took away the large attention weight, the degree of contribution of CTMT P was reduced. This effect was noticeable in the prediction of PPS and PZR P. However, considering that the prediction accuracy of PPS and PZR P is not lower than that of other variables (refer to Table III), it is difficult to say that this effect has a significant influence on the performance of the model.

4.3 Similarity between attention weights and mutual information

In the previous section, attention weight and mutual information were calculated as indicators for feature importance used to predict a specific target TH variable. To prove that the attention weight of the attention-based model reflects the phenomenological explanation, it must be shown that the distributions of the attention weight and mutual information are similar. A commonly used indicator to show similarity between the two rank matrices is 'cosine similarity'. Here, cosine similarity (S_C) is estimated to measure the similarity between the mutual information matrix (MI) obtained from the MAAP data and the attention weight matrix (Att) obtained through the model training (see Eq. (2)). The more similar the two rank matrices are to each other, the closer the cosine similarity will be to 1.

$$S_C(MI, Att) = \frac{\langle MI, Att \rangle_F}{\|MI\|_F \, \|Att\|_F} \qquad Eq.(2)$$

Fig. 7. is a boxplot graph showing the calculated S_C values for each target TH variable. Since the attention weight is different according to the model's number of nodes, the cosine similarity was also different according to the model, so it was described in the form of a boxplot to show the distribution.

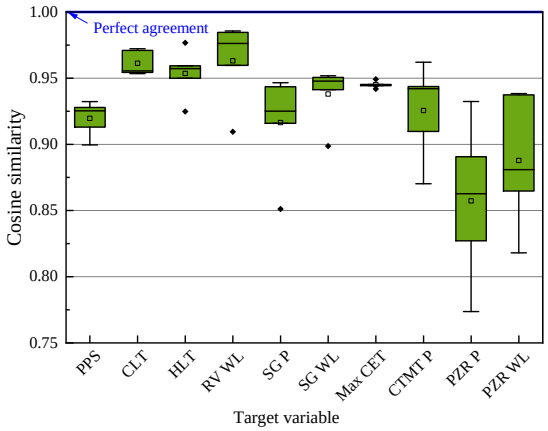

Fig. 7. Cosine similarity between attention weights and mutual information.

It is observed that the cosine similarities of all models fell in the range of 0.77 to 0.98 in all models. Thus, it can be interpreted that the proposed model learns the feature importance of the training data and embodies it as a form of attention weight. Especially for PPS, CLT, HLT, RV WL, and Max CET, the cosine similarity of all models was higher than 0.9, showing an evident potential for model explainability.

It was also investigated whether the model's performance is improved if the attention weight and mutual information are similar. Looking at RV WL and Max CET, which had the worst prediction performance, the cosine similarity is always higher than 0.9. In other words, even if the attention weight of the model is well explained in phenomenological terms, the prediction accuracy of the model may not improve. The following two reasons can be considered for this.

First, it is the characteristics of the data itself. In the previous studies where the traditional black box models were considered [7, 8], the prediction accuracies of RV WL and Max CET were also notably lower than those of other variables. Thus, it is expected that the RV WL and Max CET have characteristics that are difficult to learn by machine learning.

The second reason stems from the 'model accuracy and explainability trade-off'. As mentioned earlier, the better the explainability of the model, the lower the prediction accuracy of the model tends to be. Based on this concept, it is understood to some extent that the attention weight in the model predicting Max CET or RV WL has high explainability by reflecting the mutual information well, but the prediction accuracy is rather poor. Through this, searching for a balance between accuracy and explainability seems to be an essential component in developing an XAI for nuclear safety.

5. Conclusions and Further Works

In this study, a model that can predict the progression of a LOCCW accident was developed using the concept of explainable AI. The main conclusions of this study can be summarized as follows.

- By employing an attention-based architecture, deep learning models have the potential to be explainable for their predictions.
- Attention-based models do not show a noticeable improvement in prediction accuracy compared to traditional black box models (LSTM), but they still have a reasonable accuracy.
- In other words, it is possible to develop an AI-based AMST predictor model with both explainability and high accuracy. Such potential is expected to serve as a lubricant in applying AI models to severe accident management.

Future research will explore strategies to enhance the predictive accuracy without compromising the model's interpretability. This approach aims to achieve a balance between the explainability and accuracy of the XAI model. Also, the potential of applying other XAI techniques to predict severe accidents will be assessed.

This study used ten thermal hydraulic variables that are observable at the MCR as the input variables. It is expected that the performance of the model can be improved by adding more MCR monitoring variables.

Acknowledgment

This work was supported by KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2020-Tech-01).

REFERENCES

[1] M. Saghafi, M. B. Ghofrani, Accident management support tools in nuclear power plants: A post-Fukushima review, Progress in Nuclear Energy 92, 2016.
[2] J. H. Park, Y. J. An, K. H. Yoo, M. G. Na, Leak flow prediction during loss of coolant accidents using deep fuzzy neural networks. Nuclear Engineering and Technology, 2021.
[3] M. Lin, J. Li, Y. Li, X. Wang, C. Jin, J. Chen, Generalization analysis and improvement of CNN-based nuclear power plant fault diagnosis model under varying power levels. Energy, 282, 128905, 2023.
[4] A. Vaswani et al., Attention is all you need, Advances in Neural Information Processing Systems 30, 2017.
[5] S. Wiegreffe, Y. Pinter, Attention is not not explanation, Conference on Empirical Methods in Natural Language Processing, 2019.
[6] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, G. Cottrell, A dual-stage attention-based recurrent neural network for time series prediction, International Joint Conference on Artificial Intelligence, 2017.
[7] Y. Lee, Development of accelerated prediction method using artificial neural network for Nuclear Power Plant Severe Accident application, Master's thesis, Korea Advanced Institute of Science and Technology, 2022.
[8] S. Joo, S. H. Song, Y. Lee, J. I. Lee, S. J. Kim, Accelerated prediction of severe accident progression: Sensitivity of deep neural network performance to time resolution, Transactions of the Korean Nuclear Society Autumn Meeting, Gyeongju, Korea, October 26-27, 2023.