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

Nuclear Power & Propulsion Lab

Semin Joo, Seok Ho Song, Yeonha Lee, Jeong Ik Lee*

Department of Nuclear and Quantum Engineering N7-1 KAIST 291 Daehak-ro, Yuseong-gu,
Daejeon, Republic of Korea 34141, semin8504@kaist.ac.kr
**Corresponding author: jeongiklee@kaist.ac.kr*

2024.05.09

Contents

1. Introduction
2. Description of the Accident Dataset
3. Model Development
4. Results and Discussion
5. Conclusions and Further Works

Appendix

References

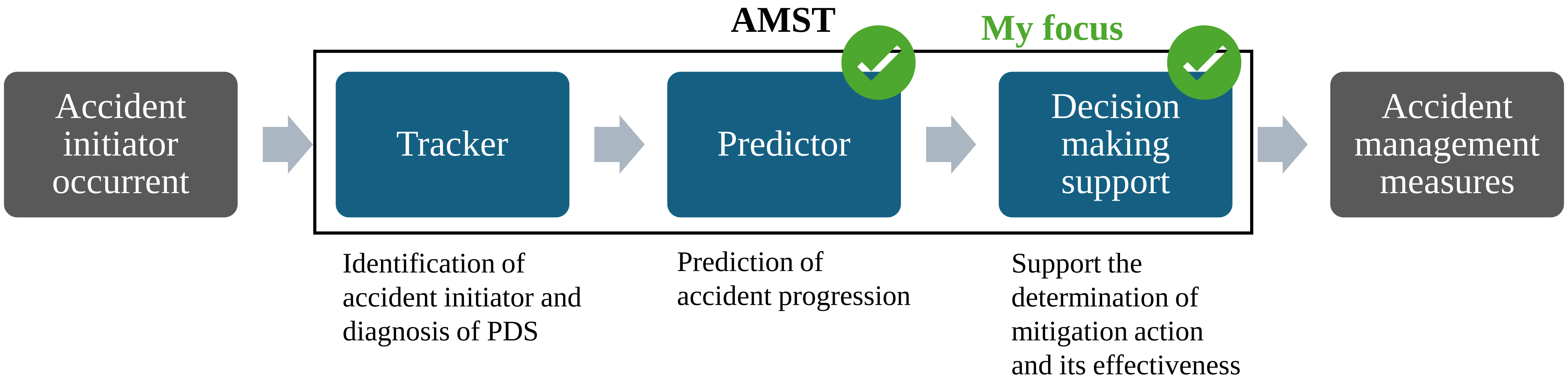
01

Introduction

Motivation

❑ Challenges in predicting and managing severe accidents

- Severe accidents are highly nonlinear and chaotic.
- DSA & PSA-based methods require large computational resources.
- Need to develop an Accident Management Support Tool (AMST) based on advanced computing methods, such as machine learning.



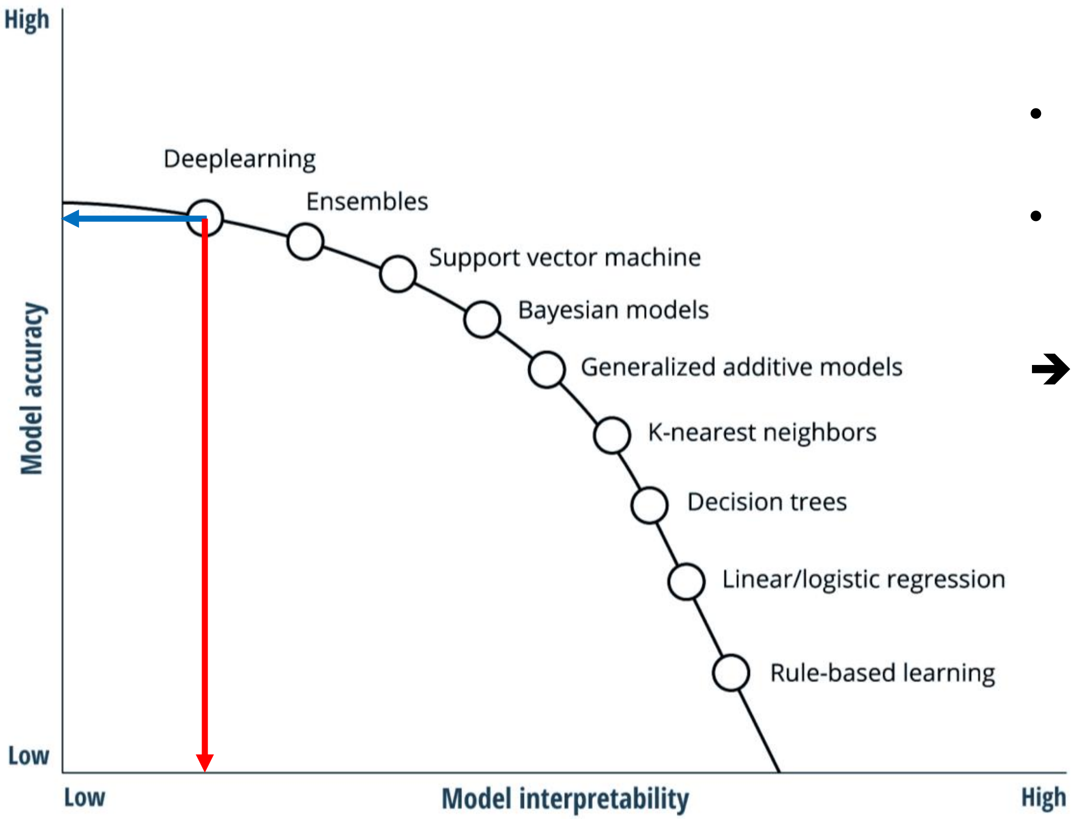
AMST
My focus
✓
✓
Accident initiator occurrent
Tracker
Predictor
Decision making support
Accident management measures
Identification of accident initiator and diagnosis of PDS
Prediction of accident progression
Support the determination of mitigation action and its effectiveness

Fig. 1. General structure of an AMST*

*M. Saghafi, M. B. Ghofrani, Accident management support tools in nuclear power plants: A post-Fukushima review, Progress in Nuclear Energy 92, 2016.

Motivation

❑ There are inherent trade-offs between the ML model's accuracy and interpretability.



Source: DPhi, "Importance of Human Interpretable models & Explainable AI," video featuring Dipanjan (DJ) Sarkar, 29:02, February 13, 2021.

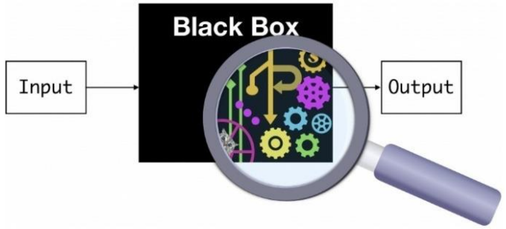
Fig. 2. Trade-offs between model accuracy and interpretability

- Deep learning models are excellent at learning the non-linear, complex patterns of the given data.
- However, the model directly maps input data to output predictions without explicitly decomposing the problem into interpretable subtasks.

➔ Deep learning methods have high prediction accuracy but lack 'interpretability'.

➔ Need for a method that can **explain/interpret** the deep learning models

Explainable AI (XAI)



❑ What is XAI?

- A subfield of AI that focuses on creating AI models whose actions can be easily understood by humans
- Goal: build trust in AI systems to make them more useful across various fields

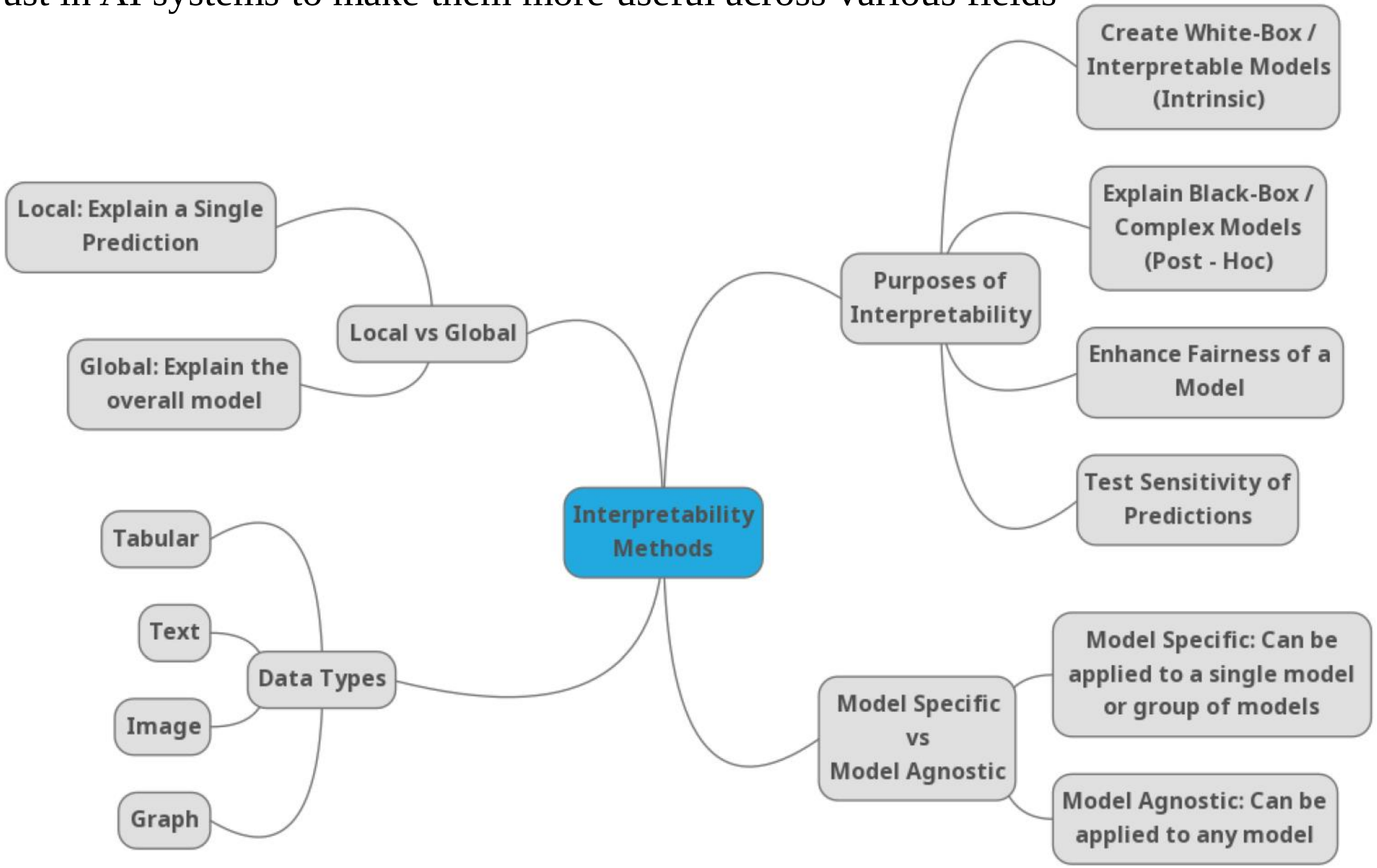
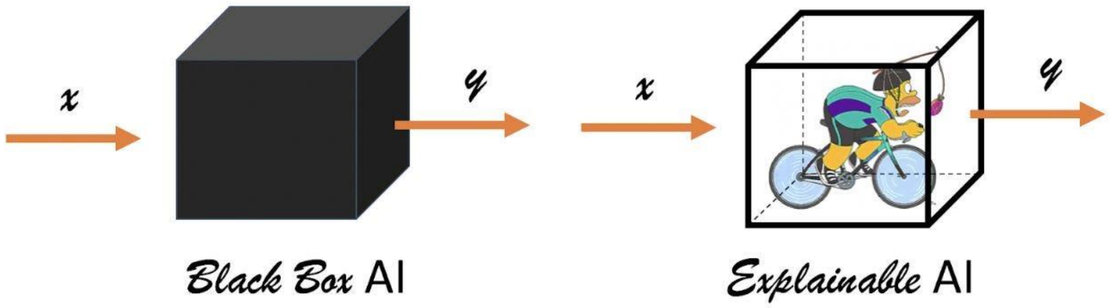

Interpretability Methods
Local vs Global
Local: Explain a Single Prediction
Global: Explain the overall model
Data Types
Tabular
Text
Image
Graph
Purposes of Interpretability
Create White-Box / Interpretable Models (Intrinsic)
Explain Black-Box / Complex Models (Post - Hoc)
Enhance Fairness of a Model
Test Sensitivity of Predictions
Model Specific vs Model Agnostic
Model Specific: Can be applied to a single model or group of models
Model Agnostic: Can be applied to any model

Fig. 3. Types of interpretability methods for XAI [1]

Research Objectives

1. Explore the feasibility of integrating XAI into AMSTs.
2. Develop a model that can predict the progression of a severe accident scenario based on an attention mechanism, which is one of the XAI techniques.
3. Prediction accuracy and the explainability of the proposed model will be assessed in comparison with the black box models.



02

Description of the Accident Dataset

Accident scenario

❑ Description of the accident scenario

- Reference reactor type: OPR1000
- Subsets of Total Loss of Component Cooling Water (TLOCCW) accident
  - Multiple failures in the safety components lead to reactor core damage (Fig. 4)
  - Various mitigation strategies are applicable (Table 1)
- Duration of a single accident scenario: 72 hr (=PSA mission time)

*RCP = Reactor Coolant Pump
*HX = Heat Exchanger
*HPI = High-Pressure Injection
*LPI = Low-Pressure Injection
*CSS = Containment Spray System
*MDAFW = Motor-Driven Auxiliary Feedwater
*CHP = Charging Pump

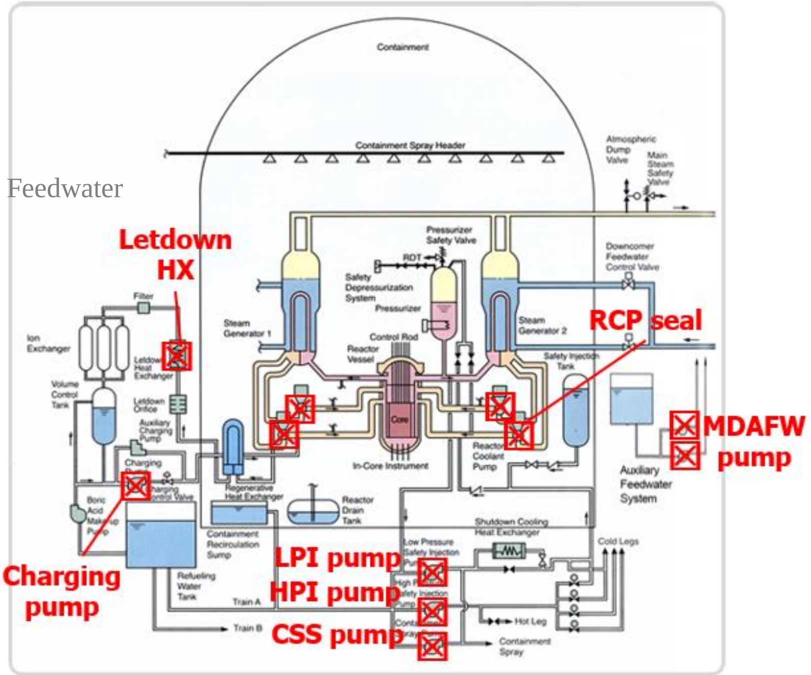


Fig. 4. Locations of component failure at OPR1000 system

Table 1. Types of mitigation strategies
(OPR1000 Severe Accident Management Guidelines)

#	Mitigation Strategy
SAMG-01	Steam generator (SG) external injection
SAMG-02	Reactor coolant system (RCS) depressurization
SAMG-03	RCS external injection

Dataset production

- ❑ Dataset production

Component failure/ SAMG activation

Fail mostly at t = 1hr	RCP seal LOCA
Fail or not?	HX
Fail or not?	HPI pump
Fail or not?	LPI pump
Fail or not?	CSS pump
Fail or not?	MDAFW pump
Fail or not?	CHP
Activate or not?	SAMG-01
Activate or not?	SAMG-02
Activate or not?	SAMG-03

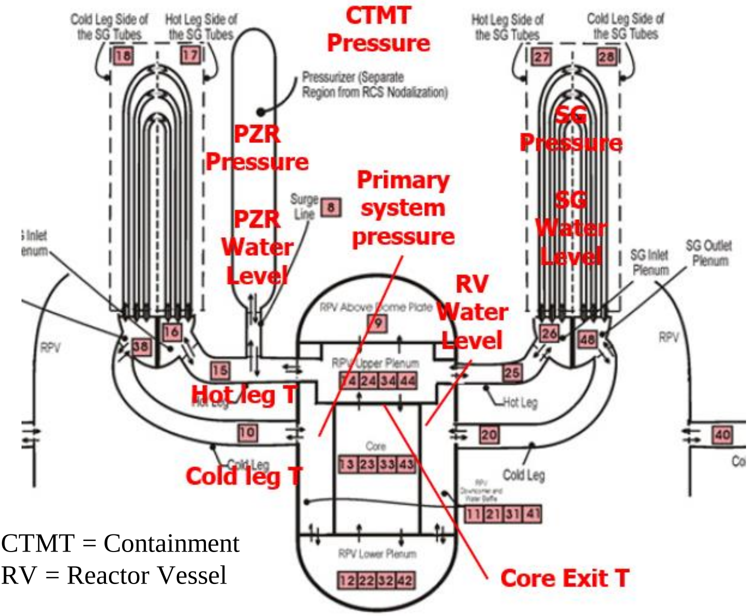
→ 11,000 accident scenarios → MAAP 5.03 Code →

TH variables

- Primary system pressure
- Cold leg temperature
- Hot leg temperature
- RV water level
- SG pressure
- SG water level
- Max. core exit temperature
- Containment pressure
- Pressurizer pressure
- Pressurizer water level

*Observable from the MCR and SAMG supervisory variables

*Appendix: Examples of accident dataset



*CTMT = Containment
*RV = Reactor Vessel

Fig. 5. Location of TH variables at OPR1000

03

Model Development

Model Development

❑ Input and output structure

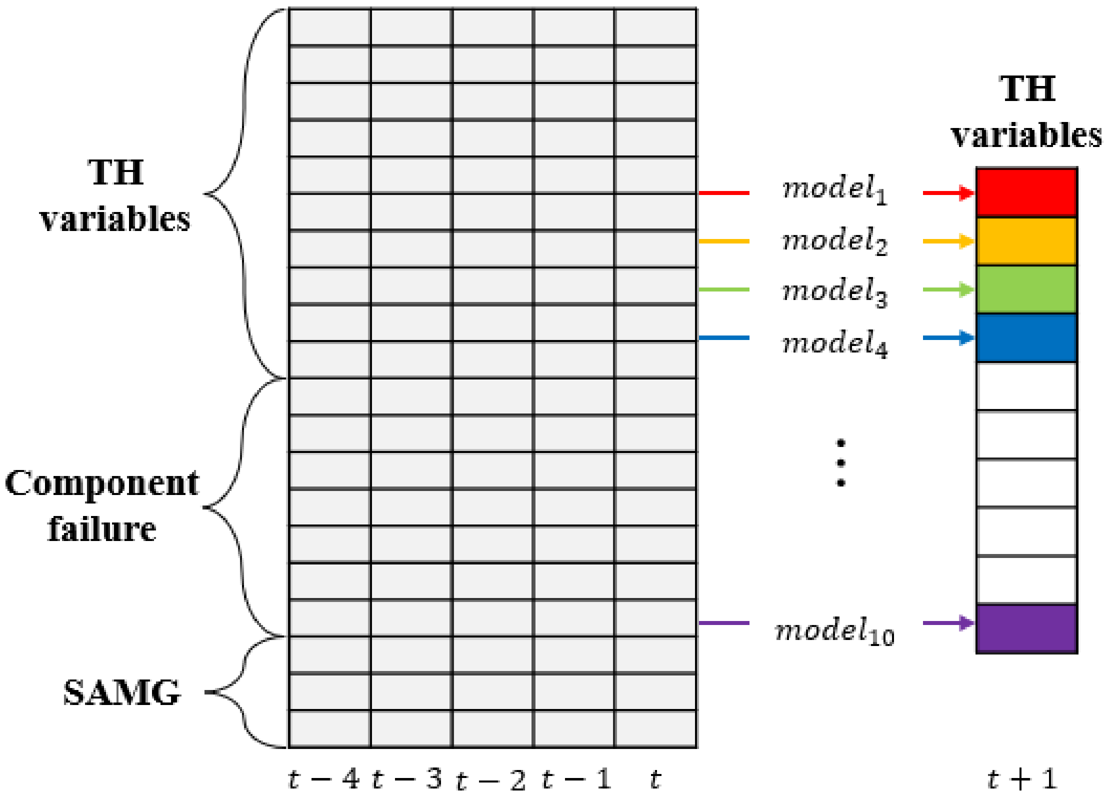


Fig. 6. Input and output structure of the prediction model

Model Development

- ❑ Model comparison: Blackbox model vs XAI model
  - ▪ XAI model - **Dual-stage Attention Recurrent Neural Network (DA-RNN)***
    - Devised for multivariable time series forecasting
    - Input attention – selectively weights the importance of input features
    - Temporal attention – selectively weights the importance of each time step

    ➔ The attention weights can be an explanation for the feature importance!

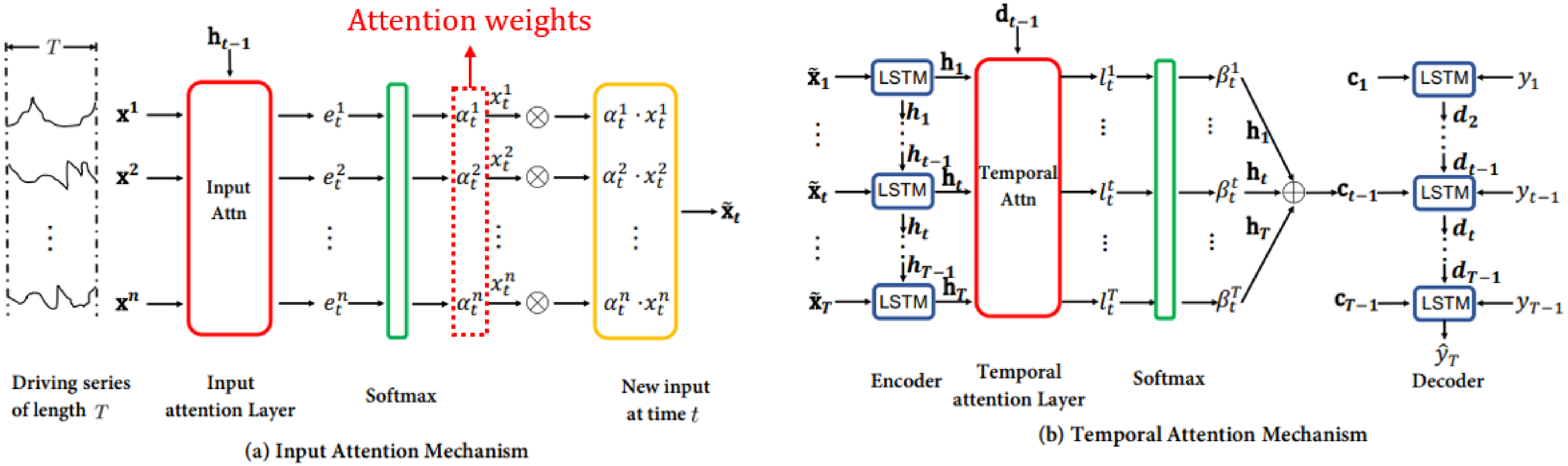


Fig. 7. DA-RNN architecture*

*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.

Model Development

- Model comparison: Blackbox model vs XAI model
  - Blackbox model - **Long Short-Term Memory (LSTM)**
    - A classic deep learning architecture for time series forecasting
    - In our previous studies, LSTM models have shown excellent regression performances [9].
    - The performance of the LSTM will be also evaluated for comparison with the XAI model.

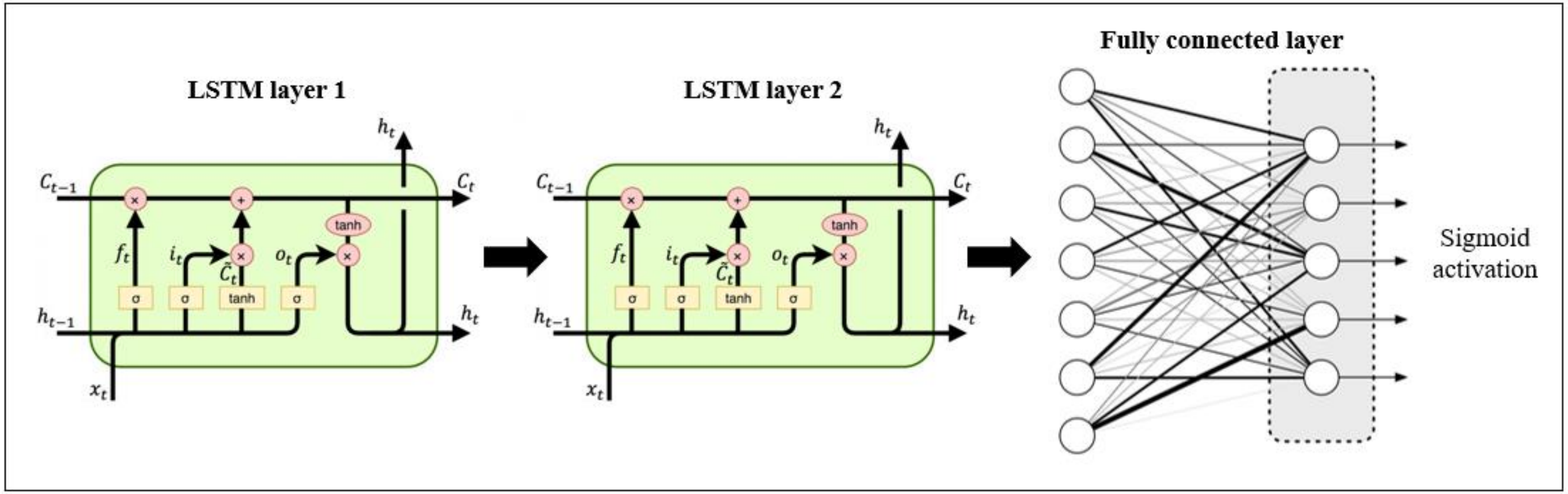
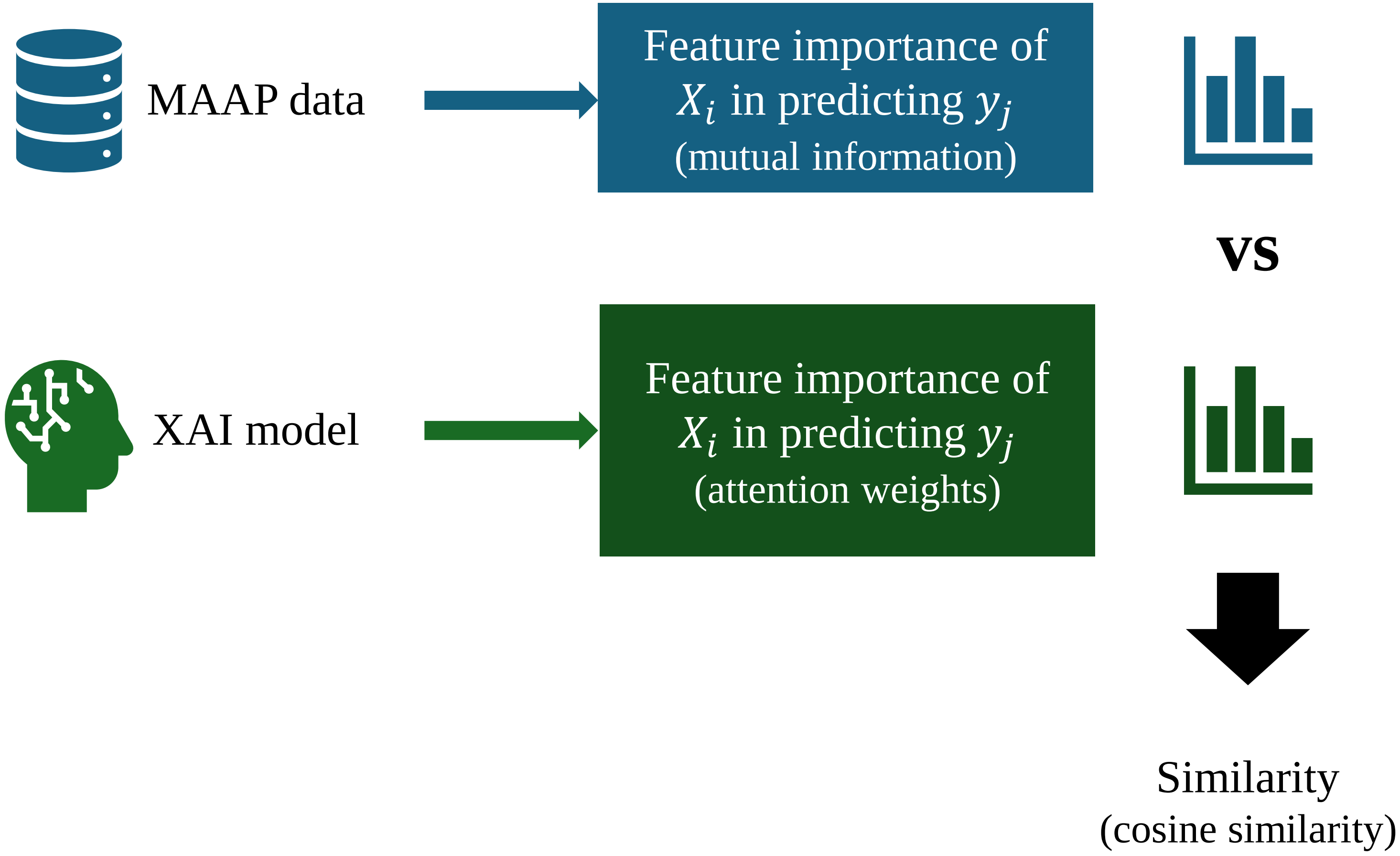


Fig. 8. Architecture of the LSTM model

Performance evaluation

❑ Model explainability



MAAP data

Feature importance of X_i in predicting y_j (mutual information)

vs

XAI model

Feature importance of X_i in predicting y_j (attention weights)

Similarity
(cosine similarity)

04

Results and Discussion

Prediction accuracy

❑ Hyperparameter test

- Number of nodes in the LSTM unit: 8, 16, 32, 64, 128
- 8 to 64: RMSE decreased by a small degree

⇒ Number of nodes in the LSTM unit does not have a marked influence on the model's regression performance.

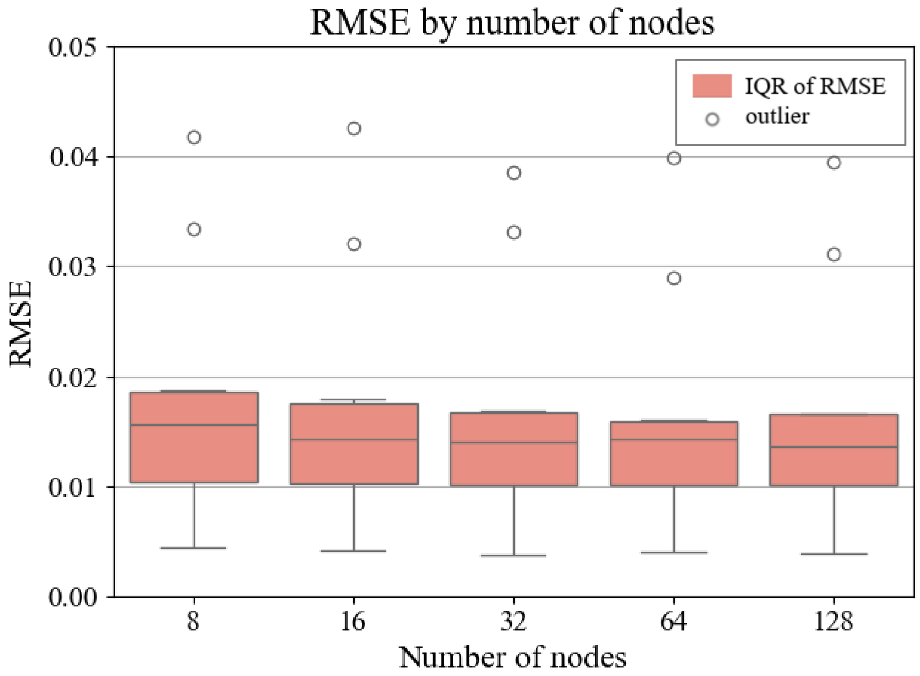


Fig. 9. MAE values for different target TH variables by the number of nodes in the LSTM unit. The IQR refers to the RMSE of models with various target TH variables.

❑ TH variable dependency

- Best vs. Worst: CTMT P vs. MAX CET
- The RMSE of predicting MAX CET was about 10 times larger than that of CTMT P.

⇒ the type of target TH variable has a significant effect on the models' performance.

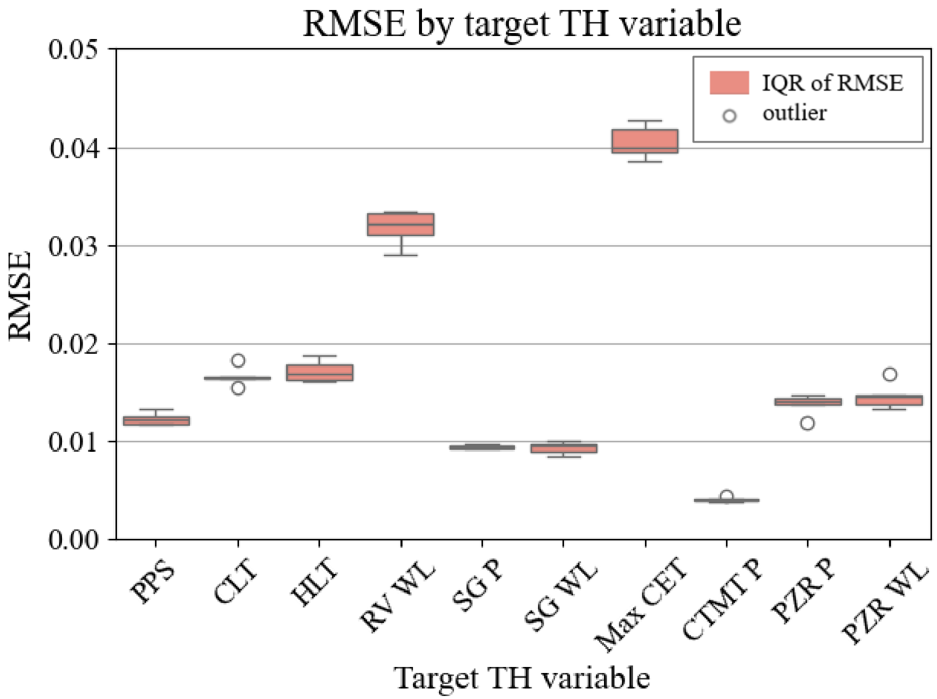


Fig. 10. MAE by the model's target TH variable. The IQR refers to the RMSE of models with various number of nodes (8, 16, 32, 64, 128)

Prediction accuracy

- ❑ Comparison of LSTM vs DA-RNN
  - ▪ LSTM model had smaller RMSE values on average.
  - ▪ Thus, the prediction accuracy did not improve by employing the attention mechanism.
  - ▪ However, the number of parameters in each model are different.
    - ➔ We cannot simply jump to a conclusion that LSTM model is superior.

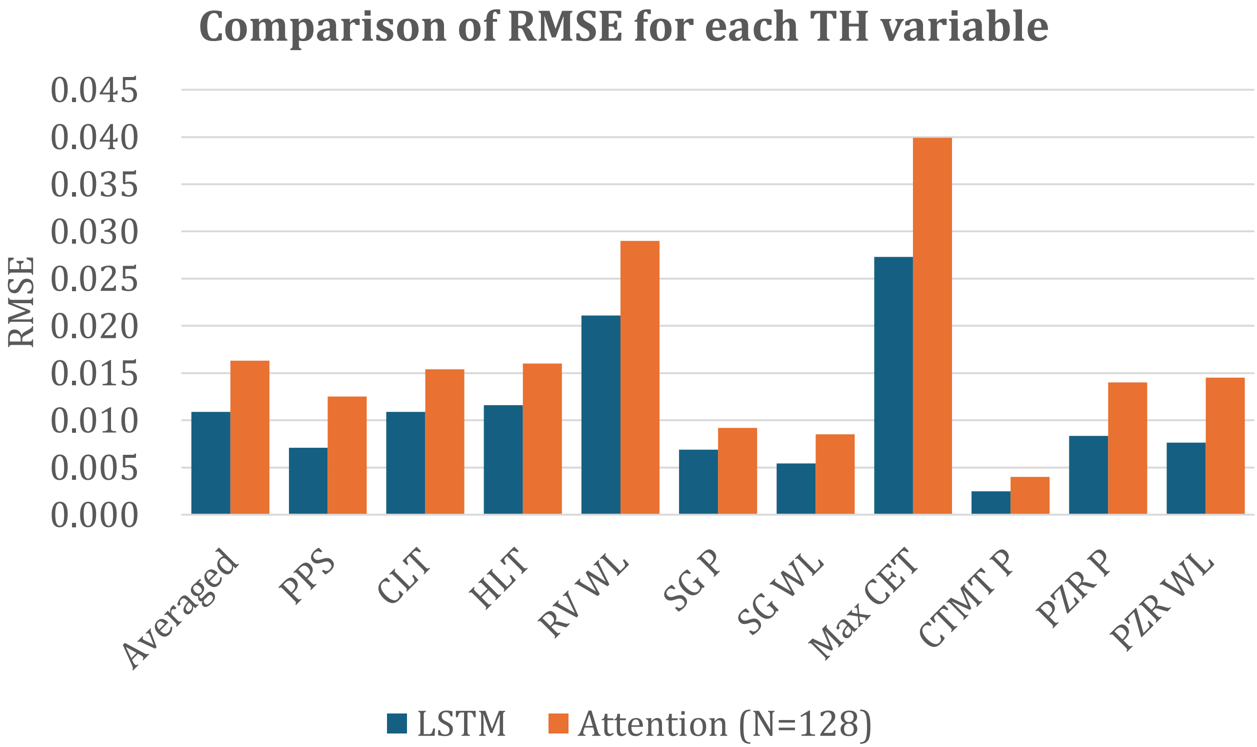
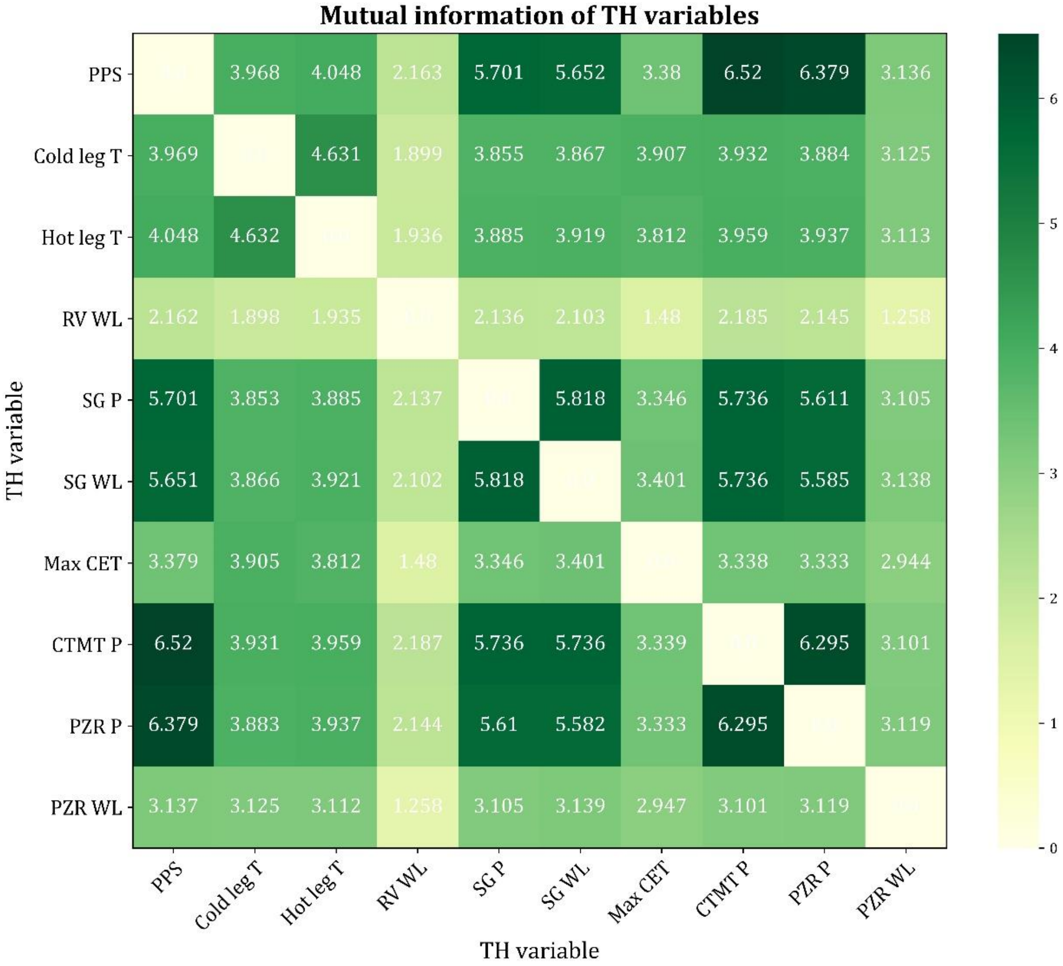


Fig. 11. RMSE of LSTM models and attention-based models

Model explainability

- ❑ Mutual Information
  - **Feature importance** represents the importance of an input parameter X for predicting a target variable Y.
  - **Mutual information** (MI): the amount of information obtained about one random variable by observing the other.
  - Comprehends the nonlinear relationship within the data. (vs. Pearson, Spearman correlation)
  - The MI values are then *normalized* so that their sum equals to one.

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



Mutual information of TH variables

	PPS	Cold leg T	Hot leg T	RV WL	SG P	SG WL	Max CET	CTMT P	PZR P	PZR WL
PPS	[illegible]	3.968	4.048	2.163	5.701	5.652	3.38	6.52	6.379	3.136
Cold leg T	3.969	[illegible]	4.631	1.899	3.855	3.867	3.907	3.932	3.884	3.125
Hot leg T	4.048	4.632	[illegible]	1.936	3.885	3.919	3.812	3.959	3.937	3.113
RV WL	2.162	1.898	1.935	[illegible]	2.136	2.103	1.48	2.185	2.145	1.258
SG P	5.701	3.853	3.885	2.137	[illegible]	5.818	3.346	5.736	5.611	3.105
SG WL	5.651	3.866	3.921	2.102	5.818	[illegible]	3.401	5.736	5.585	3.138
Max CET	3.379	3.905	3.812	1.48	3.346	3.401	[illegible]	3.338	3.333	2.944
CTMT P	6.52	3.931	3.959	2.187	5.736	5.736	3.339	[illegible]	6.295	3.101
PZR P	6.379	3.883	3.937	2.144	5.61	5.582	3.333	6.295	[illegible]	3.119
PZR WL	3.137	3.125	3.112	1.258	3.105	3.139	2.947	3.101	3.119	[illegible]

Fig. 12. Heatmap of the mutual information

Model explainability

- ❑ Mutual Information vs Attention weights
  - ▪ Example: **Cold leg temperature**
    - PPS, HLT, and Max CET seems to have relatively high importance in its prediction.
    - Explanation 1) Cold leg, core exit, and hot leg all constitute the primary flow together.
    - Explanation 2) The temperatures of the primary coolant are thermo-physically correlated with its pressure (PPS).

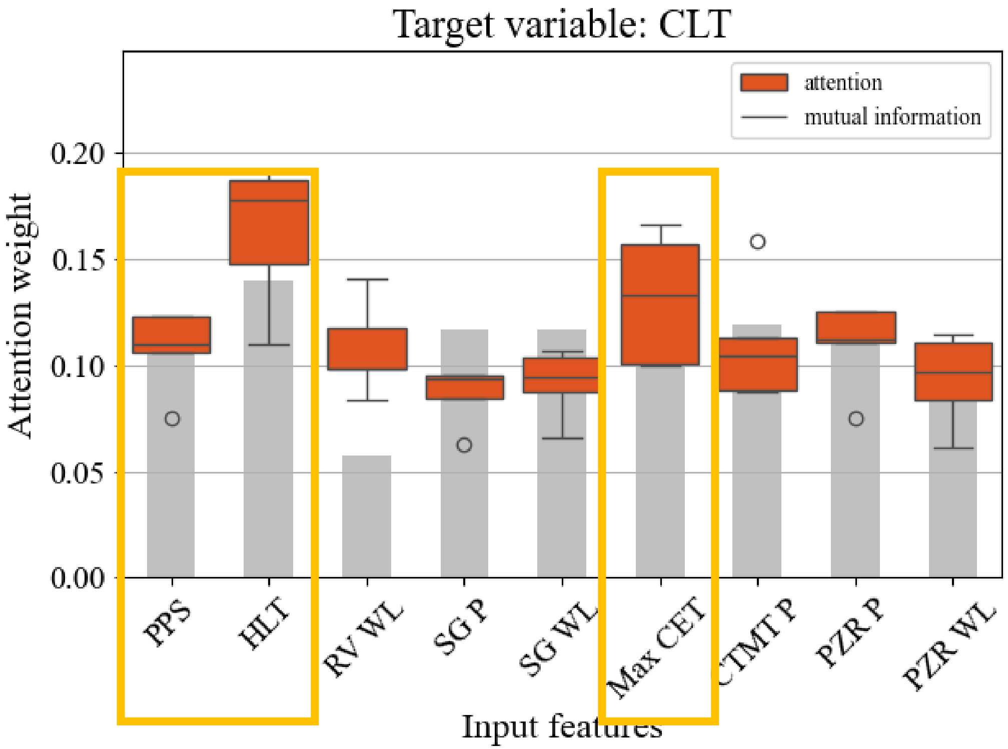


Fig. 13. Boxplot of attention weights for predicting cold leg temperature, compared to the mutual information (gray bar)

Model explainability

❑ Mutual Information vs Attention weights

- Measure the **cosine similarity** (S_C) between the mutual information matrix (MI) obtained from the MAAP data and the attention weight matrix (Att) obtained through the model training.

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

- High cosine similarity: 0.77 ~ 0.98

  ➔ *The proposed model learns the feature importance of the training data and embodies it as a form of attention weight.*

- RV WL, Max CET: worst prediction performance, but the cosine similarity > 0.9.

  ➔ *even if the attention weight of the model is well explained in phenomenological terms, the prediction accuracy of the model may not improve.*

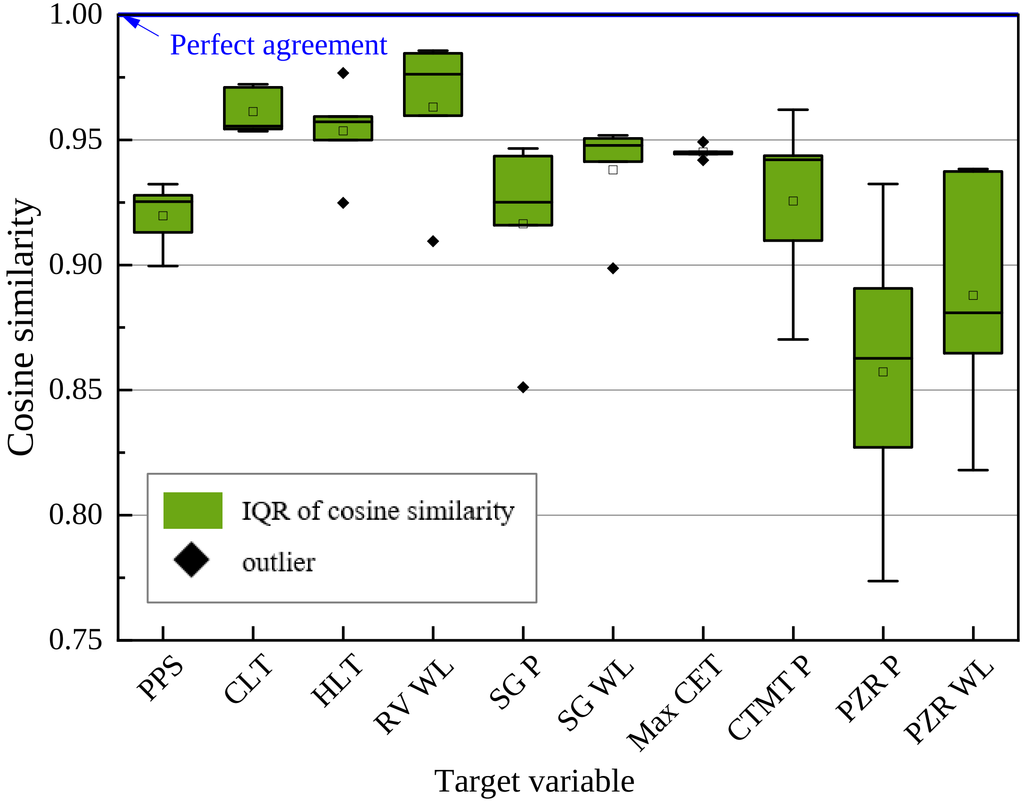


Fig. 14. Cosine similarity between attention weights and mutual information. The IQR refer to the range of cosine similarity of models with various number of nodes (8, 16, 32, 64, 128).

05

Conclusions and Further Works

Summary

1 Feasibility of the XAI model as an AMST predictor

- The XAI concept is expected to serve as a lubricant in applying AI models to Accident Management and Support Tool (AMST) development.
- Due to the ‘accuracy vs interpretability trade-off’ of deep learning models, developing an accurate XAI model is especially important.

2 Explainability of the DA-RNN model

- By learning the attention weight during the training process, the model searches for the importance of each input feature.
- The attention weights show a similar distribution with the mutual information of the original MAAP dataset.

3 Comparison between a black-box vs XAI model

- 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.
- Thus, it is possible to develop an AI-based AMST predictor model with both explainability and high accuracy.

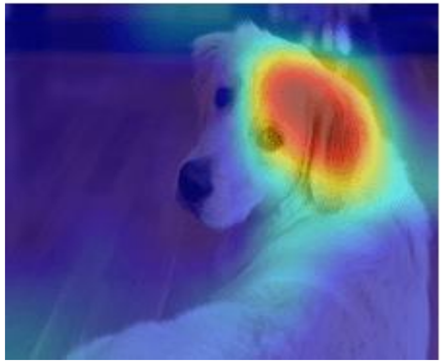
Limitations and Further Works

1. 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.

2. Investigate other index for representing the explainability of the models.

3. The potential of applying other XAI techniques to predict severe accidents will be assessed. (e.g., Grad-CAM, SHAP)

Grad-CAM



SHAP

Q & A

Appendix: Accident dataset

❑ Example of accident dataset

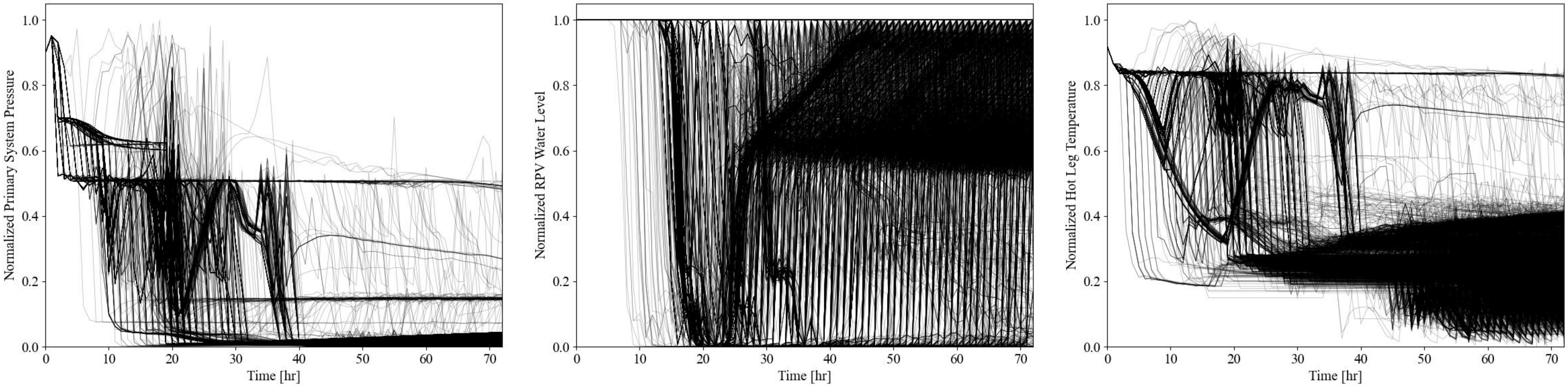
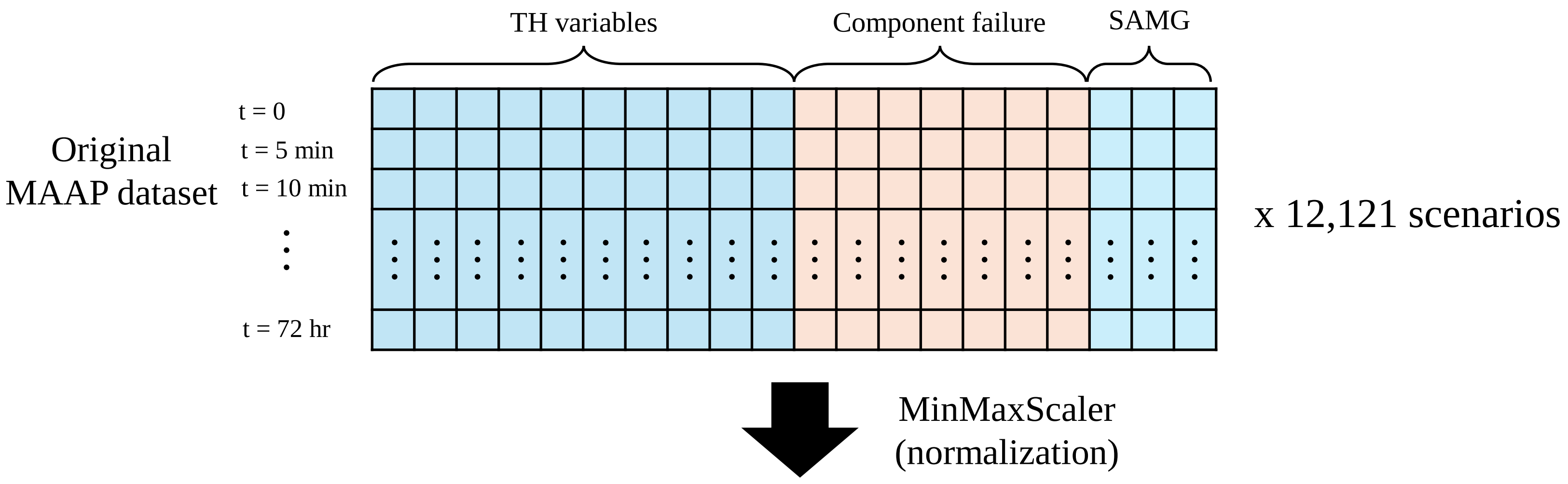


Fig. 27. Progression of TH variables for 72 hours in the produced accident scenarios.
Primary system pressure, reactor vessel water level, hot leg temperature (from left to right)

Appendix: Post-processing of MAAP dataset

❑ Example of accident dataset



Process the time series datasets into time steps of $\Delta t = 60$ min

Fig. 28. Post-processing of MAAP-generated datasets

Forecasting mechanism

❑ Forecasting mechanism

- The model forecasts the 72-hour accident scenario using the 'rolling window forecasting' method.
- Using the plant's states at the previous five time steps, the model predicts the plant's state at the next time step.
- This calculation is repeated 72 times → completes a time series of $t = 0$ to $t = 72$ hr.

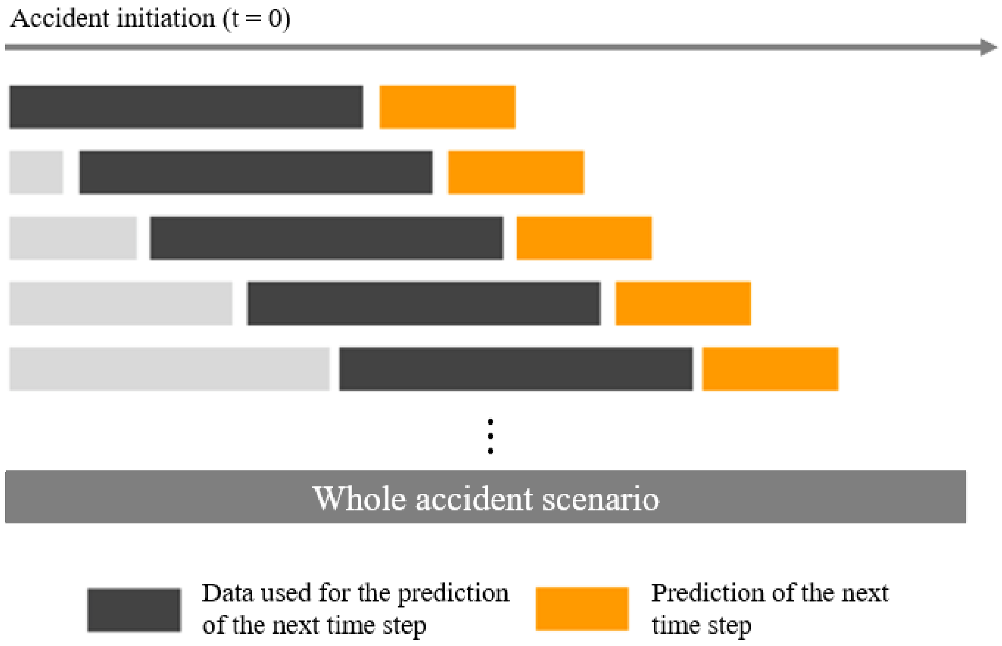

Accident initiation (t = 0)
Whole accident scenario
Data used for the prediction of the next time step
Prediction of the next time step

Fig. 12. Schematic of rolling window method in forecasting the accident scenario.

Training method

❑ Training method

- Divide the accident datasets into train (70%), validation (20%), test (10%) datasets.
- Criteria for stop training: validation loss does not decrease for 50 epochs

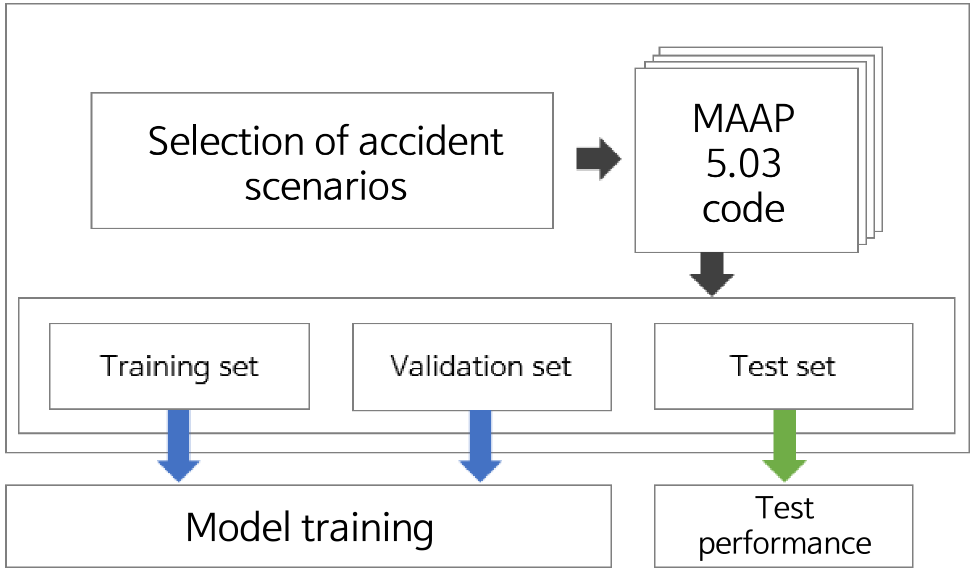


Fig. 10. Dividing MAAP datasets into train, validation, test datasets

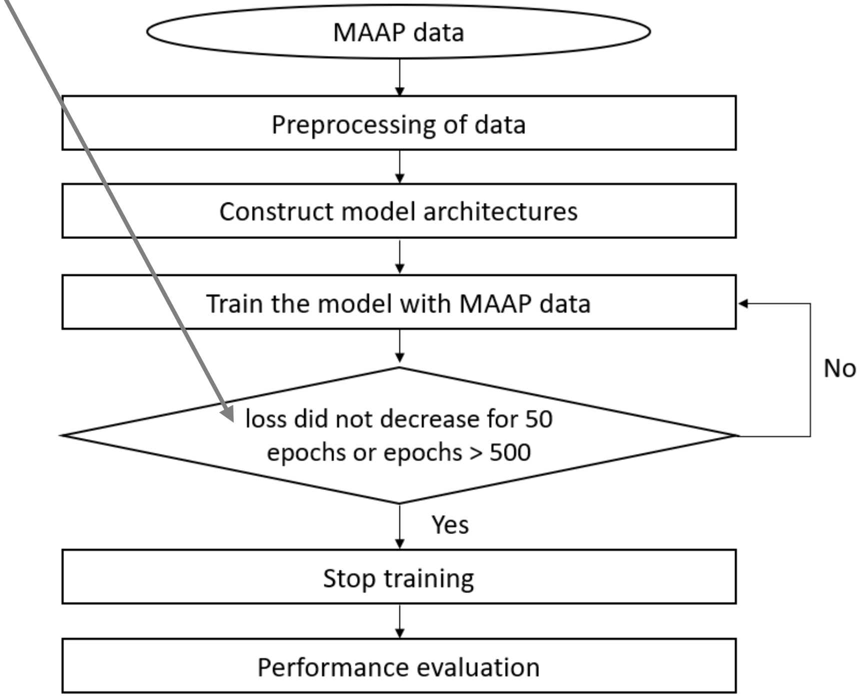

MAAP data
Preprocessing of data
Construct model architectures
Train the model with MAAP data
loss did not decrease for 50 epochs or epochs > 500
No
Yes
Stop training
Performance evaluation

Fig. 11. Training procedure

References

EXPLAINABLE AI TECHNIQUES

[1] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.

[2] 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.

[3] A. Vaswani et al., Attention is all you need, Advances in Neural Information Processing Systems 30, 2017.

[4] S. Wiegreffe, Y. Pinter, Attention is not not explanation, Conference on Empirical Methods in Natural Language Processing, 2019.

ACCIDENT MANAGEMENT SUPPORT TOOLS

[5] M. Saghafi, M. B. Ghofrani, Accident management support tools in nuclear power plants: A post-Fukushima review, Progress in Nuclear Energy 92, 2016.

[6] 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.

[7] 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.

PREVIOUS STUDIES

[8] 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.

[9] 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