Inferring Severe Accident Scenarios in Nuclear Power Plants with Reinforcement Learning with (RL) and Supervised Learning (SL) Approaches: Part 3 Sensitivity of RL to SL

Semin Joo, Seok Ho Song, Yeonha Lee, Jeong Ik Lee[a]*, Sung Joong Kim[b]
[a]*Dept. Nuclear & Quantum Eng., KAIST, 291, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea*
[b]*Dept. Nuclear Eng., Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea*
**Corresponding author: jeongiklee@kaist.ac.kr*

1. Introduction

Since the Fukushima-Daiichi accident, it has been recognized that beyond-design-basis-accidents (BDBAs) need more attention. BDBA is defined as an accident that has consequences beyond the design limit of a nuclear power plant (NPP). If the BDBA progresses until the core is damaged or melted, it is called a ‘severe accident’.

Severe accidents are non-linear and chaotic in nature, making it more challenging to predict their progression. The International Atomic Energy Agency (IAEA) recommends that the prediction and assessment of severe accidents be performed by a combination of Deterministic Safety Assessment (DSA) and Probabilistic Safety Assessment (PSA), as it provides insights into the progression of severe accidents and containment performance [1]. However, this method requires big computational resources and relies on highly conservative assumptions of severe accident scenarios due to large uncertainties. Hence, there is a need to develop a new approach that requires less computational resources and can incorporate uncertainties more easily.

From this background, an artificial neural network-based method is developed that predicts the progression of a severe accident in an accelerated manner. Two machine learning techniques are utilized to predict the progression of severe accident scenarios that can be generated by altering the total loss of component cooling water (TLOCCW) scenario. First, supervised learning is used to predict seven important thermal-hydraulic (TH) variables during a 72-hr accident scenario [2]. This is presented in the Part 1 companion paper. Using this supervised learning model as a surrogate model, a reinforcement learning (RL) that predicts an accident scenario that induces the most severe accident scenario has been developed [3]. That is, the RL agent is trained to choose a component failure time that accelerates the reactor pressure vessel’s (RPV) failure. This is presented in the Part 2 companion paper.

Three surrogate models have been developed, and the performances of the RL agent will be investigated when three different surrogate models are coupled to the RL environment. The main goal of this study is to elucidate the effect of the performance of a surrogate model on the performance of the RL agent. The detailed methodology will be discussed in the following sections.

2. Methodology

Fig. 1 summarizes the process of developing an RL that predicts the component failure time that accelerates the progression of an accident. The details of each process will be discussed in the subsections.

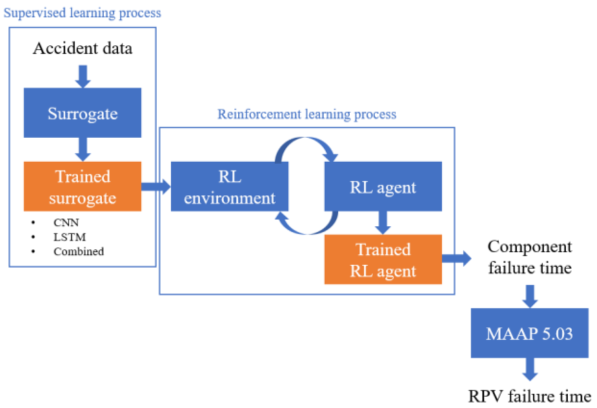

Fig. 1. Process of developing a reinforcement learning that can generate accident scenario with the most severe consequence.

2.1 Selection of Accident Scenario

Among various severe accident scenarios, the Total Loss of Component Cooling Water (TLOCCW) accident has been selected to demonstrate the performance of the proposed methodology. As severe accidents are triggered by multiple failures of components, the list of safety components that could cause a LOCCW needs to be identified. In this study, (1) Reactor Coolant Pump (RCP), (2) 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, and (7) Charging Pump (CHP) have been chosen. Assuming that these seven components can fail with a uniform probability of 1/2 within 72 hours accident progression time, 10,679 accident scenarios were generated.

Using these scenarios, Modular Accident Analysis Program (MAAP) 5.03 was used to simulate the progression of each accident scenario for 72 hours. It predicts the change in the selected TH variables (e.g., primary system pressure, cold leg temperature, core exit temperature). These outcomes are then used to train the surrogate models.

2.2 Surrogate Models

A surrogate model is a special case of supervised machine learning to predict an outcome using a data-driven, bottom-up approach. It simulates the behavior of a complex system in a time-efficient manner. The surrogate model is coupled to the RL environment to predict the TH variables of an NPP at the next time step.

It is known that the convolutional neural network (CNN) and long-term short memory (LSTM) models are specialized in predicting time-series variables. As the TH variables of an NPP during an accident scenario are time-series variables, these two models are considered appropriate for constructing the surrogate model. Thus, three neural network models have been developed: convolutional neural network (CNN), long short-term memory (LSTM) network, and a combined network composed of one CNN layer and two LSTM layers. The schematic of the combined network is described in Fig. 2.

The accident datasets generated from the MAAP code are fed into each surrogate model. Each dataset, or episode, consists of 73 time steps – 0 to 72 hours in 1-hour intervals. The surrogate models are trained to predict the TH variables at the next time step $(t+1)$, using the TH variables and mitigation strategies at the three previous time steps $(t-2, t-1, t)$.

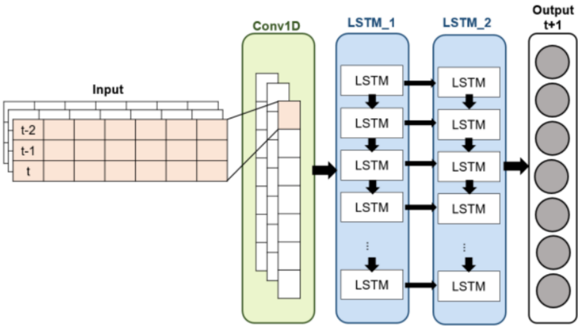

Fig. 2. Artificial neural network structure of the combined network model. It consists of a one-dimensional CNN layer and two LSTM layers [4].

The trained surrogate models interact with the RL environment to provide responses of the system to the RL agent after an action is taken by RL. Most importantly, it is expected that the performance of the RL model highly depends on the accuracy of the surrogate model. Thus, the accuracy of each surrogate model is checked and discussed together with the RL's performance.

2.3 Reinforcement Learning

In this study, an RL model is trained to choose the failure time of seven safety important components that maximize the damage to the NPP (i.e. early failure of RPV). This approach could be meaningful to developing prevention or mitigation measures in response to the worst-case scenario.

The RL is optimized by clipped proximal policy optimization (PPO). This method is advantageous because rapid convergence is possible using a clip function. When the RL agent takes a certain action (in this case, choosing which component will fail or not), the environment returns a set of reward and states to the agent. The reward system is described in Eq. (1). If the agent's action increases the pressure of the primary system (P_1), a reward is given proportional to the amount of increase in pressure, ΔP_1, and the remaining time, $72-t$. If not, there is no reward. This reward system was established based on the logic that over-pressurization of the primary system may cause a breach in the RPV. The RL agent is trained to pick an action that maximizes the reward, meaning that the agent will select the component failure time that increases P_1. It is noted that all variables handled in the environment have been normalized to have a value between unity and zero, thus the reward is non-dimensional.

$$reward = \begin{cases} \Delta P_1 \cdot (72-t), & \Delta P_1 \geq 0 \\ 0, & \Delta P_1 < 0 \end{cases} \quad (1)$$

The RL environment is coupled to three different surrogate models that were discussed in the previous subsection. All three RL agents are trained by datasets from 1,000 randomly chosen episodes (i.e. scenario). As one episode ends, the state and reward are reset. After the training is done, the RL agents are tested 1,000 times to select the component failure time. From the test results, the most frequently chosen component failure time sets are extracted. Based on the component failure time sets, the MAAP 5.03 code is used to validate the prediction from the surrogate model regarding the time of RPV failure. This approach is expected to provide insight for how each safety component failure time affects the RPV failure time.

3. Results and Discussion

3.1 Performance of Surrogate Models

Table I summarizes the accuracy of the three surrogate models. The mean absolute error (MAE) is used as a performance indicator, defined in Eq. (2). MAE refers to the mean absolute difference between the TH variables predicted by each surrogate model $(y_{pred,i})$ and those predicted by MAAP 5.03 code $(y_{MAAP,i})$. The MAEs of both valid datasets and test datasets are organized in Table I. It was observed that the most satisfying performance is demonstrated with the combined network model and the least performing model is the LSTM model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{pred,i} - y_{MAAP,i}\right| \quad (2)$$

Table I: Mean absolute errors of each surrogate model

	CNN	LSTM	Combined

Valid	0.01013	0.01425	0.00822
Test	0.01029	0.01444	0.00843

3.2 Performance of RL Agents

The trained RL agents were tested 1,000 times to select the component failure times. Fig. 3 summarizes the most frequently chosen failure times of the seven components selected by three RL agents. It is observed that the LSTM-based RL agent tends to select a significantly delayed failure time for HPI pump compared to other cases.

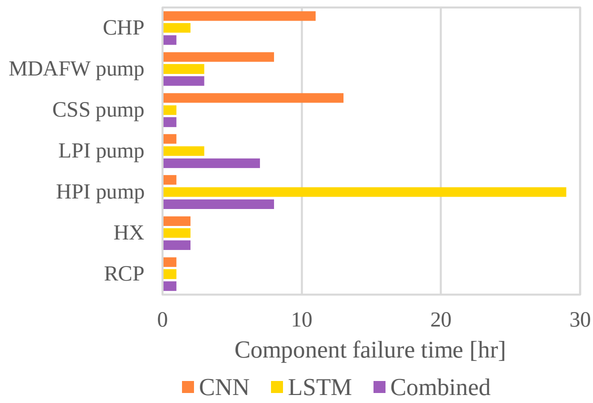

Fig. 3. Component failure times that were the most frequently chosen by three different RL agents.

Then, MAAP 5.03 code was used to validate the RPV failure time from the surrogate model prediction based on these component failure time sets (i.e., scenario). In Fig. 4, the RPV failure times and the performance of each surrogate model (MAE values in Table I) are plotted together. The component failure time sets chosen by CNN, LSTM, and the combined network model have predicted RPV failure times of 19.23, 38.06, and 18.36 hours, respectively. Although the component failure times selected by the CNN-based and CNN-LSTM combined network-based RL agents differed from each other, the RPV failure times validated by MAAP 5.03 code were comparable. It is clearly shown in Fig. 3 that the failure times of individual safety components in the two models are different. However, the times when all the safety components responsible for cooling the reactor core (e.g., HPI, LPI, MDAFW) fail appeared to be similar. This explains the similar RPV failure times between the CNN-based and the combined network-based RL models.

Also, the RPV failure time is especially delayed in the LSTM model, possibly because the LSTM-based RL agent tends to choose delayed HPI pump failure time.

The reason behind this is due to the nature of the LSTM model. Although LSTM is known to be specialized in sequence modeling, the vanishing gradient problem is deeply rooted in it. Since the LSTM performs consecutive matrix multiplications, the network cannot be trained sufficiently if the amount of update or gradient is small [5]. As the LSTM surrogate returns incorrect states to the RL environment, the agent receives erroneous rewards, thus hindering the learning process. Judging from the performance results, it seems that the performance of RL is improved by combining the CNN layer with the LSTM layer.

Most importantly, it was observed that if the MAE of the surrogate model is smaller, the RL agent develops a scenario that can produce earlier RPV failure consequences. Thus, it is concluded that for the RL agent to be trained in the direction of the original intention, the surrogate model interacting with the RL should have high performance.

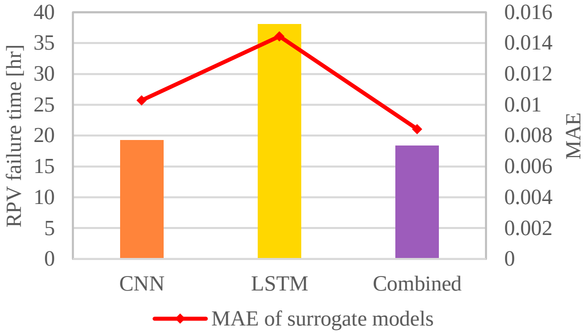

Fig. 4. RPV failure time and MAE performance for tested SL models.

3.3 Uncertainty of RL models

Finally, the distribution of the component failure times selected by the RL agent is analyzed. The standard deviation of the component failure time is evaluated, as it is an indirect measure of the uncertainty of each RL model. For example, if the variance of the failure time of a certain component is large, it implies that the RL does not recognize that the corresponding component failure time significantly contributes to advancing the RPV failure time. On the other hand, if a component plays a significant role in accelerating the RPV failure time, the component failure time selected by the RL will be concentrated in a specific time period. Furthermore, if the variance of the failure times of the safety components with a strong correlation to RPV failure time (e.g., HPI, LPI, MDAFW) is small, it means that the RL model is well-trained as originally intended.

Fig. 5 shows the standard deviation of the distribution of component failure times selected by the three RL agents. The models that showed a small variance in failure time of HPI, LPI, and MDAFW are the CNN-based and the combined network-based RL models. On the contrary, in the case of the LSTM-based RL model, the standard deviations were larger on average compared to the other models. That is, it indirectly implies that the uncertainty of the RL model based on LSTM is large and, not only do the CNN-based and the combined network-based RL models perform better but also they have smaller uncertainties for making decision to minimize the RPV failure time.

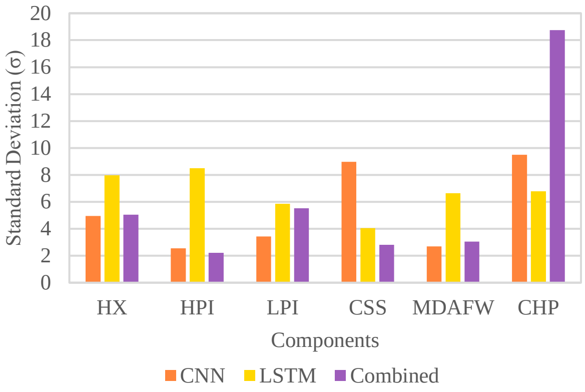

Fig. 5. Standard deviation of the component failure time distribution

4. Conclusions and Further Works

An artificial intelligence-based system was designed to predict severe accident scenarios in a pressurized water reactor (PWR). Using the MAAP 5.03 code, datasets used for surrogate model training were generated. The surrogate model was developed to predict the TH variables at the next time step in a 72-hour-long TLOCCW accident. Three surrogate models were developed, all of which are constructed with different types of layers: CNN, LSTM, and CNN+LSTM combined approach. Each surrogate model is connected to the RL environment and the overall performance was tested. The RL agent interacts with the surrogate model and is trained to select the reactor component failure time that can induce the RPV failure earliest.

The main objective of this study is to investigate the effect of the performance of surrogate models on the performance of RL agents. As a result, it was confirmed that as the surrogate predicts the TH variables with better performance, the RL agent is more likely to take action that can induce RPV failure earlier.

Based on this observation, further studies can be proposed from various perspectives. As the results of this study emphasize the performance of the surrogate model, the surrogate models can be further improved for predicting the TH variables.

Also, a different RL reward system can be searched for. The reward system used in this study relies on a simple assumption that the increase in P_1 will cause the RPV to fail. However, as RPV failure is a complex and non-linear phenomenon, a such simple reward design may not be sufficient to construct the reward system.

Lastly, a more comprehensive systematic method to quantify various uncertainties associated with an RL model is required. These uncertainties are as follows:

i. Uncertainties associated with MAAP code and input: The thermal-hydraulic variables predicted by the MAAP 5.03 code bear uncertainties, which stem from the uncertainties of the MAAP program and input.

ii. Uncertainties associated with the surrogate model: One is the epistemic uncertainty, which roots in the lack of data. The other is related to the error of the constructed neural network, which can be quantified by the dynamic time warping between the training set and the predicted values.

iii. Uncertainties associated with the RL model: The RL environment has been constructed with the surrogate model, meaning that the RL model essentially bears all the uncertainties that have been discussed above. In addition, there are aleatoric uncertainties embedded in the RL model. One intuitive measure of quantifying this uncertainty would be to calculate the variance of the learned distribution of the component failure times.

Acknowledgment

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

REFERENCES

[1] International Atomic Energy Agency, Safety Assessment for Facilities and Activities, General Safety Requirements Part 4.

[2] 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.

[3] S. H. Song, A study of Reinforcement Learning Implementation and Building Reinforcement Learning Environment Using Supervised Learning for Generating Nuclear Power Plant Accidents Scenario, Master's thesis, Korea Advanced Institute of Science and Technology, 2022.

[4] Y. Lee, K. Song, and J. Lee, Time-series forecasting of thermal-hydraulic variables during severe accidents in nuclear power plants under various scenarios using machine learning technique, International Congress on Advances in Nuclear Power Plant (ICAPP-2023), April.23-27, 2023, Gyeongju, Korea.

[5] S. Bai, J. Z. Kolter, and V. Koltun, An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, International Conference on Learning Representations (ICLR), 2018.