Key Considerations for Improving the Reliability of Deep Learning Models in Instrumentation and Controls

Jaekwan Park [a*], SeoRyong Koo [a]
[a] *Korea Atomic Energy Research Institute, 111, Daedeok-daero 989 beon-gil, Yuseong-gu, Daejeon, Republic of Korea*
[*] *Corresponding author: jkpark183@kaeri.re.kr*

Keywords : verification, artificial intelligence, deep learning model, software development process

1. Introduction

Artificial intelligence technology has made incredible strides in recent years and is expected to be applied to many industries in the coming decades. However, as vehicle accidents using self-driving technology have been reported, there are growing concerns about their trustworthiness. Various studies are being conducted to apply these AI technologies to the field of nuclear power plant instrumentation and control, but there is no system for checking the appropriateness of the development process and testing the performance of the final implementation. Therefore, we need to discuss systematic ways to increase the reliability of this new software. This paper proposes key considerations for software reliability of deep learning models from two perspectives: development process and testing. Referring to documents software V&V [1] and machine learning testing literature [2], this study identified key verification activities in the development process and key testing items for the final deep learning models.

2. Key Considerations for AI Software Reliability

2.1 Key Verification Activities

The existing system of reviewing the output of the software development process by a V&V team independent of the development team can also be applied to AI software. For AI software, the V&V team needs to review the essential items shown in Figure 1.

First, the V&V team reviews the data, design, implementation, and testing plans for the deep learning model development process. Additionally, plans for interactions between automated systems and human operators should be accompanied by additional design plans to comply with the guidance of NUREG-0700 [3], "Section 7. Automation system." Next, the V&V team reviews the *acceptance criteria* for the deep learning models. For example, the performance of deep learning models used in classification problems is commonly measured by several performance metrics such as precision, recall, and F1 score.

Data labeling is a very essential step in the training process of deep learning models in the supervised learning. This process involves adding tags or labels to raw data to specify the classes to which the data belong. However, as most datasets collected from nuclear power plants comprise raw data without labels, developers must often implement a labeling method to assign labels to each data point and thereby complete the dataset. Inaccurately labeled datasets lead to deep learning models that make incorrect decisions. Therefore, the labeling criteria, methods, and tools employed must be agreed upon by both developers and plant operators, and the V&V team must perform an adequate evaluation of these items.

As *datasets* are essential for the development of deep-learning models, a plan must be established for obtaining data from various sources including field operations, simulations, and analyses. If a dataset is insufficient or biased, the development team may be asked to supplement the dataset plan.

If unsuitable hyperparameter values are selected, the model may suffer from overfitting or underfitting. To address this issue, appropriate values for the hyperparameters must be determined through extensive experimentation and the resulting decisions documented. The process of determining these values, known as *model specification*, considers various factors related to the model architecture and configuration.

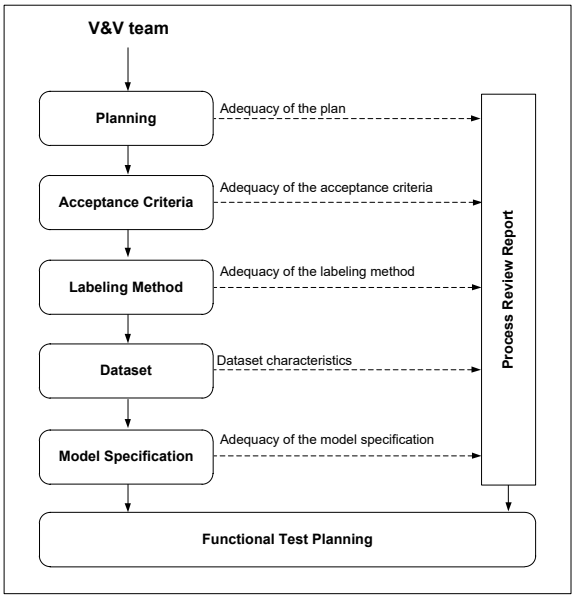

Fig. 1. Key verification activities for deep learning models.

2.2 Key Testing Items

Test cases should be prepared independently of the training dataset used by the development team so that the model can be evaluated using unseen data. If test cases are prepared by the team that created the training dataset, there is a risk that the model may perform well simply because similar data was used during training, which could misrepresent the model's true performance on new, unseen data. An independent set of test cases that does not reflect the developer's intentions can provide a more reliable assessment of model performance.

The evaluation of a deep learning model starts with *performance testing*, which is the most important evaluation of the model's capabilities. Accordingly, test cases are prepared to comprehensively cover the target operation range. Test cases for evaluating data labels are also considered to provide detailed performance results based on random, target range, and data labels. The V&V team then evaluates whether the model meets acceptance criteria in terms of average accuracy. The performance tests also evaluate the response time in terms of the time required for the deep learning model to process the given input and provide the output. According to NUREG-0700 [3], all guidance related to the main control room and information provided to operators should be updated at least once per second. If the response time does not meet this guideline, it can be improved by simplifying the model structure or reducing excessively time-consuming tasks within the model.

Deep learning models should not be more complex than necessary to complete the intended task, or they may not work effectively on real-world instrument signals that contain uncertainty. Model *generalization test* is used to evaluate how well a deep learning model fits the training data set. Poor model generalization is often the result of overfitting, which occurs when the model is too complex for the size of the dataset. An overfitted model is highly customized to the training dataset and may not perform well on new or unseen data.

In the context of AI technology, *robustness* refers to the ability of a deep learning model to produce accurate results even in the presence of invalid or noisy inputs. This is particularly important in nuclear power plant I&C systems, which must be able to operate effectively even in the presence of outliers due to aging, failure, or signal noise. To ensure the robustness of deep learning model, the V&V team should perform tests that cover the full range of normal and abnormal instrument signals according to the IEEE 1008 [4] unit test guidelines approved by RG 1.171 [5]. This ensures that the deep learning model can withstand abnormal changes in instrument signals while still producing accurate results.

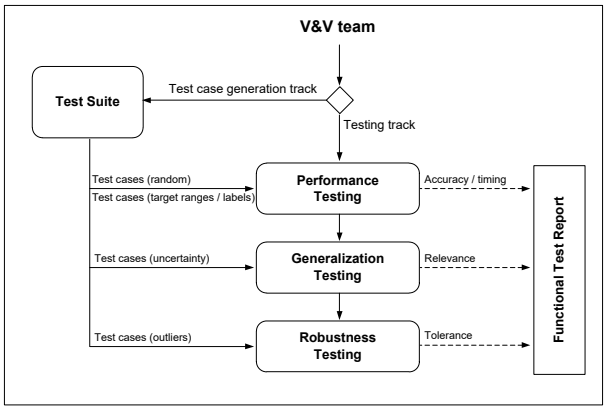

Fig. 2. Key testing items for deep learning models.

3. Conclusions

In this paper, we propose key considerations for applying deep learning models in nuclear power plant instrumentation and control to improve reliability. First, we propose verification activities to ensure the appropriateness of the development process. We propose three types of testing: performance, generalization, and robustness testing. These allow the V&V team to verify the correct behavior of the deep learning model on normal and abnormal inputs.

Acknowledgements
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2022-00144150) and the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry, & Energy (MOTIE) of the Republic of Korea (No. 20224B10100130).

REFERENCES

[1] IEEE Std 1012, IEEE Standard for Software Verification and Validation, Institute of Electrical and Electronics Engineers (IEEE), 2004.
[2] Zhang, J.M., Harman, M., Ma, L., and Liu, Y., Machine learning testing: Survey, landscapes and horizons, IEEE Transaction of Software Engineering, Vol.48, p.1–36, 2022.
[3] NUREG-0700, Human–System Interface Design Review Guidelines, United States Nuclear Regulatory Commission (US NRC), 2020.
[4] IEEE Std 1008, IEEE Standard for Software Unit Testing, Institute of Electrical and Electronics Engineers (IEEE), 1987.
[5] RG 1.171, Software Unit testing for Digital Computer Software Used in Safety Systems of Nuclear Power Plants, 2013.