Lifecycle processes and checklists for AI data

Jang-Yeol Kim [a]*, Jong-Gyun Choi [a]
[a]*Advanced Instrumentation and Control Lab., Korea Atomic Energy Research, 111, Daedeok-daero, 989beon-gil, Yuseong-gu, Daejeon, 34057, Korea*
[*]*Corresponding author: jykim@kaeri.re.kr*

***Keywords : AI, Big Data, Life Cycle, Check Lists, Data Science**

1. Introduction

AI processed data is data used to train AI models such as machine learning and deep learning. This data is collected from original data and processed in various types. The purpose of AI processed data is to provide efficient and accurate data for learning of AI models. AI processing methods include refining, transforming, and labeling. AI processed data has a significant impact on the performance of AI models. If the processed data is not accurate and efficient, the AI model may learn incorrectly. In this paper proposed the basic and extended AI data lifecycle phase associated with AI processed data. It also provided related checklists. The proposed AI data lifecycle phase and checklist are expected to be helpful for AI applications in the nuclear industry.

2. AI Data Lifecycle

2.1. Basic Lifecycle of AI data

The basic process of AI data lifecycle is divided into data collection, data preprocessing, learning, data evaluation, and data monitoring as shown in Figure 1. First, the data collection phase checks the validity, sufficiency, and diversity of data. Second, data preprocessing phase checks for errors, imbalances, and anomalies in the data. Third, the data training phase checks for model performance, overfitting, and data bias. Fourth, data evaluation phase checks the performance, prediction, and efficiency of the model. Finally, data monitoring phase checks the performance, stability, and security of the model.

2.1. AI Data Lifecycle Extension Process

Based on the basic process of AI data lifecycle in Figure 1, this paper proposes an extended process of AI data lifecycle as shown in Figure 2.[1] This process is defined as a planning-preparation-analysis-development-expansion process similar to the existing traditional method, and then a new set of interconnected processes is created by reflecting and expanding the important characteristics of AI data. In the expansion AI data lifecycle process is set up so that it can be feedback during each lifecycle processes. It is also linked the interface to the application of the model and the establishment of the operation plan.

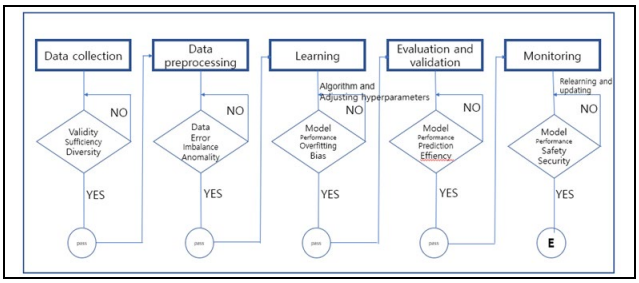

Figure 1. Basic process of AI data lifecycle

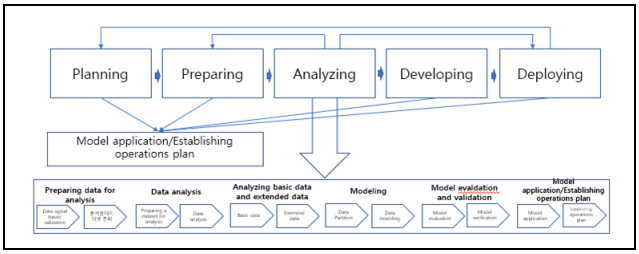

Figure 2. Expansion process of AI data lifecycle

3. Checklist of analysis steps of AI data lifecycle extension process

Since AI is a technology that learns and develops based on big data, data is the most important and serves as the basis for learning and development of models. Therefore, the quantity and quality of data have a great impact on the performance of AI models. In order to improve the performance of AI models, it is essential to collect various data and manage data quality. In order to manage data quality, it is necessary to thoroughly check the data analysis and data specifications in the requirement phase of the AI software development lifecycle. In this paper, we propose a checklist for the most important requirement phase of the AI data lifecycle expansion process as shown in Table 1 and Table 2 respectively [2], which are to check the suitability of data requirements on the suitability of design and implementation, the suitability of data duplication, and the suitability of data licenses in the requirement phase.

Table 1. Data Requirement Definition Adequacy Checklist

Requirement Definition Adequacy	**1.0 Have data processing requirements been appropriately managed?**
	1.1 Are the preparing of data processing targets (requirements) and acceptance criteria without omissions?
	1.2 Are the acceptance criteria (inspection methods, pass/fail criteria) specific and clear?

Table 2. Data Specification Checklist

Design and implementation data adequacy	**2.0 Do the data processing requirements and data contents of the data specification each other?**
	2.1 Is the processed data implemented according to the processing methods and procedures included in the requirements and acceptance criteria?
	2.2 Is the processed data implemented according to the data type, format, and columns (attributes) defined in the data specification?
Adequacy of data duplication check	**3.0 Is the processed data not duplicated with publicly available data?**
	3.1 Was the source data collected independently?
	3.2 Is the processed data not duplicated with public data provided by AI hubs homepages and public data portals?
Adequacy of Data licenses	**4.0 Do you have a valid license for the dataset you are processing?**
	4.1 Does the dataset use data that is already publicly available?

4. Conclusions

This paper proposed checklist and basic and extended AI data lifecycle process for AI models. By effectively utilizing the checklists and proposed AI data lifecycle process, you can prevent various problems that may occur during the development and deployment of AI models. It can also improve the performance of AI models. If the proposed checklists, basic and extended processes AI data lifecycle are fully utilized, it is expected to be a new turning point in the development of AI applications in the nuclear industry area.

REFERENCES

[1] Jang-Yeol Kim, Korean Institute of Communications and Information Sciences (KICS), “The Trends in Artificial Intelligence Development for Nuclear Applications based on IEC International Standards”, Energy and AI Safety Workshop, November 29 - November 30, 2023

[2] Data Voucher Business Management System, https://kdata.or.kr

[3] AI-Hub, https://aihub.or.kr

[4] Open Government Data Portal in Korea https://data.go.kr