Study on the Effectiveness of Time Series Data Augmentation for Anomaly Detection in Measurement and Control Systems

Han Gil Lee[a], Seung Hyeok Yang[a], Sang Beom Kang[a], Dae Young Lee[a*]
[a]*FNC Technology Co., Ltd., 13 Heungdeok 1-ro, 32F, Giheung-gu, Yongin-si, Gyeonggi-do, 16954, Korea*
[*]*Corresponding author: ldy242@fnctech.com*

***Keywords : Artificial Intelligence, Timeseries Data Augmentation, Anomaly Detection, Reliability, Accuracy**

1. Introduction

In recent years, along with the dazzling progress of AI (Artificial Intelligence) technology, significant efforts have been made to successfully integrate AI technology into the nuclear power industry. By leveraging the fault data obtained from the damage diagnosis testbed of the measurement and control system, successfully predicting the lifespan of components in the measurement and control system through artificial intelligence technology would greatly contribute to enhancing the safety of nuclear power plants by proactively addressing the replacement of equipment with a high likelihood of failure.

To achieve this, it is important to demonstrate the accuracy and reliability of artificial intelligence software. However, the evaluation methods for the accuracy and reliability of artificial intelligence technology applied in the nuclear power industry are still inadequate. Furthermore, in the actual environment of the measurement and control system, various time series data containing noise are generated, and there are practical limitations to securing all of this data [1].

In this study, as a solution to this issue, we deliberated on improving the accuracy and reliability of artificial intelligence technology from the perspective of time series data design. We constructed suitable datasets by applying time series data augmentation techniques to the fault data obtained from the testbed, based on the standard of balanced time series datasets presented by TTA (Telecommunications Technology Association) [2]. Subsequently, we examined whether the balanced datasets could improve the performance through the application of a deep learning anomaly detection model.

2. Background

The use of deep learning in time series data analysis is valuable for recognizing and classifying complex patterns in data, extracting useful information, as well as detecting and identifying abnormal patterns to assist in real-time decision-making. Among these the task of time series anomaly detection involves detecting all types of anomalies that deviate from normal patterns, making it a challenging field for artificial intelligence due to the multitude of patterns present in the data. Therefore, our research focused on improving the reliability and accuracy of the applied deep learning models from the perspective of dataset design.

The reliability and accuracy of artificial intelligence software are crucial not only for the performance of the applied models but also for the balance and suitability of the dataset. If specific areas in the utilized time series dataset are overrepresented, it can lead to sampling biases, sampling errors, and data overfitting, all of which can significantly reduce the reliability and accuracy of the software. Therefore, it is essential to secure unbiased noise-balanced data to assess the level of trust.

Accordingly, TTA has defined guidelines for designing balanced datasets that reflect various characteristics of time series data [2]. We explored suitable balanced datasets based on these standards. The following are the review criteria for noise balance in time-series datasets presented by TTA standards.

Table I. Considerations for Noise Balance in Time Series Data

Balance List	Equation for Noise	Detail
trend noise	$\boldsymbol{Y'} = (aX + b) + \boldsymbol{Y}$	- $\boldsymbol{Y}$: Original time series data - a : $\pm$ gradient - b : bias
seasonal noise	$\boldsymbol{Y'} = a \times sin(bX + c) + \boldsymbol{Y}$	- a: Amplify/attenuate original time series data - b: Cycle/c: x-axis movement
outlier noise	$if\ rand(X) > a,$ $then\ \boldsymbol{Y'} = \boldsymbol{Y} + randbetween(b, c)$ $else\ \boldsymbol{Y'} = \boldsymbol{Y}$	- a : Selector for applying outlier values - b, c : Random outlier values min/max range
sudden change noise	$if\ a \leq (X) \leq b,$ $then\ \boldsymbol{Y'} = \boldsymbol{Y} + c$ $else\ \boldsymbol{Y'} = \boldsymbol{Y}$	- $\boldsymbol{Y}$: Original time series data - a, b: Abrupt Changes, starting and ending points to reflect - c: Sudden change

		(positive, negative)
distributed noise	$\boldsymbol{Y'} = \boldsymbol{Y} + a \times random()$	$-random$: Random value between 0 and 1 Uniform Distribution $-a$: Noise intensity amplification/attenuation

To assess whether artificial intelligence software that uses Time series data operates reliably, the above items must be checked and reflected in the data [2]. Each noise type in the list contributes to diversifying the time series data, enabling the model to learn various patterns. We applied time series data augmentation techniques to the existing data to incorporate noises listed in Table 1. And, we evaluated the impact of using this designed balanced dataset on anomaly detection performance for new data through experiments.

3. Experiment

In this section, we describe the Time series data augmentation techniques utilized in our study and present the test results of the anomaly detection model.

3.1 Time Series Data Augmentation Techniques for Balanced Dataset Design

Time series data typically exhibits temporal dependency, and depending on the task at hand, there is a tendency for it to be dependent on the passage of time. Due to these characteristics, appropriate data augmentation tailored to each transformed domain is necessary. Therefore, in this study, we applied Time series data augmentation specifically for the purpose of anomaly detection tasks.

Time series Data differs from image and text data in that it can be divided into time and frequency domains. The time domain represents the area containing information about time, and intuitive augmentation techniques are often used in this domain. The frequency domain, on the other hand, contains information about the frequency of the data. It consists of the Amplitude Spectrum, which represents the magnitude, and the Phase Spectrum, which represents the position along the time axis [3].

We explored techniques that allow us to inject noise into the original data without completely losing its characteristics. Consequently, we selected intuitive methods such as slicing, shuffling, and wrapping in the time domain to introduce noise to certain segments. Additionally, in the frequency domain, we opted to apply Gaussian noise across the entire range using Fourier transformation. The following are the types and application methods of augmentation techniques for each domain.

Table II. Augmentation Characteristics Based on the Domain of Time Series Data

DOMAIN	METHOD	APPLICATED
TIME	- Window slicing & Shuffle - Window warping	- Crop randomly a portion - Compress or expand a specific time range
FREQUENCY	- AAFT (Amplitude Adjusted Fourier Transform) - APP (Amplitude and Phase Perturbations)	- Apply Fourier transformation and shuffle, Inverse Fourier transformation - Apply Gaussian noise to some Amplitude & Phase values

We obtained balanced datasets by injecting noise into the original data using the techniques and application methods outlined in Table 2.

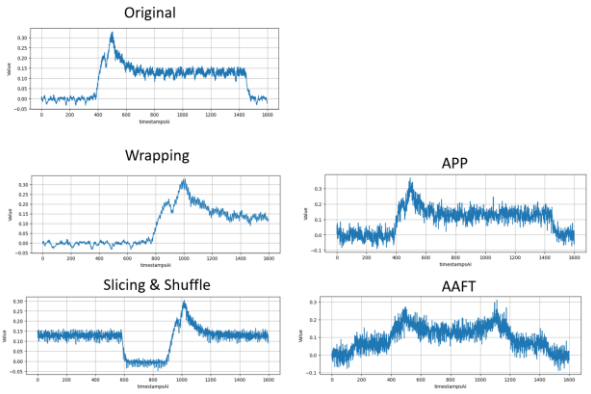


Fig. 1. Time series data augmentation results

Fig. 1, depicts the results of applying time series data augmentation techniques to the original data. We applied the time domain techniques described in Table 2, such as Wrapping, Slicing & Shuffle, as well as Frequency domain techniques like APP and AAFT. Comparing each of the new datasets to the original data, it's evident that they contain various types of noise.

3.2 Check anomaly detection results of time series balanced dataset

To conduct performance testing of the anomaly detection model on the balanced dataset, we selected a deep learning model called TadGan. TadGan utilizes GAN (Generative Adversarial Network), a generative deep learning model, to learn and restore patterns in the data. It predicts injected data and detects anomalies by identifying areas with significant errors, demonstrating

higher detection performance compared to traditional autoencoder-based models [4][5].

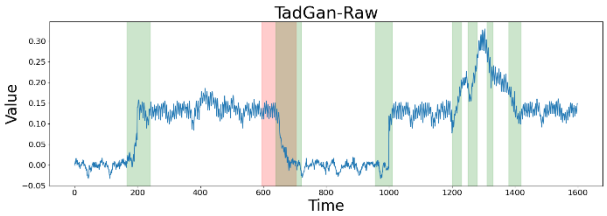
In Section 3.1, we designed three datasets using the generated data and applied them to train the TadGan model, resulting in the creation of pretrained models for each dataset. First, let's outline the composition of the training datasets for each model.

Table III. Training dataset for Pretrained Model

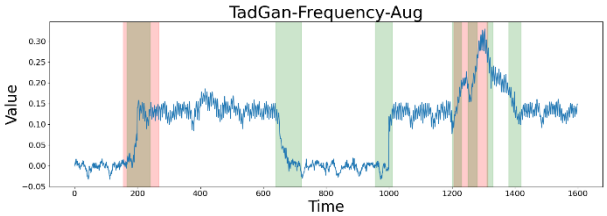
Pretrained Model	Dataset
Model 1	Original Data 1~3
Model 2	Original Data1, Frequency Aug(APP, AAFT)
Model 3	Original Data1, Time Aug(Slicing & Shuffle), Frequency Aug(APP)

The following are the anomaly detection results for the test data of each pretrained model.

Model 1:



Model 2:



Model 3:

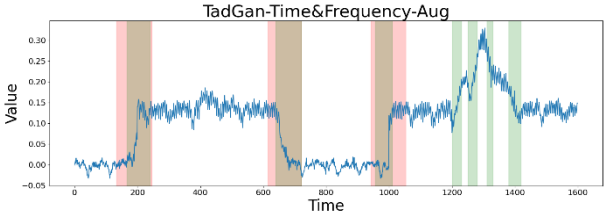


Fig. 2. Time series anomaly detection results by case

In the graphs for Models 1 to 3, the green regions represent the areas labeled as anomalies, while the red regions represent the anomalies detected by the model. The higher the overlap between the green and red regions, the higher the accuracy of the model can be considered. Each model has an equal amount of dataset, and the test data consists of data not used for training.

Table IV. Confusion Matrix score by Dataset Case

Dataset Case	Precision	Recall	F1
Model 1	0.589	0.195	0.293
Model 2	0.613	0.409	0.491
Model 3	**0.641**	**0.632**	**0.636**

Table IV. presents the Confusion Matrix Score (Precision, Recall, F1) for each model based on the anomaly detection results obtained from Fig. 2.

The metrics used quantify accuracy by utilizing actual and predicted data. The values are calculated based on the labeled data from Fig. 2, where the labeled values are considered as actual data, and the regions detected as anomalies by each model are considered as predicted data. Higher values for each metric indicate higher accuracy. As a result, Model 3, which includes augmented data from both Time and Frequency domains, achieved the highest Confusion Matrix Score. This indicates that the balanced time series dataset researched in this paper has an impact on enhancing the anomaly detection performance of artificial intelligence technology.

4. Conclusion

We conducted research to enhance the accuracy and reliability of artificial intelligence anomaly detection tasks using the measurement and control system testbed. To achieve this, we reviewed the time-series type balanced dataset defined as a standard by TTA and designed a balanced dataset using time-series data augmentation techniques. Subsequently, we applied deep learning anomaly detection models to the designed datasets and observed improvements in model performance. The data obtained from the measurement and control system damage diagnosis testbed has limitations in achieving an appropriate noise balance compared to real-world data. Therefore, it is expected that applying the balanced dataset design approach for time series data, as conducted in this study, would positively impact the accuracy and reliability of artificial intelligence anomaly detection tasks.

ACKNOWLEDGMENTS

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning(KETEP) and the Ministry of Trade, Industry & Energy(MOTIE) of the Republic of Korea (No.20224B10100120).

REFERENCES

[1] UNIST, Yong Kyung Oh, Simulation-based Anomaly Detection in Nuclear Reactors, Journal of the Korean Institute of Industrial Engineers, 47(2), 130-143, 10.7232/JKIIE.2021.47.2.130, 2021.
[2] TTA, A Method for Evaluating the Reliability of Artificial Intelligence Software Based on the Balance of the Validation

Dataset - Part 3: Design of Time Series Type Balanced Data, TTAK.KO-11.0280-Part32021.
[3] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, Huan Xu, Time Series Data Augmentation for Deep Learning: A Survey, arXiv:2002.12478, 2022.
[4] MIT Cambridge, Alexander Geiger, Dongyu Liu, Sarah Alnegheimish, Alfredo Cuesta-Infante, Kalyan Veeramachaneni, TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks, arXiv:2009.07769, 2020.
[5] ETH Zrich, Oliver Ammann, Gabriel Michau, Olga Fink, Anomaly Detection And Classification In Time Series With Kervolutional Neural Networks, arXiv:2005.07078, 2020.