Evaluation of deep autoencoder based anomaly detection with cold neutron source facility in HANARO

Seunghyoung Ryu [a], Byoungil Jeon [a], Minwoo Lee [b], Yonggyun Yu [a*]

[a]*Applied Artificial Intelligence Lab, Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon 34057, South Korea*
[b] *HANARO Management Division, Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon 34057, South Korea*
[*]*Corresponding author: ygyu@kaeri.re.kr*

1. Introduction

High-flux advanced neutron application reactor, namely HANARO, is a 30 MW scale multi-purpose research reactor in Korea. As the only domestic research reactor, it is actively utilized in the field of neutron science. Due to its importance and wide applicability to academic and industrial application, robust and reliable operation is required. HANARO is firmly protected by multiple safeguards and operated by experienced professionals, however, more stringent safety standards are required after the Fukushima Daiichi nuclear disaster in 2011. In this circumstance, deep learning methodologies are actively being studied for reactor safety improvement, for example, event diagnosis [1, 2, 3], reactor anomaly detection [4], sabotage detection [5], signal reconstruction [6], and probabilistic safety assessment [7].

In this research, we evaluate the feasibility of applying deep learning to anomaly detection with monitoring data from HANARO. Specifically, autoencoder based unsupervised anomaly detection is applied to cold neutron source (CNS) facility. Then we evaluate detection performance with synthetic abnormal data which is formed by injecting arbitrary bias.

2. Anomaly detection with autoencoder

Autoencoder is a class of neural network architecture composed of encoder and decoder networks. The encoder performs nonlinear dimension reduction and input data is converted to low dimensional latent vector. Then the decoder performs nonlinear dimension expansion and reconstructs original input data from a latent vector obtained from the encoder. Unsupervised deep anomaly detection can be conducted with autoencoder by learning latent vector of a normal dataset. Because autoencoder is trained to minimize reconstruction error on normal samples, autoencoder is able to successfully reconstruct normal input data. On the other hand, autoencoder fails to reconstruct abnormal samples resulting in high reconstruction error. Therefore, abnormal data can be classified by setting a threshold on reconstruction error (i.e., anomaly score).

CNS is an add-on facility of HANARO that produces cold neutrons by decelerating thermal neutrons. CNS is operated in conjunction with the reactor, thus abnormal behavior of CNS may induce sudden shutdown or trip of the reactor.

In this research, we train autoencoder with 43 CNS variables. From the historical database, we collect normal CNS data from the normal operation cycle when the reactor successfully finished its operation. Then we evaluate anomaly detection performance in terms of area under the receiver operating characteristic (AUROC) based on the synthetic abnormal data.

3. Experiment setup

3.1 Data preparation

We obtained anonymized CNS variables from 13 operation cycles. Then we divide 11 cycles for the training set and the most recent two cycles for the test set, respectively. There are 340,440 samples in the training set and randomly selected 300,000 samples are used for training and the rest 40,440 samples are used for validation. All data are normalized before training; z-score normalization is applied per channel.

Table 1. Data configuration

	Number of data	Cycle index
Training	300,000	69~89 (11 cycles)
Validation	40,440	
Test	41,760	91
	41,760	92

3.2 Autoencoder

We utilize simple autoencoder structure with 6 hidden layers and its details are described in Table 2.

Table 2. Autoencoder configurations

Model	Structure
Encoder	Input(43)-Dense(256)-LReLU-Dense(128)-LReLU-Dense(30)-LReLU
Decoder	Input(30)-Dense(128)-LReLU-Dense(256)-LReLU-Dense(43)

Note that the numbers in parenthesis indicate the number of neurons in the corresponding layer. The network is trained with adam optimizer with a learning rate of 0.001. The model is based on *PyTorch.*

3.2 Evaluation setup

To evaluate anomaly detection performance, we generate synthetic abnormal data by adding bias β to target channel c in test sample. If we denote input vector as $\mathrm{x} = [x_1, \dots, x_d] \in \mathbb{R}^{\mathrm{d}}$, synthetic anomaly data $\hat{\mathrm{x}}$, $\beta =$ 0.1 and $c = 3$, for example, is $[\mathrm{x}_1, \dots, x_3 + 0.1, \dots, x_d]$. Synthetic anomalies are created with different β and c. We change β from -1 to 1 (increased by 0.1) and c for 0 to 43, respectively. Then we compared the synthetic anomaly $\hat{\mathrm{x}}$ can be distinguished to normal sample x according to the reconstruction error. In doing this, we calculated two anomaly score: channel-wise and sample-wise anomaly score. Channel-wise anomaly score is squared error of specific channel c where β is added. Sample-wise anomaly score is mean squared error of sample thus errors on other channels are included.

Based on two anomaly score, we calculate AUROC which is a measure of separability between two distribution, i.e., anomaly score distribution of normal and abnormal data. ROC curve is a graph of true positive rate (y-axis) versus false positive rate (x-axis) according to the varying threshold; perfect classifier will be located at (0, 1). AUROC is the area under ROC curve. Because the perfect classifier located at (0,1), AUROC score close to 1 indicates two classes can be easily separated. On the other hand, 0 indicates that two classes are classified as completely opposite. Note than AUROC of uniform random classification is 0.5.

4. Experiments

4.1 Reconstruction on test set.

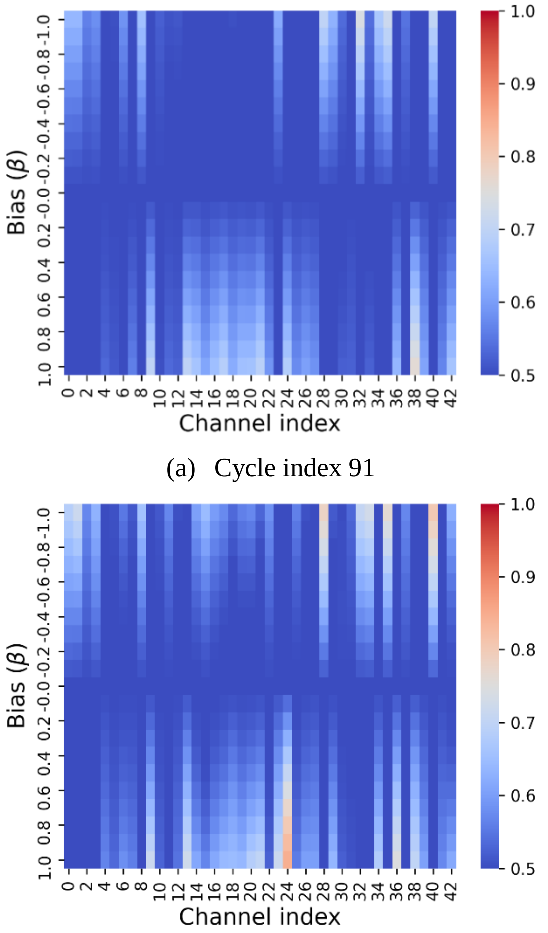
Below Table 3 describes reconstruction error (average and standard deviation in parenthesis) of two test cycles in terms of mean squared error (MSE), and mean absolute error (MAE). Compared to the errors on training and validation set, test set shows higher reconstruction errors. The result indicates different insights on generalization performance of trained model; the variation of observed data for each cycle may be greater than expected, or the model is overfitted. In application of anomaly detection, distinguishing anomaly is more important than improving generalization performance, hence we analyze detection performance with synthetic anomaly in following subsection.

Table 3. Error comparison

Error		MSE	MAE
Training		0.02 (1.21)	0.06 (0.05)
Validation		0.02 (1.38)	0.06 (0.05)
Test	91	0.91 (6.87)	0.48 (0.22)
	92	0.42 (0.11)	0.46 (0.07)

4.2 Anomaly detection result

We calculate AUROC of varying biases and channels, and visualize the result in forms of heat map. Figs 1 and 2 are heatmaps of AUROC score where x axis represent target channel c and y axis represents bias β. The color of the corresponding cell indicates the AUROC value.



(a) Cycle index 91

(b) Cycle index 92

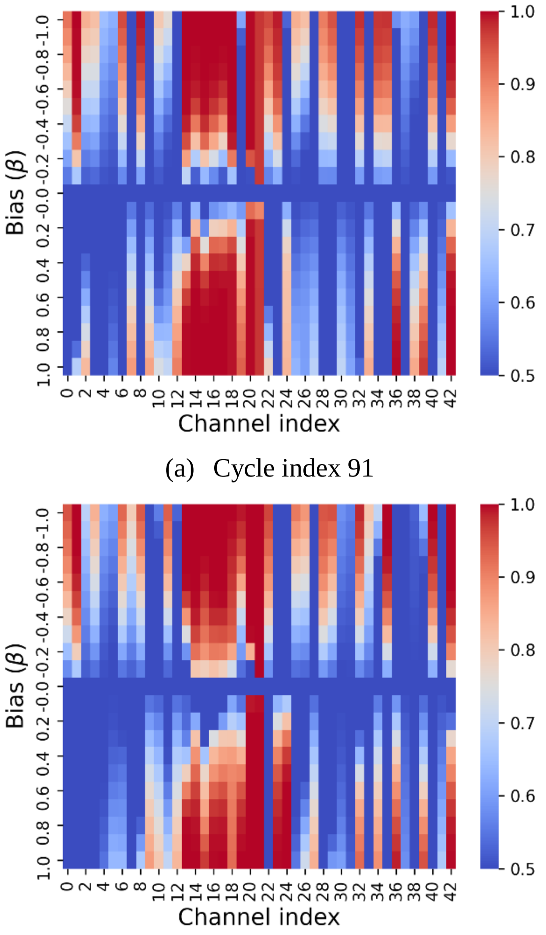
Fig.1. AUROC heat map on sample-wise score

As can be seen, AUROC of channel-wise score is higher than sample-wise score because errors on other channels (channels where bias is not added) reduces the impact of synthetic bias on target channel.

In addition, the impact of synthetic bias is differ according to the channel and some channels are more sensitive. For example, channel 12 to 21 shows high AUROC score compared to other channel, and channels 0, 3, 4 shows lower AUROC scores on negative bias.

5. Conclusions

We propose autoencoder based anomaly detection framework for CNS facility in HANARO. We train deep autoencoder with observed CNS variables in normal operation cycles so as to learn normal latent features from input data. Then, anomaly detection performance is evaluated with synthetic abnormal data by adding bias per channel. The result show that due to averaging effect on reconstruction error, channel-wise anomaly score is more sensitive to synthetic bias. In this regard, For the large scale system where the number of channel is a lot higher, channel-wise anomaly detection can be more effective.



(a) Cycle index 91

(b) Cycle index 92

Fig.2. AUROC heat map on channel-wise score

For future works, sensitivity analysis in original domain (because bias is added after data normalization) and improvement in deep neural network architecture is required.

ACKNOWLEDGEMENTS

This research was supported by a grant from Korea Atomic Energy Research Institute (KAERI) R&D Program (No. KAERI-524450-21)

REFERENCES

[1] T.-H. Lin, T.-C. Wang, S.-C. Wu, Deep learning schemes for event identification and signal reconstruction in nuclear power plants with sensor faults, Annals of Nuclear Energy 154 (2021) 108113.

[2] M. C. dos Santos, V. H. C. Pinheiro, F. S. M. do Desterro, R. K. de Avellar, R. Schirru, A. dos Santos Nicolau, A. M. M. de Lima, Deep rectifier neural network applied to the accident identification problem in a PWR nuclear power plant, Annals of Nuclear Energy 133 (2019) 400–408.

[3] M.I. Radaideh, C. Pigg, T. Kozlowski, Y. Deng and A. Qu, Neural-based time series forecasting of loss of coolant accidents in nuclear power plants, Expert Systems with Applications 160 (2020) 113699.s

[4] F. Caliva, F. S. De Ribeiro, A. Mylonakis, C. Demazi'ere, P. Vinai, G. Leontidis, S. Kollias, A deep learning approach to anomaly detection in nuclear reactors, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8.

[5] S. Chen, K. Demachi, Proposal of an insider sabotage detection method for nuclear security using deep learning, Journal of Nuclear Science and Technology 56 (2019) 599–607.

[6] S. G. Kim, Y. H. Chae, P. H. Seong, Development of a generative-adversarial-network-based signal reconstruction method for nuclear power plants, Annals of Nuclear Energy 142 (2020) 107410.

[7] H. Kim, J. Cho, J. Park, Application of a deep learning technique to the development of a fast accident scenario identifier, IEEE Access 8 (2020) 177363–177373. H. Kim, J. Cho, J. Park, Application of a deep learning technique to the development of a fast accident scenario identifier, IEEE Access 8 (2020) 177363–177373.