

A Case Study on Performance Evaluation of AI-Based Regression Models in Nuclear Instrumentation and Control

Jackwan 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 :** Artificial Intelligence, Regression, Instrumentation and Control, Performance Metrics, Model Evaluation

1. Introduction

Artificial intelligence (AI) models that use multivariate time-series signals as data are being researched and developed in the field of nuclear instrumentation and control (I&C) to support plant condition monitoring and operator decision-making. Among them, regression models can predict changes in instrumentation signals and can therefore be used to forecast changes in plant operating conditions. Before deployment in the field, the performance of developed AI models needs to be confirmed through appropriate testing [1]. Reference [2] introduced representative performance metrics applicable to classification and regression models and presented a case study on the performance evaluation of a classification model. Based on these previous studies, this paper proposes a practical performance evaluation method applicable to a time-series prediction model being developed to support nuclear power plant operation.

Using a single metric may not be sufficient to evaluate a regression model. MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) express errors in the physical units of the data, making direct comparison difficult across signals with different scales. MAPE (Mean Absolute Percentage Error) and the coefficient of determination (R^2) are ratio-based metrics, but under certain conditions they may overstate errors. Therefore, rather than uniformly applying a single metric, it is desirable to use two or more metrics and refer to the metric appropriate to the situation [3].

2. Performance Evaluation Method for Regression Models

2.1 Performance Metrics and Limitations

As performance metrics have not yet been formally defined in standard documents, this study selects MAE, RMSE, MAPE, SMAPE (Symmetric MAPE), and R^2 , which are widely used in the IT field, as performance metrics. For n samples, let the actual value be y_i and the predicted value be $\hat{y}_i$. The five performance metrics mainly used in this analysis are defined as follows.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

$$SMAPE = \frac{100}{n} \sum_{i=1}^n \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

MAE most directly represents the average prediction error in physical units, while RMSE places greater weight on large errors. MAPE expresses error as a percentage and is therefore suitable for comparing output signals with different magnitudes. However, when y_i approaches zero, the denominator can cause MAPE to become excessively large [4]. R^2 indicates how well the predictions explain the variability of the actual values, but for actual-value signals with very low variability, the denominator of R^2 becomes very small. In this case, even if the absolute error is small, R^2 may take a large negative value, thereby overstating the apparent error. SMAPE can be used as an auxiliary metric that partially mitigates the limitations of MAPE.

2.2 Selection of Performance Metrics and Performance Analysis Strategy

Nuclear I&C signal data are multivariate time series. In addition, during normal operation, there are many signals whose actual values are zero and many signals that exhibit little variation. Based on these characteristics, when the actual values are sufficiently greater than zero, MAPE, which is widely used for regression models, is used as the primary evaluation metric. However, when the actual values include near-zero values, MAPE is excluded from aggregation for the corresponding test case. In such cases, other auxiliary metrics are used for evaluation.

R^2 is used as an auxiliary metric for checking trend reproducibility only when the actual-value signal has sufficient variance. When the variance of the actual values is very small, R^2 is excluded from performance aggregation, and MAPE, MAE, and other metrics can be

used for evaluation. These exclusion criteria are intended to prevent incorrect conclusions arising from inherent characteristics of the performance metrics.

Because AI applied in nuclear I&C can be used for plant condition monitoring and operator decision support, it is difficult to determine the reliability of an AI application solely from the overall average performance of the model. Therefore, testing of AI regression models for nuclear applications should generate not only overall prediction performance results but also performance results by model output signal and by scenario. Based on these data, detailed evaluations are required, such as assessing the suitability of using the model for its assigned function, evaluating appropriateness through analysis of each scenario covered by the model, and assessing the reliability of each output signal through analysis of its influence on model performance.

Accordingly, this paper presents the actual values and model-predicted values from regression-model testing as graphs so that reviewers can visually and readily inspect the error magnitude or error rate. In addition, evaluation results are recorded in files hierarchically at the individual-signal, test-file, model, and final-summary levels so that, if model performance degradation occurs, reviewers can readily identify which signal or scenario data caused it.

3. Case Study of Regression Model Evaluation

3.1 Evaluation Target and Test Configuration

The evaluation target is an AI regression model for progression prediction of major trip variables in a nuclear power plant using multivariate time-series data. For each test case, the input consists of 60 time intervals of 254 signals, and the model was developed to predict 240 future time intervals for 10 output signals. The output signals are RCS T cold 1B, RCS T cold 1A, Var Overpower, SG 1 Press, SG 2 Press, SG 2 W Lvl, SG 1 W Lvl, PZR Press, CNMT Press Hi, and DNBR. The prediction model employs a hybrid Autoformer-TCN-VAE architecture. It combines moving-average-based trend-seasonality decomposition and an Autoformer encoder-decoder with a temporal convolutional network (TCN) to extract both long-term dependencies and local temporal patterns from multivariate time-series data. A variational autoencoder (VAE) is incorporated to represent latent features, and the final dense layer generates predictions for the 10 output signals over 240 future time steps. The regression-model test and evaluation tool loads a file containing the trained model structure and 92 test case files for 41 abnormal scenarios, performs the tests, calculates performance metrics for each output signal, saves the results to files, and displays predicted-versus-actual trend graphs.

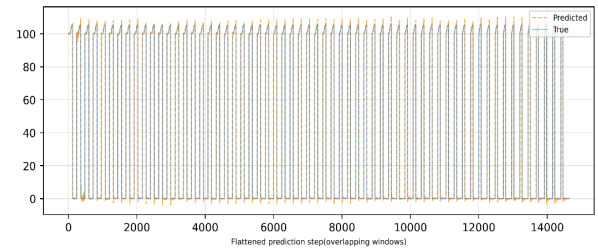
3.2 Evaluation Results and Interpretation of Performance Metrics

The mean MAPE calculated from the performance results for the 41 scenarios was 2.13%, and the median MAPE was 1.46%. MAPE was below 5% for 35 scenarios and between 5% and 10% for 6 scenarios,

while no scenario had a model-level average MAPE of 10% or higher. Although the overall results appear favorable, detailed signal-level analysis identified recurrent or scenario-dependent problems that were difficult to identify from model-level averages alone.

The performance of RCS T cold 1B, RCS T cold 1A, SG 1 Press, SG 2 Press, and PZR Press was below 5% MAPE in all cases. In contrast, the median MAPE of CNMT Press Hi was 10.43%; among the 41 scenarios, 28 were 5% or higher and 21 were 10% or higher, with a maximum of 68.13%. SG 2 W Lvl and SG 1 W Lvl each exceeded 10% in two scenarios, with maximum values of 26.68% and 26.93%, respectively. Cases in which MAPE was excluded due to near-zero actual values were also present, including for DNBR; among the remaining valid scenarios, DNBR exceeded 10% in one case, reaching 20.65%. In particular, for Var Overpower, the percentage error was calculated to be very large in intervals where the actual signal approached zero, even though the prediction trend itself was favorable.

Figure 1 presents two cases illustrating why performance metrics should be interpreted according to the conditions. In Figure 1, the model output variable Var Overpower follows the actual-value trend very closely, with $R^2=0.9993$ and $MAE=0.3376$; however, MAPE is calculated to be abnormally large because the actual values repeatedly approach zero. Therefore, in this case,



MAPE is excluded from aggregation because a calculation error occurs, rather than being used as direct evidence of model performance degradation. Figure 2 shows an actual prediction-performance degradation case for the CNMT Press Hi variable. Although the small MAE may make the performance appear favorable, MAPE provides a better basis for judgment in this small-value range. MAPE reaches 21.83%, and the model fails to reproduce the consistent variation in the actual values. Fig. 1. Case of MAPE calculation error and actual prediction performance (MAPE = 1,325,879.69%, $R^2 = 0.9993$, MAE = 0.3376).

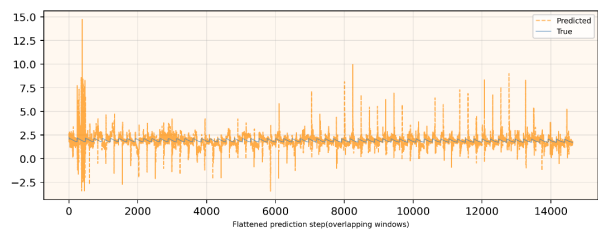


Fig. 2. Case of MAE limitation and MAPE-based judgement (MAPE = 21.83%, $R^2 = -34.8651$, MAE = 0.4218).

Analysis of the scenario-specific MAPE test-result summary file in Figure 3 provided detailed information

on factors affecting performance. It was found that the prediction error rates of the SG 1 W Lvl and SG 2 W Lvl variables were problematic in scenarios 3522-03-04 and 3531-01-02. In addition, CNMT Press Hi was identified as a cause of performance degradation in many scenarios. This information can be used by model developers to select additional training data for model improvement and to improve the model structure for the corresponding output.

scenario	RCS T cold 1B	RCS T cold 1A	Var Overpower	SG 1 Press	SG 2 Press	SG 2 W Lvl	SG 1 W Lvl	PZR Press	CNMT Press Hi	DNBR
3431-01-07	0.17	0.16	N/A	0.55	0.56	1.02	1.04	1.92	38.02	3.65
3431-02-04	0.04	0.05	0.14	0.14	0.14	0.10	0.10	0.48	11.30	0.30
3431-05_2	0.02	0.02	0.05	0.05	0.07	0.05	0.09	0.04	5.02	0.07
3431-05_3	0.01	0.01	0.03	0.03	0.04	0.04	0.04	0.02	2.02	0.04
3431-05_4	0.02	0.02	0.04	0.04	0.05	0.06	0.03	0.02	2.90	0.05
3431-09_2	0.05	0.05	0.79	0.16	0.16	0.18	0.16	0.30	6.48	0.31
3431-09_3	0.09	0.08	N/A	0.14	0.15	0.51	0.47	0.57	14.52	0.76
3431-09_4	0.05	0.08	N/A	0.15	0.14	0.92	0.79	0.41	10.43	N/A
3431-13_2	0.06	0.06	N/A	0.59	0.53	1.16	1.10	1.40	1.59	7.10
3431-13_3	0.04	0.04	0.16	0.11	0.09	0.09	0.08	0.26	9.05	0.19
3440-01-02	0.13	0.18	N/A	0.19	0.21	0.57	0.55	0.19	21.32	0.81
3451-01_2	0.25	0.22	1.17	0.55	0.69	1.60	0.82	0.25	37.26	0.54
3451-01_3	0.16	0.16	0.55	0.46	0.35	0.34	0.25	0.22	18.81	0.66
3451-02_2	0.07	0.07	0.20	0.16	0.17	0.12	0.14	0.13	9.95	0.25
3451-04-02	0.03	0.04	1.42	0.20	0.18	0.26	0.23	1.46	62.54	1.17
3451-04_3	0.03	0.04	0.06	0.06	0.09	0.06	0.12	0.04	5.69	0.14
3451-05_3	0.01	0.01	0.05	0.05	0.05	0.03	0.03	0.26	1.55	0.10
3461-00-6	0.12	0.10	0.29	0.15	0.26	0.25	0.13	0.11	16.68	0.23
3461-00_2	0.04	0.03	0.10	0.07	0.08	0.08	0.09	0.10	7.32	0.09
3461-00_3	0.02	0.02	0.07	0.02	0.03	0.02	0.02	0.36	1.60	0.08
3462-01_2	0.22	0.22	0.43	0.41	0.45	0.41	0.38	0.37	30.89	0.95
3462-01_3	0.01	0.01	0.02	0.01	0.02	0.02	0.02	0.01	1.24	0.02
3514-00_2	0.17	0.19	5.48	0.65	0.63	1.17	1.02	0.60	21.72	0.68
3515-00_3	0.09	0.10	3.21	0.56	0.21	0.33	0.34	0.54	17.02	0.60
3515-02_2	0.31	0.33	6.46	1.15	0.54	0.66	0.52	0.52	52.57	0.83
3521-01_2	0.47	0.31	1.04	1.02	0.99	1.29	0.95	0.46	16.66	1.02
3521-02_2	0.04	0.04	0.31	0.17	0.18	0.04	0.04	0.06	2.96	0.35
3522-02_3	0.45	0.40	N/A	1.27	1.24	6.15	6.73	0.57	68.13	5.45
3522-03-04	0.42	0.43	N/A	2.43	2.41	20.95	21.00	0.91	8.75	8.03
3530-01-02	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.02	1.00	0.06
3531-01-02	0.58	0.56	N/A	2.58	2.46	26.68	26.93	1.24	46.76	6.18
3531-01_5	0.08	0.06	0.15	0.12	0.10	0.12	0.18	0.09	13.53	0.17
3535-01-02	0.47	0.45	N/A	1.98	1.99	4.12	4.11	1.25	30.10	20.65
3540-01_3	0.12	0.16	0.27	0.19	0.21	0.32	1.99	0.20	18.46	0.36
3541-05_2	0.01	0.01	0.03	0.03	0.02	0.18	0.17	0.02	2.10	0.03
3541-06_2	0.02	0.03	0.05	0.04	0.05	0.10	0.17	0.03	3.41	0.13
3551-00_2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.76	0.01
3711-01_2	0.03	0.02	0.06	0.03	0.06	0.03	0.03	0.02	3.48	0.06
3711-01_5	0.93	1.01	4.60	4.27	4.19	2.14	2.13	0.58	52.07	5.94
3731-01_2	0.18	0.17	0.90	0.36	0.28	0.27	0.35	0.26	25.72	1.60
3741-03_3	0.08	0.09	0.67	0.47	0.29	0.10	0.12	0.13	4.90	1.72

Fig. 3. Summary table of MAPE results for output signals by scenario

4. Conclusions

This paper proposed a practical performance evaluation method applicable to AI-based regression models for nuclear power plant operation support. Scale-independent MAPE was used as the primary metric, while MAE and R^2 were interpreted as auxiliary metrics according to signal characteristics. MAPE was excluded when the actual values were near zero, and R^2 was excluded when the variance of the actual-value signal was very small.

In the evaluation case study, most scenarios showed low MAPE overall; however, detailed analysis revealed recurrent weaknesses in CNMT Press Hi and performance degradation of steam-generator water-level signals in specific scenarios. It was also confirmed that a very large MAPE or a large negative R^2 does not always indicate poor prediction. Therefore, before making a final judgment at the model level, it is necessary to confirm the validity of the performance metrics and to review the scenario- and signal-level results together.

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 Climate, Energy & Environment (MCEE) of the Republic of Korea (No. 20224B10100130).

REFERENCES

- [1] ISO/IEC TR 29119-11, Software and systems engineering - Software testing - Part 11: Guidelines on the testing of AI-based systems, 2020.
- [2] J. Park and S. Koo, A Case Study on Performance Evaluation of AI Models in Nuclear Instrumentation and Control, Transactions of the Korean Nuclear Society Autumn Meeting, Changwon, Korea, October 30-31, 2025.
- [3] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Computer Science, Vol. 7, e623, 2021.
- [4] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," International Journal of Forecasting, Vol. 22, No. 4, pp. 679–688, 2006.