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

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 : Artificial Intelligence, Instrumentation and Control, Performance Metrics, Model Evaluation

1. Introduction

Artificial intelligence (AI) model development is rapidly advancing in the field of nuclear instrumentation and control (I&C), where signals from numerous sensors are utilized for monitoring and decision-making. While many studies report model performance using the most intuitive metric, accuracy, this metric alone is insufficient[1-2]. Accuracy does not adequately address the problems of class imbalance and fails to identify the types of errors, such as false positives and false negatives, that may critically affect decision-making in nuclear applications.

To overcome these limitations, it is necessary to adopt a wider set of performance metrics appropriate to the model type. International standards and technical documents, such as ISO/IEC TS 4213[3] and ISO/IEC TR 29119-11[4], provide guidelines for evaluating AI model performance.

In nuclear I&C, AI models typically use multivariate time series data as input and are designed to solve either classification or regression problems. This paper presents a case study of performance evaluation for AI models under development for nuclear decision support, using multiple performance metrics to ensure robust and reliable assessments.

2. Performance Evaluation of Nuclear AI Models

This paper introduces representative performance metrics applicable to two major types of AI models: classification and regression. In addition, a classification model currently being developed for supporting nuclear power plant decision-making is selected, and its performance is analyzed from multiple perspectives using these metrics.

2.1 Types of AI Models: Classification and Regression

AI models can be broadly categorized into classification models and regression models. While both types aim to predict outputs based on input data, they differ fundamentally in the nature of the output they predict. Classification models are used to assign input data into one of several predefined categories. Common examples include spam detection in emails or diagnosing diseases based on patient data. Regression models are used to predict continuous numerical values. Examples include predicting house prices, stock market indices, or future values of sensor signals.

2.2 Performance Metrics for Classification Models

The performance of classification models is typically measured using metrics such as accuracy, precision, recall, and F1-score.

- ✓ Accuracy: Proportion of correct predictions among all samples.
- ✓ Precision: Proportion of correctly predicted positives among all positive predictions.
- ✓ Recall: Proportion of correctly predicted positives among all actual positives.
- ✓ F1-score: Harmonic mean of precision and recall, useful when seeking a balance between the two.

These metrics are derived from the confusion matrix, which summarizes the test results. Figure 1 illustrates a confusion matrix for a classification model, where the rows represent the actual values (positive or negative) and the columns represent the predicted values (positive or negative).

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="4">Model Prediction</td><td rowspan="2">Metrics based on Actual State</td></tr>
<tr><td colspan="2">Positive</td><td colspan="2">Negative</td></tr>
<tr><td rowspan="4">Actual State</td><td rowspan="2">Positive</td><td colspan="2" rowspan="2">True Positive (TP)</td><td colspan="2" rowspan="2">False Negative (FN)</td><td>Recall</td></tr>
<tr><td>False negative rate(FNR), miss rate</td></tr>
<tr><td rowspan="2">Negative</td><td colspan="2" rowspan="2">False Positive (FP)</td><td colspan="2" rowspan="2">True Negative (TN)</td><td>Fallout, False positive rate(FPR)</td></tr>
<tr><td>Specificity, Selectivity, True negative rate(TNR)</td></tr>
<tr><td colspan="2">Metrcis based on Model Prediction</td><td>Precision, Positive predict value(PPV)</td><td>False discovery rate(FDR)</td><td>False omission rate(FOR)</td><td>Negative predictive value(NPV)</td><td>Accuracy, F1-score</td></tr>
</table>

Fig. 1. Structure of the confusion matrix and evaluation metrics for classification models

Once the test cases are applied to the model and the confusion matrix is populated, various performance metrics can be calculated based on the results.

- ✓ The overall average performance of the model across the test cases, accuracy is calculated as $\frac{TP+TN}{TP+TN+FP+FN}$.

- ✓ Precision, which focuses on reducing false positives, is calculated as $\frac{TP}{TP+FP}$.
- ✓ Recall, which emphasizes reducing false negatives, is calculated as $\frac{TP}{TP+FN}$.
- ✓ Finally, the F1-score, a balanced metric computed as the harmonic mean of Precision and Recall, is calculated as $2 \times \frac{Precision \times Recall}{Precision+Recall}$.

These metrics are essential for assessing performance, especially in multiclass or imbalanced classification problems.

2.3 Performance Metrics for Regression Models

Regression models are evaluated based on how accurately they predict numerical values, using the following metrics:

- ✓ MAE (Mean Absolute Error): Average of absolute differences between predictions and actual values. Interpretable with original data units.
- ✓ RMSE (Root Mean Squared Error): Square root of average squared errors. Penalizes larger errors more.
- ✓ MAPE (Mean Absolute Percentage Error): Average of percentage errors; useful for comparing relative accuracy.
- ✓ R^2 Score (Coefficient of Determination): Indicates how much of the output variance is explained by the model. A value close to 1 indicates high explanatory power.

The formulas for computing these performance metrics are given below.

- ✓ $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- ✓ $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- ✓ $\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$
- ✓ $R^2 = 1 - \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$

(n denotes the number of test cases, y_i is the i-th actual value, $\hat{y}_i$ is the i-th predicted value by the model, and $\bar{y}$ represents the mean of the actual values.)

These metrics allow a comprehensive evaluation of regression models in terms of absolute accuracy, sensitivity to large errors, and explanatory strength.

2.4 A Case Study: Performance Evaluation of a Classification Model

This paper presents the performance evaluation results of a specific classification model developed for nuclear power plant operation support. The model is designed for the early diagnosis of abnormal plant conditions and uses one normal state and 51 abnormal states as output labels.

The model input consists of a multivariate time series with 759 sensor signals over 120 time steps, and the output is a single classification value representing the plant condition. Performance testing was conducted using 91 test case files that were not included in the training process. Each file contained data representing normal conditions in the initial phase and abnormal conditions in the later phase.

The detailed test results are shown in Figure 2. Overall, the model demonstrated high performance, achieving an accuracy of 97.79%, precision of 95.37%, recall of 95.25%, and F1-score of 95.0%. However, label-wise analysis revealed that several abnormal states, including "3451-02_3" and "3741-03_3," exhibited relatively high misclassification rates. Specifically, in test samples belonging to five abnormal labels, between 18% and 40% of the inputs were incorrectly classified into other abnormal categories. These findings provide valuable feedback to model developers and serve as a basis for improving model performance. The case study illustrates that issues not captured by accuracy can be more precisely uncovered through label-level analysis in combination with metrics such as precision, recall, and F1-score.

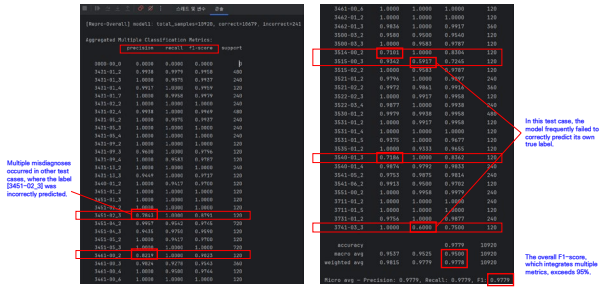

Fig. 2. Test results conducted on 91 test case files.

3. Conclusions

This study highlights the importance of applying diverse and context-appropriate performance metrics when evaluating AI models for nuclear instrumentation and control. Sole reliance on accuracy is insufficient, particularly in safety-critical area where class imbalance and error type are of high concern. By applying a combination of classification and regression metrics, a more comprehensive understanding of model behavior can be obtained. The case study demonstrated that while the classification model achieved high overall performance, detailed metric-based analysis revealed areas requiring improvement.

As future work, we plan to conduct detailed performance analysis and evaluation of regression models. In addition, studies will be carried out on testing model performance under data drift conditions and assessing model robustness against adversarial attacks.

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). This work was also supported by the Nuclear Safety Research Program through the Regulatory Research Management Agency for SMRS (RMAS) and the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No. 1500-1501-409). Finally, this work was supported by Korean Institute of Energy Technology Evaluation and Planning (KETEP), and Ministry of Trade, Industry, and Energy (MOTIE) of the Republic of Korea (No. 20224B10100130).

REFERENCES

[1] Amalia Luque, Alejandro Carrasco, Alejandro Martín, Ana de las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, Pattern Recognition, Vol.91, p.216-231, 2019

[2] Ghanem M, Ghaith AK, El-Hajj VG, Bhandarkar A, de Giorgio A, Elmi-Terander A, Bydon M. Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review. Brain Sci., 16;13(12):1723. doi: 10.3390/brainsci13121723, 2023.

[3] ISO/IEC TS 4213, Information technology — Artificial intelligence — Assessment of machine learning classification performance, 2022.

[4] ISO/IEC TR 29119-11, Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems, 2020.