Feasibility Study of an Explainable AI-based Anomaly Detection for Nuclear Reactor Core Operation in PWRs

Hanjoo Kim [a], Sang-Rae Moon [b], Deokjung Lee [a*]
[a] *Department of Nuclear Engineering, Ulsan National Institute of Science and Technology UNIST-gil 50, Ulsan, 44919*
[b] *Core Analysis Group, Korea Hydro & Nuclear Power Central Research Institute (KHNP-CRI), Daejeon, 34101*
[*] *Corresponding author: deokjung@unist.ac.kr*

1. Introduction

This study explores the feasibility of applying explainable AI (XAI) technology to detect anomalies in nuclear reactor core.

The operation of pressurized water reactors (PWRs) has recently increased in cycle length and power, leading to a greater need for anomaly mitigation technologies such as predicting axial offset anomaly (AOA). Additionally, the importance of nuclear energy has grown in response to the need to address climate change and stabilize energy supply. Accordingly, the need for enhanced safety measures in nuclear energy has also increased. As part of these efforts, studies on AI-based nuclear reactor core anomaly detection have been conducted [1][2][3]. It has constructed a simulation-based machine learning (ML) technology to diagnose and predict control rods mis-location, coolant inlet temperature asymmetry, CRUD induced power shift (CIPS), in-core detector signal error, and so on by using a nuclear reactor analysis code called RAST-K [4].

Despite the superior performance of AI technology, the application of AI in mission-critical industries such as financial services, healthcare, and nuclear power is limited due to its black box nature which refers to the complex decision-making processes of MLs, where the inputs and outputs of the model are known, but the internal workings of the model and how it arrives at its decisions are opaque and not easily explained [5]. To enhance the practical applicability of AI-based anomaly detection for nuclear reactor core operation industry, feasibility of applying explainable AI technologies was studied.

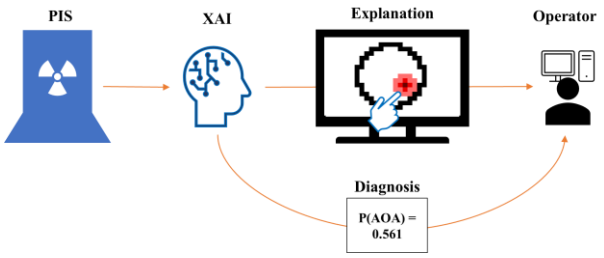


Fig. 1. Concept of XAI-based nuclear reactor core anomaly detecting system during operation.

2. AI-based core anomaly detection method

In this section, a framework of implementing ML system for core anomaly detection is described.

2.1 Data Acquisition

Collecting operational data of a nuclear core, especially regarding core anomalies, can pose a challenge due to its potential danger and the high cost involved. Furthermore, the strict safety regulations in place at nuclear power plants make it difficult to gather data from actual reactor operations. To overcome these difficulties, a simulation-based data generation method was adopted to generate synthetic data that represents a range of operating conditions and scenarios. This synthetic data can then be used to train and validate machine learning models. The RAST-K [4] nuclear reactor core analysis code was utilized to generate this synthetic reactor operation data, based on an OPR-1000 reactor core. The procedure for generating this data is outlined as follows:

① Building a nuclear reactor core model by sampling input parameters.
② Conducting core calculations using RAST-K on the model created in step ①.
③ Extracting relevant parameters from the output text file.

By repeating steps ① to ③, output parameters of various core model are collected in a single text file formatted as ‘csv’, which is then directly used for training. The data is labeled as normal or anomaly based on the core conditions determined by the input parameters sampled in step ① or the power shape calculated in step ③. A simulation-generated operation data includes class label, control rods position signals which is indicated by operator (73 features, feature 0~72), in-core instrument (ICI) signals (225 features, feature 73~297) and ex-core detector signals (18 features, feature 298~315), representing power distribution inside a reactor vessel. Figure 2 shows an example of dataset.

The feasibility study for XAI-based core anomaly detection focused on Axial Offset Anomaly (AOA) caused by CRUD-Induced Power Shift (CIPS). The CIPS model simulates the buildup of CRUD that can result in an axial power tilt during operation. A snapshot of core operation data is labeled as 'CIPS' if a severe AOA occurs within the next 30 days, where the Axial Shape Index (ASI) deviates from the design and operating values by more than 3%. The generated synthetic data represents

operating conditions at 100% power level, at the middle of cycle (MOC) with a burnup of 8.0-9.0 GWd/MTU.

2.2 ML Model Description

Since the generated training data has labels, the model belongs to the supervised learning. Target of the learning is to categorize a core operation state as either belonging to normal or abnormal, thus making it a task of classification learning. An ensemble ML model based on decision tree called Random Forest was used as classifier.

Random Forest (RF) is a machine learning algorithm that combines multiple decision trees (DTs) using a technique called bagging to improve accuracy and reduce overfitting [7]. In a RF, final prediction is made by aggregating the prediction of all decision trees in the forest. Each tree in the forest is built independently using a different subset of the training data and input features. It is a powerful and versatile algorithm that is useful for a variety of tasks. To construct the random forest model, 200 decision trees were employed and trained on subsets of the data containing 16 features, with a maximum depth of 16.

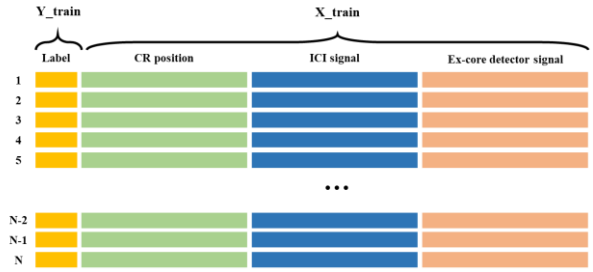


Fig. 2. Example of dataset for training and testing nuclear reactor core anomaly detecting ML models.

3. Explainable AI methods

Various explainable AI techniques were applied to the ML model developed for core anomaly detection to assess the feasibility of its applicability. This section describes XAI methods applied to the ML model.

3.1 Mean Decreased Impurity

Mean decreased impurity (MDI) is a feature importance evaluation method that measures how much a feature reduces the impurity of a DT in a RF. The importance of each feature is calculated by averaging its reduction in impurity across all trees in the forest. Calculation of impurity reduction by a i^{th} feature during node splitting in a DT is shown as Eq. 1 and Eq. 2.

Gini impurity at node j is defined as the probability of each class $p(k)$:

$$G(N_j) = 1 - \sum_{k=1}^{K} p(k)^2 \cdot \quad (1)$$

Reduction of impurity during splitting a node j by the i^{th} feature into two nodes, called left node and right node, is as follow:

$$I_i(N_j) = \omega_j G(N_j) - \omega_{j,l} G(N_{j,l}) - \omega_{j,r} G(N_{j,r}) \cdot \quad (2)$$

The MDI can guide feature selection and engineering efforts but may not be accurate in models with highly correlated features.

3.2 Permutation Importance

Permutation importance [8] evaluates feature importance in a machine learning model by randomly shuffling values of individual features and observing changes in the model's performance as written in Equation (3). Features that are important to the model will cause significant decreases in performance when shuffled. Since it calculates difference between performance on original data and shuffled data, it can be performed with a single model without retraining or cross-validation. While it is useful for identifying critical features, it may not capture interactions or accurately measure highly correlated feature sets.

$$I_i = A(Original) - A(Permuted) \cdot \quad (3)$$

3.3 Local Interpretable Model-agnostic Explanations

Local interpretable model-agnostic explanations (LIME) [9] is a methodology for XAI that explains any machine learning model's predictions by generating a surrogate model around a specific data point of interest. To generate a local explanation with LIME, many synthetic samples or perturbations are generated by randomly modifying the features. The synthetic samples are used to train simple, interpretable surrogate model, which is then used to identify the features that are most important for the original model's predictions at the selected instance. LIME adopts a model-agnostic approach, and it is useful for interpreting complex models, identifying biases, and increasing transparency and accountability in AI systems.

3.4 Shapley Additive Explanation

Shapley additive explanation (SHAP) [10] is a methodology for XAI that explains individual predictions made by machine learning models. It assigns an importance value to each feature based on its contribution to the prediction, using game theory, thus computing SHAP values of each feature. SHAP values can be used to improve model accuracy, identify biases, and increase transparency and accountability in AI systems. It has been widely applied in various fields and is particularly useful for high-dimensional models.

Shapley value ϕ_i of a feature i can be calculated as:

$$\phi_i = \sum_{S \subseteq F \setminus i} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \cup i}(x_{S \cup i}) - f_S(x_S)] \,, \quad (4)$$

where F is subsets of all features, S is all possible subsets of F excluding the feature i, $f_{S \cup i}(x_{S \cup i}) - f_S(x_S)$ is a difference of predictions from two models, a model $f_{S \cup i}$ is trained with feature i present and the other model is trained with the feature withheld.

4. Results

4.1 Classifier Model Performance

Performance of the ML model learned with the simulation-based training data is written in Table I with evaluation metrics such as accuracy, precision, recall scores, which are defined as Eq. 5~7, and Receiver Operating Characteristic (ROC) curve in Figure 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$precision = \frac{TP}{TP + FP} \quad (6)$$

$$recall = \frac{TP}{TP + FN} \quad (7)$$

Table I. Classification evaluation metrics for the RF model on detecting CIPS.

	Accuracy	Precision	Recall
CIPS	0.8767	0.4231	0.9698

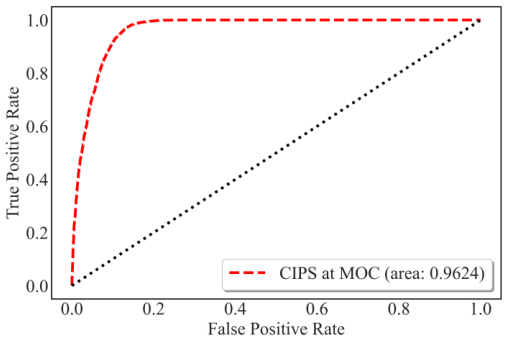


Fig. 3. ROC curve of the RF model on detecting CIPS

Table I shows the evaluation metrics for the ML model used to detect CIPS, including accuracy, precision, and recall scores of 87.7%, 42.3%, and 97.0%, respectively. It's important to note that the labeled data used for this model is imbalanced, which can cause a discrepancy between the precision and recall scores. The imbalanced ratio for the dataset is 0.10, which means that the proportion of instances in the minority class to the majority class is very small. Despite this challenge, the classifier is able to predict the occurrence of CIPS up to 30 days in advance with 97% recall score. This is a promising result and suggests that the model is effective in detecting CIPS, even in an imbalanced dataset.

4.2 Global Feature Importance

The importance of each feature, including 73 control rod positions, 225 ICI signals, and 18 ex-core detector signals, was evaluated using both Mean Decreased Impurity (MDI) and Permutation Importance (PI) in relation to the CIPS detection model and the dataset. Figure 4 show importance of each feature estimated by both methods. Figures 5 and 6 shows the radial distribution of feature importance obtained by integrating importance of five axial ICI signals and three ex-core detector signals each. In contrast, Figure 7 shows the feature importance of ICI signals from five axial levels obtained by integrating the radial signals from the same level.

PI considers the change in model performance when a feature is randomly shuffled. If shuffling a feature leads to an increase in performance, then that feature is assigned a negative importance score as shown in Figure 4. In both methods, features 1 to 73, which represent control rod position, are shown to have negligible importance. This is because control rod position is fixed in generating core models with full power operation and does not affect model performance.

Accounting for radial feature importance with MDI shows that the ICI signals at the periphery region, where fresh fuel is loaded, are relatively important. Fresh fuel assembly leads to higher power, thus making CRUD more likely to form and causing power depression. On the other hand, results from the PI method show ambiguous due to the high correlation among features in the reactor core system.

However, both methods agree on features at the top level are evaluated to the most important since CRUD deposit accumulates at the top of the core.

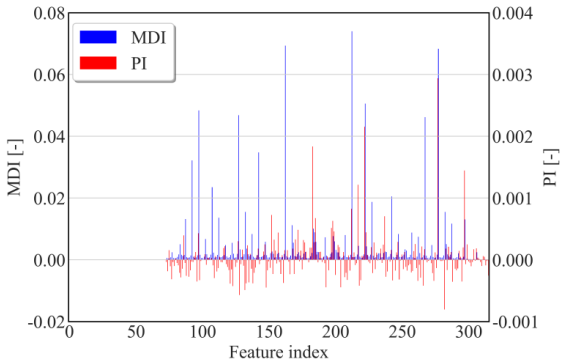


Fig 4. Feature importance estimated by mean decreased impurity (MDI) and permutation importance (PI)

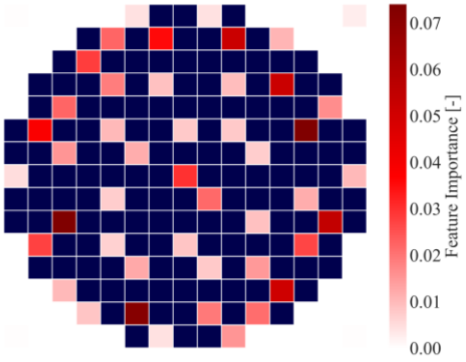


Fig. 5. Axially integrated feature importance of ICI and ex-core detector signal evaluated by MDI

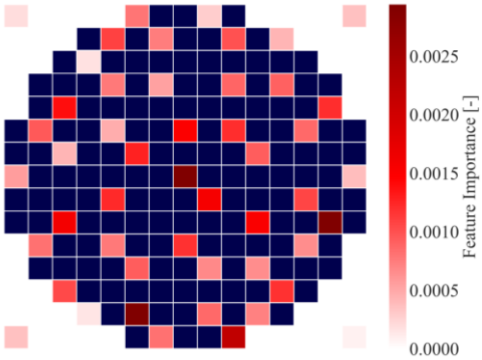


Fig. 6. Axially integrated feature importance of ICI and ex-core detector signal evaluated by PI

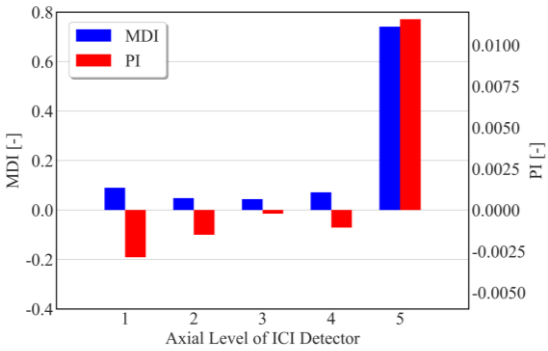
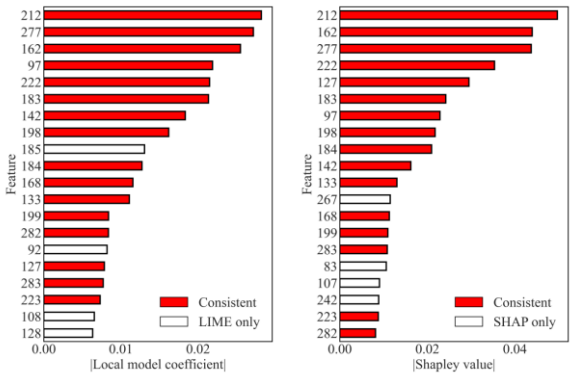


Fig 7. Radially integrated feature importance of 5 different axial levels of ICI detector

4.3 Interpreting Model Predictions by Instance

This section explores the interpretability of an ML model using two popular explainability methods: LIME and SHAP. LIME and SHAP are both post-hoc explainability techniques that can be used to help understand how the model is making its predictions. The results of the analysis using these methods are presented in this section, along with a comparison of their effectiveness in explaining the model's predictions on individual instances of the CIPS data. To accomplish this, labeled CIPS data with a label of '1' and normal data with a label of '0' were used.

The features were ranked based on the absolute value of local model coefficient and Shapley value. Out of 316 features, top ranked 20 were identified as the most important features by each interpreting methods as shown in Figure 8 and 10, for the prediction of 'CIPS' and 'Normal', respectively. Among these 20 important features, 16 (14 features for prediction on 'Normal') were identified as important by both methods for ML model's prediction of 'CIPS', showing consistency of explanation between the two methods. The important features consist of ICI signals, and their location are shown in Figure 9 and Figure 11. All of them are located at top (5th level) or bottom (1st and 2nd level) of fresh fuel assemblies at periphery and center of the core, where the appearance of CRUD affects most.



(a) LIME (b) SHAP

Fig. 8 Top 20 important features contributing for ML to predict as 'CIPS' evaluated by LIME and SHAP

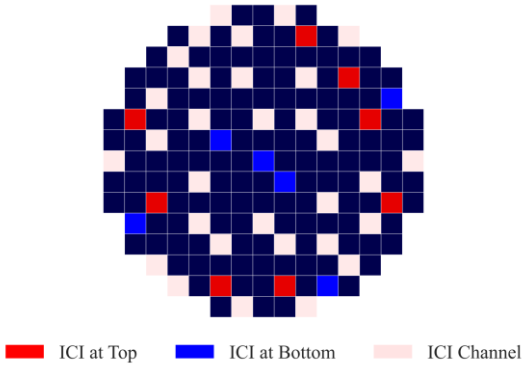
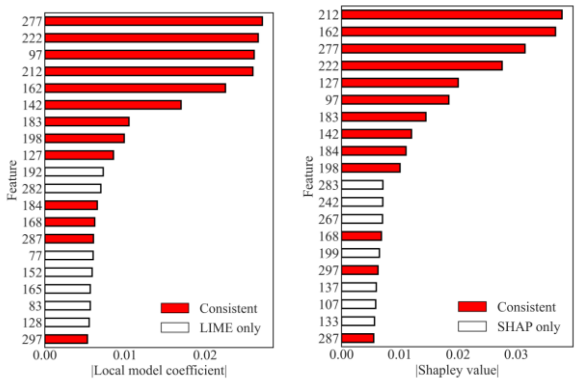


Fig. 9. Location of ICI channel whose signals are evaluated as important by both LIME and SHAP on ML model's prediction on 'CIPS'.



(a) LIME (b) SHAP

Fig. 10. Top 20 important features contributing for ML to predict as 'Normal' evaluated by LIME and SHAP

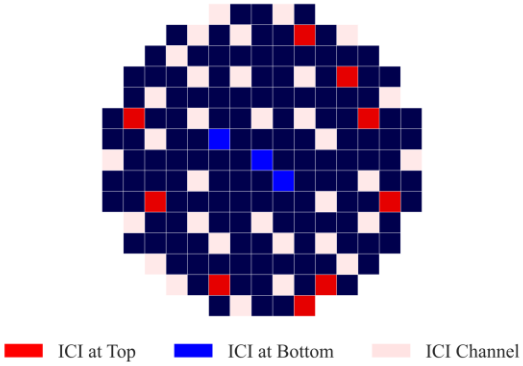


Fig 11. Location of ICI channel whose signals are evaluated as important by both LIME and SHAP on ML model's prediction on 'Normal'.

5. Conclusions

This study utilized different explainable AI (XAI) techniques to examine the feasibility of applying them on a machine learning (ML) framework for detecting anomalies in a reactor core. Specifically, CRUD-induced power shift (CIPS) was used to evaluate the effectiveness of the XAI methods, taking account into its nature, it is known that CRUD is more likely to appear in the top region of the core and near high power fuel assemblies such as fresh fuel assembly, resulting in local power deviation at those regions and their opposite regions.

A random forest classifier (RF) trained with simulation-based dataset was employed as the CIPS predictive model with over 90% accuracy. Local and global analyses of feature importance were conducted to interpret the ML model's prediction. Among the input parameters, control rod position and ex-core detector signals are relatively insignificant according to all methods. The global feature importance analysis identifies that features from the top of the core are important, while local interpreting methods identify signals from both the top and bottom of the core as important. This is because LIME and SHAP can handle correlated features by considering their interactions and dependencies, thus identifying features at the bottom region as important, which is proper understanding of the nature of CIPS. MDI, LIME, and SHAP agree that features near the fresh fuel region where CRUD appears are important.

This study explored feasibility of using XAI methods on ML-based core anomaly system. The XAI methods were able to explain the nature of CIPS well, even without information related to the state. In future studies, further investigation on the explainability of the ML model for nuclear reactor core anomaly detection will be conducted to improve its reliability and will be expanded to various core anomaly situations.

Acknowledgement

This work was supported by KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2022-Tech-13)

REFERENCES

[1] Kim, H., Yun, D., Shin, H., Moon, S., & Lee, D. (2020, July). Feasibility study on machine learning algorithm in nuclear reactor core diagnosis. In Proceedings of the Transactions of the Korean Nuclear Society Virtual Spring Meeting, Korea (online)
[2] Oh, Y. Kim, H., Lee, D. and Kim, S. (2021). "Simulation-based Anomaly Detection in Nuclear Reactors", Journal of the Korean Institute of Industrial Engineers, 47.2: 130-143.
[3] Kim, H., Jo, Y. and Lee, D. (2021). Feasibility study on AI-based prediction for CRUD induced power shift in PWRs., In Proceedings of the Transactions of the Korean Nuclear Society Virtual Autumn Meeting, Korea (online)
[4] Park, J. et Al. (2020). RAST-K v2—Three-Dimensional Nodal Diffusion Code for Pressurized Water Reactor Core Analysis. Energies, 13(23), 6324.
[5] Adadi, A., Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.
[6] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
[7] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth International Group.
[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2021). Permutation feature importance: A simple and reliable method to improve black box models. Interpretable Machine Learning, 1(1), 6-23.
[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
[10] Lundberg, S. M. and Lee, S. I., "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, 2017.