Proceedings of the Reactor Physics Asia 2023 (RPHA2023) Conference
Gyeongju, Korea, October 24-26, 2023

A Comparative Analysis of eXplainable AI Techniques for Nuclear Reactor Core Anomalies

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*

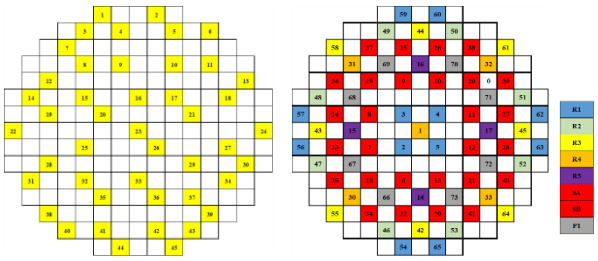
Keywords: eXplainable AI, machine learning, anomaly detection, core anomaly

1. Introduction

As the AI industry has grown, machine learning (ML) has been integrated into nuclear reactor cores for anomaly detection [1,2,3]. Despite the potential, the "black box" nature of ML models poses challenges for their adoption in the nuclear industry. To overcome this, eXplainable AI (XAI) techniques, designed to clarify the decisions of these complex models, have risen in importance. A previous study explored the feasibility of utilizing various XAI methods, including Mean Decreased Impurity (MDI), Permutation Importance (PI), Local Interpretable Model-Agnostic Explanation (LIME), and Shapley Additive Explanation (SHAP) for a model predicting axial offset anomaly [4,5,6,7,8,9]. This research extends that evaluation by applying these methods to diverse nuclear reactor core anomaly scenarios, including control rod mis-location, inlet temperature asymmetry, and cross-wiring of ICI detector signals. To concentrate solely on the explanatory power of the XAI methods, model-related uncertainties was minimized by simplifying the dataset. The effectiveness of each XAI approach was evaluated by matching their top 15 identified features against a baseline set, established based on relative feature differences.

2. Nuclear Reactor Core Anomaly Scenarios

This study utilized a dataset generated from a nuclear reactor core simulation under the core condition of the beginning of the cycle at hot full power and all-rod-out of a cycle of OPR-1000 type reactor core using RAST-K [10]. This dataset provides a snapshot of nuclear operating parameters, consisting of 225 ICI signals distributed across 45 fuel assemblies at 5 axial levels. For all dataset representing ICI signals, uniform distribution of random noise between ±0.5% was added.



(a) ICI Channels (b) CEA Assemblies

Fig. 1. Position of ICI and CEA of OPR-1000 Type Core

2.1 Control Rod Mis-location

The “Control Rod Mis-location” scenario represents a situation where a control rod position of a CEA bank is deviated from control rod assemblies in the same subgroup. To simplify the problem, CEA#6 was chosen to be deviated. Data was labeled as “anomaly” if the deviation exceeds 5 steps (1.905 cm/step). Otherwise, it was labeled as “normal”. A known consequence of control rod deviation is the power reduction in fuel assemblies near the control rod, particularly in the upper region. This decrease serves as a distinct characteristic, enabling the ML model to differentiate this situation from regular conditions. For a detailed comparison of local interpretations, two representative samples were selected: one with a 6 steps deviation and another with a 34 steps deviation. The most significant variables were selected based on the deviation of their signals from the normal core. Fig. 2 shows the top 15 ICI signals according to the baseline: (a) the average of datasets labeled as "anomaly". (b) The sample with a 6 steps deviation. (c) The sample with a 34 steps deviation." The feature index P33H5 refers to the detector signal corresponding to the 33rd ICI fuel assembly, as shown in Fig. 1(a), and the 5th axial level, which is top level.

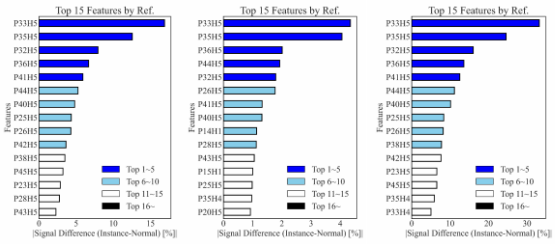
2.2 Inlet Temperature Asymmetry

The "Inlet temperature asymmetry" refers to an imbalance in quadrant inlet temperature, causing a quadrant power tilt. For simplification of dataset, the deviation of inlet temperature appeared between the left-upper section of the core and the rest. During dataset generation, temperature was sampled and when the temperature deviation exceeded 0.6 °C, it was labeled as “anomaly”, and "normal" otherwise. The effect of this temperature deviation was widespread within the core. Thus, the deviations in individual ICI signals were less pronounced compared to the more localized impact of control rod mis-location. For detailed examination, two samples were selected with temperature deviations of 1.0 °C and 2.7 °C, respectively. Fig. 3 presents the top 15 features based on the baseline.

2.3 ICI detector cross-wiring

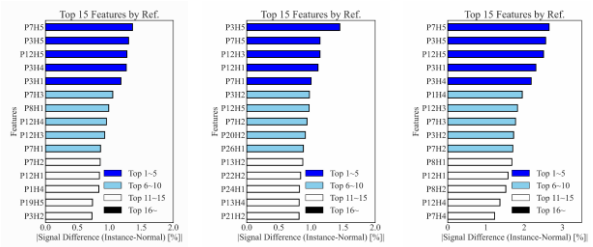
The "ICI detector cross-wiring" scenario describes a situation where two ICI detectors are mistakenly

connected. In this case, the signal wire of one ICI channel is plugged into the terminal of another detector. For a simplified dataset on ICI cross-wiring, ICI 3 was chosen as the default, while another ICI channel was selected from those located in the upper left region. The ICI cross-wiring impacts only the features of the crossed ICI channels. Contrasting other situations, ICI cross-wiring creates point anomalies, causing sudden, significant feature value shifts. Two instances were analyzed: a cross-wiring of ICI 3 with ICI 8 and another with ICI 7.



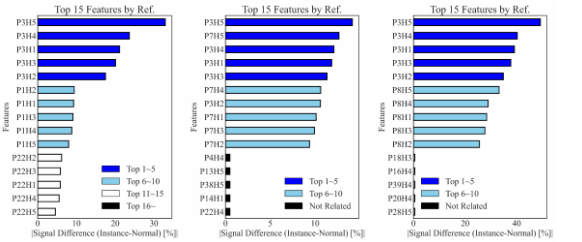
(a) Average (b) 6-step (c) 34-step

Fig. 2. Top 15 features for control rods deviation by baseline



(a) Average (b) 1.0℃ (c) 2.7 ℃

Fig. 3. Top 15 features for T inlet deviation by baseline



(a) Average (b) ICI 3↔ICI 8 (c) ICI 3↔ICI 7

Fig. 4. Top 15 features for ICI cross-wiring by baseline

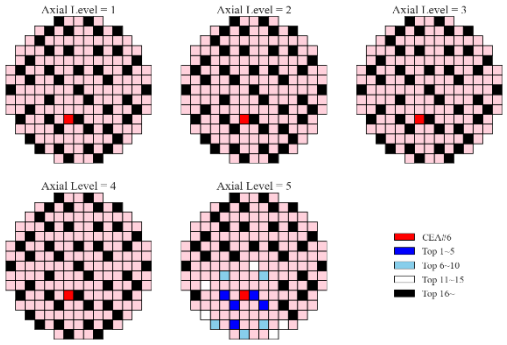


Fig. 5. Position of Top 15 features for control rod deviation (average) by baseline

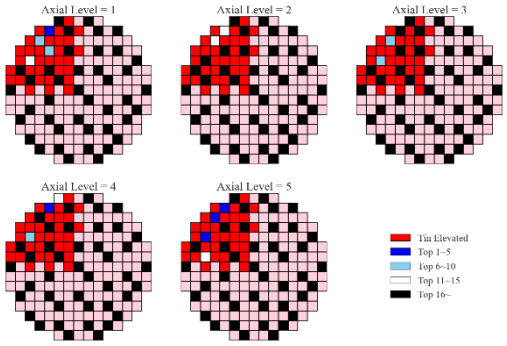


Fig. 6. Position of Top 15 features for coolant temperature deviation (average) by baseline

Table summarizes the core abnormal scenarios in terms of the distinction of features between normal and abnormal condition.

Table I: Characteristics in Data for Core Anomalies

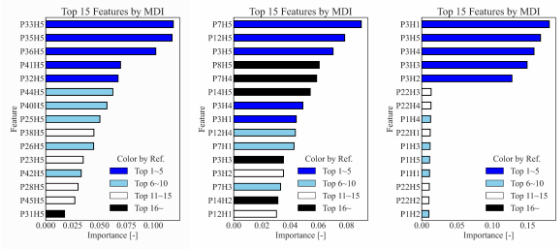
Anomaly Type	Affected Region	Max. of ICI deviation	STD of ICI deviation	Value Change
CR Mis-location	Near CR, Upper Core	16.87%	3.88%	Gradual
T Inlet Asymmetry	Global	1.36%	0.21%	Gradual
ICI Cross-wiring	Crossed ICI Channels	32.91%	8.18%	Rapid

3. Results

This study's analysis focus on the ICI detectors most affected by each anomaly with the primary features identified by each XAI method for those scenarios to assess the performance of XAI methods under different conditions, considering their approaches. A Random Forest classifier was utilized in this study.

3.1 Mean Decreased Impurity

Mean decreased impurity (MDI) [5] is method that evaluate a feature importance by measuring how much a feature reduces the impurity during a Tree-based ML's training process. As it evaluates throughout the training across the full dataset, MDI provides a global view of a feature's significance. It indicates the degree of emphasis a tree-based ML model places on a particular feature during prediction. Given that MDI captures global feature importance, the top 15 features by average effects were examined. Fig. 7 presents the comparison of top 15 features between MDI and the baseline. It was observed that the training was more concentrated on features which is more sensitive to the abnormal condition. Since tree-based learning with Random Forest implies randomness during training process, rank can vary when feature distinctions are subtle. Moreover, due to the default selection of ICI 3 for ICI cross-wiring scenario, the model was biased toward that channel.

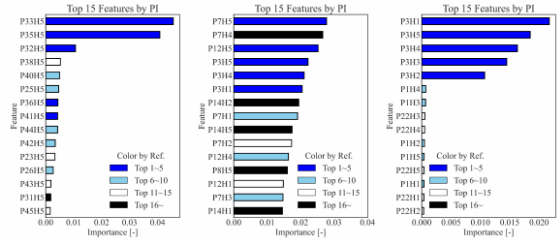


(a) CR deviation (b) Inlet asymmetry (c) ICI cross-wiring

Fig. 7. Comparison of Top 15 features: MDI vs Baseline

3.2 Permutation Importance

Permutation Importance (PI) [6] evaluates feature importance by observing the performance difference after shuffling feature values in the test dataset. Consequently, choosing an appropriate performance metric, which corresponds with the nature of the ML problem, is essential for determining PI. PI is model-agnostic, meaning it can be applied to any machine learning model. However, this method operates under the assumption that each feature is independent. As such, it may struggle when addressing the importance of correlated features. In cases where features are interrelated, a model might still generate accurate predictions even if one feature is shuffled, drawing on information from the other correlated feature. Thus, while PI may consistently identify the top 3 features in alignment with the baseline, it may struggle to accurately assess the significance of the remaining features.

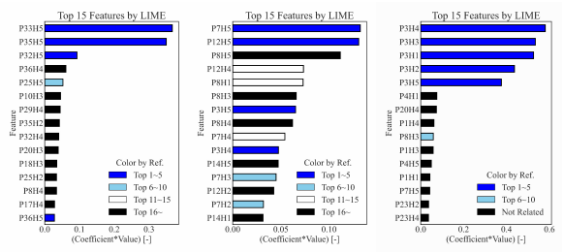


(a) CR deviation (b) Inlet asymmetry (c) ICI cross-wiring

Fig. 8. Comparison of Top 15 features: PI vs Baseline

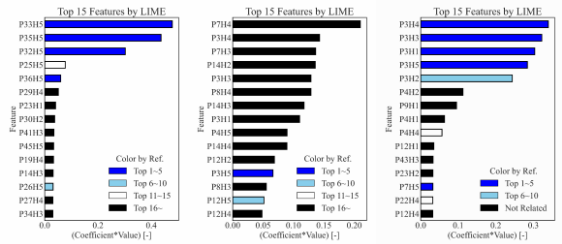
3.3 Local Interpretable Model-agnostic Explanations

Local interpretable model-agnostic explanations (LIME) [7] is a XAI method that explains any machine learning model's predictions through a surrogate model around a specific data point of interest. Key steps in LIME's process involve generating synthetic samples, predicting using the black-box model, and training a surrogate to highlight main features for that data point, assuming local linear approximation and features independence. When features correlates, it can identify most important features, but may overestimate their contribution and underestimate others. LIME struggles when a feature's value is close to noise, as its perturbation can result in information loss by merging with the noise. In cases like ICI cross-wiring, swift changes in feature values correspond to changes in data point classes. Features near these anomalies are rarely seen in training, so changing them might not give varied results, making it hard to judge their importance. LIME's correct pick of the top 5 features for ICI 3 is due to the model's bias from a simplified dataset.



(a) CR deviation (b) Inlet asymmetry (c) ICI cross-wiring

Fig. 9. Comparison of Top 15 features for a sample distant from class boundary: LIME vs Baseline

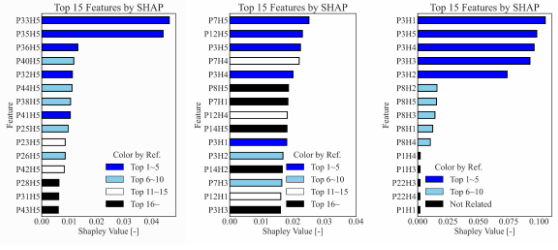


(a) CR deviation (b) Inlet asymmetry (c) ICI cross-wiring

Fig. 10. Comparison of Top 15 features for a sample close to class boundary: LIME vs Baseline

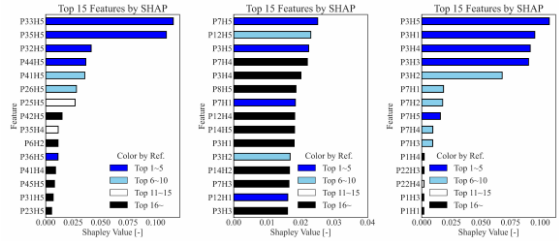
3.4 Shapley Additive Explanation

Shapley Additive Explanation (SHAP) [8] uses game theory to interpret individual machine learning predictions by allocating importance values to each feature based on its contribution. In this study, TreeSHAP [9], a variant optimized for tree-based ensemble models, was utilized to take advantage of its computational efficiency. In local interpretation, SHAP tends to align more closely with the baseline than LIME because it considers every possible feature combination to determine each feature's impact. This approach suitable for dataset with correlated features, like those in core anomaly detection. Additionally, it can identify important features in situation like ICI cross-wiring, where rapid shifts in feature values with changes in data point classes.



(a) CR deviation (b) Inlet asymmetry (c) ICI cross-wiring

Fig. 11. Comparison of Top 15 features for a sample distant from class boundary: SHAP vs Baseline



(a) CR deviation (b) Inlet asymmetry (c) ICI cross-wiring

Fig. 12. Comparison of Top 15 features for a sample close to class boundary: SHAP vs Baseline

Table II summarizes the characteristics of each XAI techniques in this study. Table III presents the count of consistently evaluated as important features across various XAI methods and baseline.

Table II. Summary of XAI techniques in the comparative analysis on nuclear reactor core anomaly detection

XAI Methods	Model Dependency	Consistency	Global/Local	Basis
MDI	Tree	Stable	Global	Tree-based Foundation
PI	Model-agnostic	Stable	Global	Statistical
LIME	Model-agnostic	Vary	Local	Local Approx.
SHAP	Model-agnostic	Consistent	Local/ Global	Cooperative Game Theory

Table III: Consistent evaluation of important features by various XAI techniques and baseline.

Anomaly Type	Sample	MDI	PI	LIME	SHAP
CR mis-location	6 steps	14	14	6	9
	34 steps			5	12
Inlet Asymmetry	1.0 °C	10	10	2	6
	2.7 °C			9	10
ICI cross-wiring	ICI 3↔8	15	15	6	10
	ICI 3↔7			6	10
Consistent		49	49	34	57
Important by Baseline		50	50	80	80

Computational complexity and cost are summarized in Table III. The complexity of each method is defined with the number of tree models(T), average depth of tree-based models(D), number of test dataset(N), number of features(M), number of permutations(P), number of synthetic samples(S), complexity of a surrogate model (C) and average number of leaves (L) in each decision tree.

Table IV: Computational complexity and time [sec] of each method on core anomaly types

	MDI	PI	LIME	SHAP
O	$O(TD)$	$O(TDNMP)$	$O(TDS+SMC)$	$O(TLM^2)$
CR mis-location	0.109	68.732	20,121.776	0.126
Inlet Asymmetry	0.109	87.957	20,587.080	0.409
ICI cross-wiring	0.108	65.887	20,914.081	0.047

4. Conclusion

This research examined various XAI techniques, including mean decreased impurity (MDI), permutation impurity (PI), LIME, and SHAP, regarding reactor core anomaly detection for scenarios like control rod mis-location, inlet temperature asymmetry, and ICI cross-wiring. To isolate the explanatory power of each method, the dataset was simplified, minimizing uncertainties from model inaccuracies. A baseline reference was established by measuring deviations between normal and abnormal datasets. Both MDI and PI offered insights into global feature importance. MDI specifically highlighted features the ML model concentrated during training. PI correctly identifies top features but tends to overestimate their importance, especially when applied to interrelated features. Both LIME and SHAP offer local explanations. LIME, which perturbs features near the target instance based on assumptions of local linear approximation and feature independence, struggles with highly correlated features and point anomalies that exhibit drastic feature changes. In contrast, SHAP more consistently matches with the baseline outcomes as it takes into account all possible combinations of features.

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). Feasibility study on machine learning algorithm in nuclear reactor core diagnosis. KNS Spring Meeting, Korea (online).
[2] Kim, H., Jo, Y., & Lee, D. (2021). Feasibility study on AI-based prediction for CRUD induced power shift in PWRs. KNS Autumn Meeting, Korea (online).
[3] Oh, Y., Kim, H., Lee, D., & Kim, S. (2021). Simulation-based Anomaly Detection in Nuclear Reactors. Journal of the Korean Institute of Industrial Engineers, 47(2), 130-143.
[4] Kim, H., Moon, S. R., & Lee, D. (2023). Feasibility Study of an Explainable AI-based Anomaly Detection for Nuclear Reactor Core Operation in PWRs. KNS Spring Meeting, Jeju, South Korea, May 18-19.
[5] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Wadsworth International Group.
[6] 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.
[7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.
[8] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.
[9] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. arXiv preprint arXiv:1802.03888.
[10] Park, J., et al. (2020). RAST-K v2—Three-Dimensional Nodal Diffusion Code for Pressurized Water Reactor Core Analysis. Energies, 13(23), 6324.