Comparative Analysis of XAI Methods for Critical Parameter Identification and Feature Compression in SMR Abnormal Diagnosis

Young Ho Chae [a*], Seung Geun Kim [a], Seo Ryong Koo [a]
[a] *Korea Atomic Energy Research Institute, 111 Daedeok-daero 989 beon-gil, Daejeon, 34057*
[*] *Corresponding author: yhchae@kaeri.re.kr*

Keywords : Feature Compression, Abnormal Diagnosis, eXplainable Artificial Intelligence

1. Introduction

Artificial neural networks (ANNs) have transformed data-driven approaches in many fields, including the nuclear industry. ANN models have been applied to tasks like equipment diagnostics and early detection of plant conditions. However, implementing ANNs for nuclear power plant(NPP) simulations is challenging due to the vast number of variables involved - over 14,000 in some cases.

This abundance of data provides a comprehensive view of plant state but also introduces over redundancy. Therefore, feature selection and dimensionality reduction are crucial. Identifying key variables can improve efficiency and accuracy in condition diagnosis while adhering to the principle of model parsimony.

Reducing variables offers several benefits beyond simplification:

1. It aligns with regulatory requirements for verifying AI systems in nuclear facilities.
2. It enhances cybersecurity by allowing focused protection of essential data.
3. It improves result explainability to operators, aiding swift decision-making.
4. It boosts operational efficiency by reducing computational overhead.

This study evaluates various feature extraction methods for nuclear plant condition diagnosis, including explainable AI (XAI) techniques, feature compression algorithms, and statistical approaches. The goal is to assess how these methods impact ANN performance when trained on only the extracted key variables.

The research takes an approach by using XAI not just to explain model decisions, but as a tool to identify important variables from a fully-trained ANN. Methods explored include gradient-based selection, layer-wise relevance propagation, DeepSHAP, integrated gradients, LIME, saliency maps, and DeepLIFT. Feature compression techniques like variational autoencoders, UMAP [1-8], and PCA are also investigated.

This approach allows for a comprehensive evaluation of different feature selection and compression techniques in nuclear plant condition diagnosis. By comparing LSTM networks trained on various subsets of variables, the study assesses the potential for reducing diagnostic system complexity without compromising accuracy or reliability.

The findings aim to provide insights into effective variable selection approaches for nuclear plant simulations, potentially enhancing the efficiency and accuracy of condition diagnosis systems. This could contribute significantly to the safety and operational effectiveness of nuclear power plants.

2. Experiment

2.1 Dataset Preparation

2.1.1 Target Simulator

To assess the effectiveness of our feature compression methods, we employed the IAEA's integral pressurized water reactor (iPWR) simulator, created by Tecnatom in 2017. The simulator is designed to model and examine Small Modular Reactor (SMR) behavior, with a particular focus on the iPWR design. The iPWR's unique feature is the incorporation of primary circuit components within the reactor pressure vessel, a design choice aimed at boosting safety and reliability by eliminating the need for external primary circuit piping.

2.1.2 Scenarios (Abnormal/Emergency)

We collected 35 distinct scenarios, including 26 abnormal situations and 9 emergency events. The 26 abnormal scenarios were broadly categorized as follows:

1. Feed water system issues:
   - Malfunctions in feed water pumps (#1, #2)
   - Feed water system pipeline rupture
   - Feed water control valve closure
2. Automatic depressurization system (ADS) problems:
   - ADS valves (#1, #2, #3) becoming stuck in an open position
3. Decay heat removal system complications:
   - Decay heat removal inlet valves (#1, #2) becoming stuck open
4. Steam generator irregularities:
   - Steam generator control valves (#1, #2) stuck in open or closed positions
   - Steam generator isolation valves (#1, #2) stuck open
5. Main steam system difficulties:
   - Main steam relief valves (#1, #2) stuck open

- Main steam isolation valves (#1, #2) stuck open
- Main steam control valve stuck closed
- Main steam turbine isolation valve stuck closed
- Main steam bypass turbine valve stuck open

6. Vacuum-related issues:
   - Loss of containment vacuum
   - Loss of condenser vacuum

To ensure a thorough analysis, we gathered 30 datasets for each abnormal scenario, varying the malfunction severity.

The 9 emergency scenarios were grouped into three primary categories:

1. Loss of coolant accident (LOCA):
   - Issues with automatic depressurization system (ADS) valves (#1, #2, #3)
   - Reactor core vessel safety valve stuck open
   - Reactor core vessel relief valve stuck open
2. Steam generator tube rupture (SGTR):
   - Steam header break
   - Steam generator tube (#1, #2) rupture
3. Main steam line break (MSLB):
   - Rupture in the main steam line

2.2 Experiment Design

Two different neural networks are designed to estimate the performance of the feature compression methods. Each feature compression methods compress features from 116 to 20. Therefore, the first neural network for the baseline has 116 input neurons, and the other network has 20 input neurons. The detailed designs are as follows.

- Neural network architecture
  - Structure: LSTM network
  - Input size: 116 (corresponding to the number of measurements), 20 (corresponding to the compressed features)
  - Output size: 35 (corresponding to the number of scenario classes)
  - Number of layers: 4
  - Number of neurons in each hidden layer: 32
  - Loss function: Cross-entropy
  - Optimizer: Adam

Also, experiment procedures can be described as follows.

Let $X \in \mathbb{R}^{n \times m}$ be the input dataset, where n is the number of samples and m = 116 is the number of state variables.

Let $Y \in \mathbb{R}^{n \times k}$ be the corresponding labels, where k is the number of classes (abnormal and emergency situations).

1. Initial LSTM training: $f_{LSTM}: X \rightarrow Y$ Where f_{LSTM} represents the LSTM model trained on all 116 variables.
2. Feature selection methods: For each XAI method M_i (i = 1, …, 7 for the different XAI methods): $S_i = M_i(f_{LSTM}, X) \in \mathbb{R}^{20}$ Where S_i represents the set of 20 most important variables selected by method M_i.
3. Feature compression methods:
   - Variational auto encoder: $E_{VAE} : \mathbb{R}^{116} \rightarrow \mathbb{R}^{20}$, $D_{VAE} : \mathbb{R}^{20} \rightarrow \mathbb{R}^{116}$, $Z_{VAE} = E_{VAE}(X) \in \mathbb{R}^{(n \times 20)}$ Where E_{VAE} and D_{VAE} are the encoder and decoder of the VAE, respectively.
   - UMAP: $f_{UMAP} : \mathbb{R}^{116} \rightarrow \mathbb{R}^{20}$, $Z_{UMAP} = f_{UMAP}(X) \in \mathbb{R}^{(n \times 20)}$
   - PCA: $f_{PCA} : \mathbb{R}^{116} \rightarrow \mathbb{R}^{20}$, $Z_{PCA} = f_{PCA}(X) \in \mathbb{R}^{(n \times 20)}$
4. Training LSTM models with reduced features: For each feature selection or compression method $j: X_j \in \mathbb{R}^{n \times 20}$ (either selected variables or compressed representations) $f_{LSTM_j}: X_j \rightarrow Y$
5. Performance evaluation: For each model f_{LSTM_j} and the original f_{LSTM} : $Accuracy_j = \sum(\hat{y}_i == y_i)/n$ Where $\hat{y}_i$ is the predicted label and y_i is the true label for sample i.

3. Results and Discussion

3.1 Experiment results

Fig. 1 depicts the overall experimental results, showing accuracy trends across 100 epochs for various methods. The findings can be categorized into three groups:

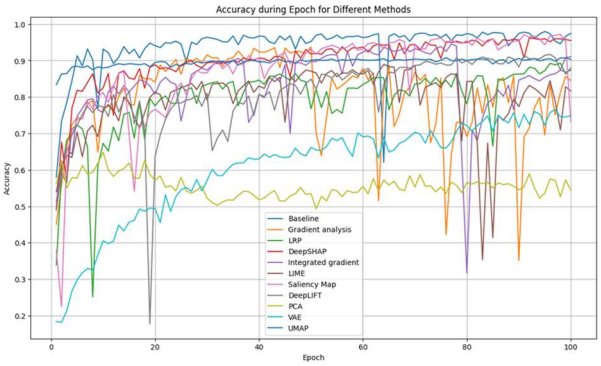


Fig. 1 Accuracy during Epochs

1. High-Performing Models: DeepSHAP, UMAP, and Saliency map consistently outperformed other methods, achieving peak accuracies above 0.90. Their success stems from their ability to capture non-linear relationships and model-specific importances. DeepSHAP and saliency map benefit from their direct link to the model's decision-making process. UMAP

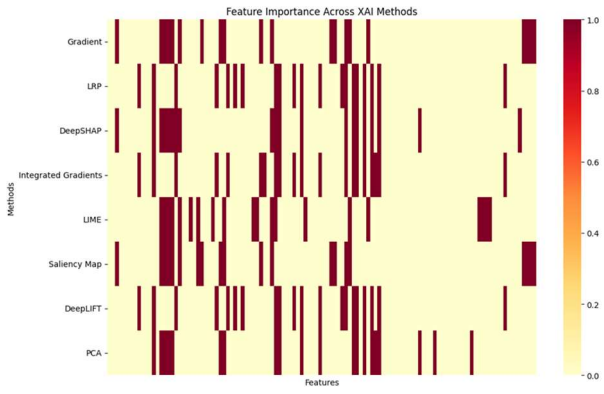
excels in maintaining the data's topological structure, demonstrating remarkable stability across epochs. This suggests robust feature representation and effective preservation of data relationships. UMAP's exceptionally high initial accuracy and consistent performance indicate its proficiency in dimensionality reduction while retaining crucial data structure. This preservation of essential variable relationships likely enables more efficient learning by the neural network, allowing it to swiftly achieve and maintain high diagnostic accuracy. UMAP's effectiveness in capturing complex, non-linear relationships in a lower-dimensional space appears to provide an ideal foundation for the diagnostic neural network, facilitating quick distinction between various plant conditions from the early stages of training.

2. Moderate Performers: Gradient Analysis, Integrated Gradient, LRP, DeepLIFT, and LIME showed suitable performance but varied in stability. Gradient-based methods (Gradient Analysis, Integrated Gradient) exhibited sensitivity to the model's current state, leading to instability. LRP and DeepLIFT demonstrated more consistent performance, likely due to their propagation-based approach. LIME's local approximation strategy resulted in moderate but unstable performance.

3. Low Accuracy Models: PCA and VAE struggled to match the performance of other methods. PCA's linear nature presents a significant limitation in this complex domain. While VAE showed some improvement, it may be hindered by the challenges of simultaneously optimizing both encoder and decoder networks for this specific task.

3.2 Frequently Selected Features

Fig. 2 depicts the selected features from the each compression methods.



Certain features consistently emerge across multiple methods, highlighting their overall significance:

1. Feature 69 (Total flow in pressure header): Identified by 5 out of 8 methods (LRP, DeepSHAP, Integrated Gradient, DeepLIFT, and PCA).
2. Features 15 (ADS1 valve flow), 16 (ADS2 valve flow), and 14 (Relief valve flow): Each selected by four methods, mainly gradient-based and saliency-based approaches.
3. Features 44 (Pump2 speed), 67 (Steam flow rate line2), 64 (% open valve turbine), and 71 (Steam pressure line2): Each chosen by 3–4 methods across various categories.

The recurring selection of these features indicates their vital role in differentiating between normal and abnormal plant conditions.

High-performing methods (DeepSHAP, Saliency Map, DeepLIFT) share some commonalities in their feature selection:

1. Features 15, 16, and 17 (ADS1, 2, 3 valve flow): Consistently chosen by these methods, suggesting their strong relevance to the model's decision-making process.
2. Feature 69 (Steam flow to turbine): Selected by both DeepSHAP and DeepLIFT, but not by saliency map, indicating its importance in propagation-based methods.

Feature 69 is crucial for distinguishing between abnormal and emergency situations, while Features 15, 16, and 17 are essential in diagnosing automatic depressurization system abnormalities and identifying LOCAs caused by ADS valve issues.

In contrast, PCA, which showed suboptimal performance, selected unique features (e.g., 84 (Flow charge), 98 (Temperature difference hotleg-coldleg)) not commonly chosen by other methods. This discrepancy might explain PCA's reduced effectiveness in capturing relevant diagnostic information.

3.3 Insights

The comprehensive analysis of various feature selection methods in nuclear power plant diagnostics reveals several key insights. The consistent identification of certain features (such as valve openings, safety system flows, and key temperature indicators) across multiple high-performing methods underscores their critical importance in accurately diagnosing plant conditions. This consensus provides a strong basis for prioritizing these parameters in monitoring systems.

XAI-based methods, particularly saliency map and DeepSHAP, demonstrate superior performance in distinguishing between different abnormal situations, highlighting the importance of non-linear feature interactions in nuclear plant diagnostics. The ability of these methods to capture complex relationships among variables is crucial for developing more accurate and responsive diagnostic systems.

Dimension reduction, as shown in this study, conducts a significant role in nuclear power plant data analytics.

Identifying a subset of highly informative features can significantly enhance the efficiency and interpretability of diagnostic models. This reduction in dimensionality offers several key advantages in the nuclear industry:

1. Cybersecurity Enhancement: Reducing monitored variables minimizes potential cyber threat attack surfaces, improving overall plant information system security.
2. Improved Human-Machine Interface: A reduced set of key variables provides more comprehensible and actionable information for human operators, potentially leading to faster and more accurate decision-making in critical situations.
3. Computational Efficiency: Focusing on fewer crucial variables can lead to more streamlined and swifter-executing diagnostic models, enabling real-time analysis and quicker response to developing situations.
4. Data Management Optimization: Prioritizing key variables can guide more efficient data collection, storage, and processing strategies, potentially reducing infrastructure costs and improving system responsiveness.

However, it's important to note that these results can not guarantee that methodologies like DeepSHAP will perform well on all diagnostic datasets. Instead, the results demonstrate the potential performance benefits of using various XAI or feature compression preprocessing modules. These methods can be effective not only for extracting useful variables but also for eliminating variables deemed unimportant by any methodology.

The feature selection and compression methodologies explored in this study show potential as effective pre-processing modules for developing robust nuclear power plant diagnostic systems. By extracting the most relevant variables from the extensive plant parameters, these methods can enhance both the accuracy and efficiency of artificial intelligence based diagnostic systems.

4. Conclusion

The study analyzes feature selection methods, particularly XAI techniques, to improve nuclear power plant condition diagnostics. It demonstrates the effectiveness of dimension reduction in capturing critical information from plant parameters, enhancing efficiency and accuracy in detecting abnormal situations.

The research applies XAI methods for feature selection, with saliency map and DeepSHAP showing superior performance in identifying key features. UMAP emerges as a promising dimensionality reduction method, outperforming traditional approaches like PCA.

Key features identified include valve openings, safety system flows, and critical temperature indicators. This consensus provides valuable insights for prioritizing monitoring efforts in plant operations.

Feature compression presents numerous advantages, including potential enhancements in cybersecurity, human-machine interfaces, computational efficiency, and data management. The methodologies investigated demonstrate significant potential as effective pre-processing modules for robust diagnostic systems, thereby improving both accuracy and efficiency.

Also, the research highlights the limitations of traditional linear methods and underscores the importance of non-linear feature interactions in developing effective diagnostic tools.

Acknowledgement

This research was supported by the National Research Council of Science & Technology(NST) grant by the Korea government (MSIT) (No. GTL24031-400).

REFERENCES

[1] Binder, A., Montavon, G., Bach, S., Müller, K.-R., Samek, W., 2016. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers.
[2] Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions.
[3] Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic Attribution for Deep Networks.
[4] Shrikumar, A., Greenside, P., Kundaje, A., 2019. Learning Important Features Through Propagating Activation Differences.
[5] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.
[6] Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.
[7] Kingma, D.P., Welling, M., 2022. Auto-Encoding Variational Bayes.
[8] McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.