Mitigating Spurious Correlations for Nuclear Power Plant Fault Diagnosis Using Adversarial Debiasing

Ji Hyeon Shin *, Jae Min Kim, Seo Ryong Koo
Advanced Instrumentation & Control Research Department, Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon, 34057
**Corresponding author: jhshin0127@kaeri.re.kr*

Keywords : operator support system, fault diagnosis, artificial neural network, adversarial debiasing

1. Introduction

The integration of Artificial Intelligence (AI) into the operational support systems of Nuclear Power Plants (NPPs) has been researched for enhancing plant safety and efficiency. While research into digital twins for NPPs is expanding, current AI models are primarily being developed to provide reliable decision support to human operators by learning physical fault signatures in simulation environments. However, high predictive accuracy alone is insufficient for deployment in safety-critical nuclear systems. For AI to be deployed as a core component of software in a nuclear environment, it must also demonstrate reliability, interpretability, and robustness across varying operational conditions.

Among them, to ensure fairness and reliable decision support, this research specifically focuses on mitigating shortcut learning, a potential issue where the model might rely on sensitive attributes like Initial Conditions (IC) rather than the underlying physical fault signatures. In an NPP context, if a model learns to associate a specific operating power level with a specific fault type simply because of unbalanced training data, it may fail catastrophically when that fault occurs at a different power level. This study approach to refine the model training process, specifically through the incorporation of adversarial training designed to enhance model fairness [1,2]. By systematically mitigating the influence of sensitive attributes, we ensure that the diagnostic logic is grounded in the actual transient signatures of the reactor, thereby facilitating more trustworthy decision support for operators.

2. Backgrounds

To ensure that AI models in nuclear systems prioritize physical transients over statistical noise, it is essential to establish a theoretical framework for predictive fairness. This section introduces the model training algorithm focusing on the suppression of sensitive attribute leakage, and fairness metrics.

2.1 Adversarial Debiasing

Adversarial debiasing is a framework, designed to prevent a classifier from "leaking" information about a protected or sensitive attribute into its predictions, $\hat{Y}$. The architecture consists of two competing networks: a classifier and an adversary. The classifier aims to predict the target label, Y, by generating a latent representation. Simultaneously, the adversary attempts to predict the sensitive attribute from the classifier output or its latent features. The classifier is trained to minimize its prediction loss while maximizing the adversary loss. To achieve this, the gradient of the classifier is modified using a projection step to ensure that the updates do not contribute to the adversary success. The modified gradient is calculated as follows [2]:

$$(1)\quad \nabla_W L_{clf} - \mathrm{proj}_{\nabla_W L_{adv}} \nabla_W L_{clf} - \alpha \nabla_W L_{adv}$$

L_{clf} is the classification loss, L_{adv} is the adversarial loss, and α is a hyperparameter controlling the strength of the debiasing.

2.2 Fairness Metrics

Demographic Parity (DP) requires the likelihood of a positive prediction to be independent of the sensitive attribute. It is measured by the difference in selection rates. Equal Opportunity (EO). This focuses on the true positive rate (TPR), ensuring that the model is equally skilled at detecting a fault across all sensitive groups. To quantitatively evaluate the fairness of the model, two primary metrics are utilized:

$$(2)\quad \Delta DP = \left| P\left(\hat{Y} = 1 \middle| A = 0\right) - P\left(\hat{Y} = 1 \middle| A = 1\right) \right|$$

$$(3)\quad \Delta EO = \left| \begin{matrix} P\left(\hat{Y} = 1 \middle| A = 0, Y = 1\right) \\ -P\left(\hat{Y} = 1 \middle| A = 1, Y = 1\right) \end{matrix} \right|$$

Let Y denote the ground-truth label, $\hat{Y}$ the predicted label, A the sensitive attribute. $P(\cdot)$ denotes probability.

3. Experimental Setup

The validation of the proposed methodology requires a controlled environment where the impact of spurious correlations can be identified. Accordingly, the following sections describe the empirical framework designed to evaluate how the refined model handles intentional data biases.

3.1 Datasets

The dataset consists of simulated fault data from an NPP [3], focusing on the Malfunction (MF) scenarios under different ICs. To simulate a worst-case shortcut learning environment, the training set was intentionally biased:

- **Training Set:** Composed exclusively of IC0-MF0 (Normal state at beginning-of-life) and IC1-MF1 (Malfunction / feedwater system pipe break at middle-of-life).
- **Test Set:** Includes cross-combinations (IC0-MF1 and IC1-MF0) to evaluate whether the model truly learned the fault physics or simply memorized the IC-MF pairing.

Table I: The Configuration of Datasets

Dataset	IC	Label	Correlation
Training set	IC0 (*BOL) IC1 (*MOL)	MF0 (Normal) MF1 (Fault)	Bias
Test set	IC0 IC1 IC0 IC1	MF0 MF1 MF1 MF1	Bias Inverted bias

* BOL, Beginning-Of-Life; MOL, Middle-Of-Life;

3.2 Model Hyperparameters

The classifier is a deep multi-layer perceptron designed to extract complex features from sensor data.

- **Architecture:** 4 dense layers with 512-256-128 hidden units.
- **Adversary:** A specialized network taking the classifier's logits and true labels as input to predict the IC.
- **Optimization:** Adam optimizer with a learning rate of 0.0001 for the classifier and 0.0005 for the adversary.
- **Debiasing strength:** α is set to 9.0. This value was empirically chosen to enforce strong invariance to the sensitive attribute, which was necessary due to the intentionally biased training setup.

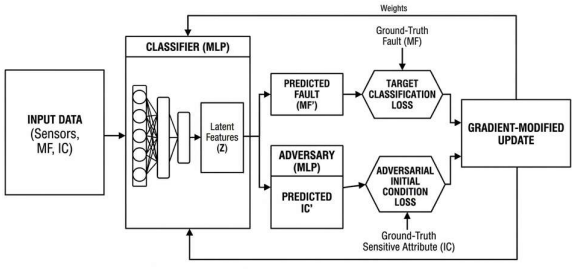

Fig. 1. The Model Training Process and Structure.

4. Results

The effectiveness of the adversarial training methodology is assessed through a comparative analysis of the model's internal representations and external diagnostic performance. We begin by examining the latent space to visually verify whether the learned features have been successfully decoupled from the initial condition biases.

4.1 Latent Space Analysis

The visualization of the latent space using t-distributed Stochastic Neighbor Embedding (t-SNE) reveals the impact of the adversarial training [4].

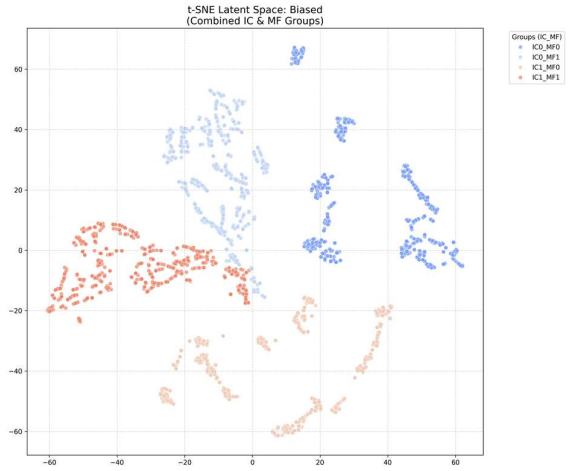

Fig. 2. The Latent Space of Biased Model.

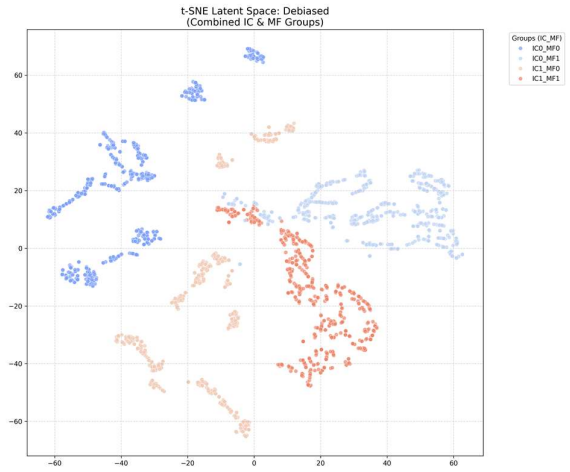

Fig. 3. The Latent Space of Debiased Model.

About the biased model, the result shows clear, segregated clusters based on the IC. Even when the target is the same, the data points are separated by their

sensitive attributes, indicating that the model is using IC as a primary feature for its decision-making. However, about the debiased model, the IC clusters are significantly more overlapped and far away. This demonstrates that the adversarial network successfully forced the classifier to discard IC-specific information, resulting in a visualization representation that focuses on the common fault regardless of its condition.

4.2 Model Performance

The numerical results demonstrate a substantial improvement in both diagnostic performance and fairness.

Table II: The Comparison Results with Fairness Metrics

	Biased model	Debiased model
Accuracy	0.8669	0.9763
TPR(IC0)	0.8348	0.9063
TPR(IC1)	1.0000	1.0000
*EOD	0.1652	0.0937
P(F|IC0)	0.4213	0.4575
P(F|IC1)	0.6859	0.5030
*DPD	0.2646	0.0456

* EOD, Equal Opportunity Difference; DPD, Demographic Parity Difference;

The accuracy increased from 86.69% to 97.63% after debiasing. This confirms that the biased model was indeed relying on shortcuts that failed on the cross-condition test data. The debiased model, by ignoring the IC, was forced to learn the robust physical features of the malfunctions, leading to better generalization.

5. Conclusions

This study demonstrates that adversarial debiasing effectively reduces shortcut learning associated with initial conditions in nuclear fault diagnosis models. The proposed framework improved diagnostic accuracy to 97.63% while substantially decreasing fairness metrics. These findings confirm that enforcing invariance to initial conditions enhances cross-condition generalization in a nuclear division. Future work will extend the approach to more diverse scenarios and fault types.

ACKNOWLEDGEMENT

This research was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (No. GTL24031-000). Also, this work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2022-00144150).

REFERENCES

[1] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M. and Lempitsky, V., Domain-adversarial training of neural networks, Journal of machine learning research, 17(59), pp.1-35, 2016.
[2] Zhang, B.H., Lemoine, B. and Mitchell, M., Mitigating unwanted biases with adversarial learning, In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp.335-340, 2018.
[3] Batra, C. Integral Pressurized Water Reactor Simulator Manual. Vienna, Austria: International Atomic Energy Agency, 2018.
[4] Van der Maaten, L. and Hinton, G., Visualizing data using t-SNE, Journal of machine learning research, 9(11), 2008.