Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26415-cd

A Bias-Reducing Framework for Enhancing Trustworthiness of AI-Based Operator Support Models in Nuclear Power Plants

Ji Hyeon Shin

Advanced Instrumentation & Control Research Department, Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon, Republic of Korea. E-mail: jhshin0127@kaeri.re.kr

Seo Ryong Koo

Advanced Instrumentation & Control Research Department, Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon, Republic of Korea. E-mail: srkoo@kaeri.re.kr

Recent advances in artificial intelligence (AI) have led to increasing research efforts to support human operators in nuclear power plants (NPPs). However, before deploying AI-based systems in actual operational environments, their trustworthiness must be thoroughly verified. Trustworthy AI requires robustness and bias mitigation to ensure reliable decision support. This study focuses on bias mitigation in AI models developed for NPP operator support and proposes a bias-reducing framework to enhance both fairness and reliability. Most existing AI-based operator support systems depend on simulation-generated data to compensate for the limited availability of real operational data. However, even when simulation diversity is considered, the individual characteristics of plant systems and components make it difficult to predict or eliminate potential bias across various operational scenarios. Furthermore, a domain gap between the simulated training domain and the actual operational domain can amplify bias and degrade model generalization performance. To address these challenges, this study proposes a bias-reducing framework that considers a fairness point into the training process of AI-based operator support models for NPPs. The proposed training stage in this framework extends beyond the main task structure to include the sub-task structure designed for domain adaptation and bias mitigation strategies in the neural network. Through this approach, the study establishes a methodological foundation for developing trustworthy AI-based operator support models that can reliably support operator decision-making in safety-critical environments such as NPP operations.

Keywords: Trustworthy AI, bias mitigation, adversarial training, nuclear power plant, operator support system.

1. Introduction

As research into Artificial Intelligence (AI) for Nuclear Power Plants (NPPs) expands, ensuring the reliability and robustness of these models has become critical for future safety-critical applications. Trustworthy AI requires robustness and bias mitigation to ensure reliable decision support. In NPP environments, most AI-based systems depend on simulation-generated data due to the scarcity of real operational data. However, simulation diversity is often limited, and individual plant characteristics can introduce potential biases across various scenarios. Furthermore, the domain gap between simulated training environments and actual operational domains can amplify these biases, degrading the model's generalization performance. Therefore, A significant challenge in data-driven diagnostics is shortcut learning, where models rely on spurious correlations—such as operational history or specific initial conditions (ICs)—rather than learning the underlying physical fault signatures.

This research addresses these challenges by incorporating a fairness-oriented training process into the neural network architecture. We propose a bias-reducing framework to enhance AI reliability by forcing the model to ignore biased environmental features and classify faults based solely on sensor dynamics, thereby preventing the learning of incorrect causal relationships.

2. Methodology

The proposed framework implements an adversarial debiasing algorithm designed to mitigate bias during the training stage. The

architecture comprises two competing neural networks: a classifier (Predictor) and an adversary.

- The Classifier aims to predict the target fault (label Y) with high accuracy from input features (X), while the Adversary attempts to infer the sensitive attribute (Z, such as Initial Conditions) from the Classifier's output logits.
- **Adversarial Debiasing Training:** The Classifier is trained to minimize its own prediction loss while simultaneously maximizing the Adversary's loss. This forces the Classifier to produce representations that are uninformative about the sensitive attribute Z.
- **Gradient Modification:** To ensure the model does not learn the bias, the gradient of the Classifier is modified during each update step. Specifically, the projection of the Adversary's gradient is removed from the Classifier's gradient:
  (i) Let g_{pred} be the gradient of the predictor and g_{adv} be the gradient of the adversary.
  (ii) The unit vector of the adversary's gradient is calculated: $u = normalize(g_{adv})$.
  (iii) The updated gradient g' becomes $g' = g_{pred} - (g_{pred} \cdot u) \cdot u - \lambda g_{\mathrm{adv}}$, where λ is the adversary loss weight.
- **Numerical Stability:** To maintain numerical stability and training stability, gradient clipping and robust normalization techniques are applied.

3. Case Study

In this section, we evaluate the effectiveness of the proposed adversarial debiasing framework using simulated nuclear power plant fault data. We first describe the experimental environment and data configuration, followed by a detailed analysis of the model's performance across various test scenarios.

3.1. *Experimental Setup*

The dataset includes simulated normal operation (MF0) and a feedwater system pipe break (MF1) under two different life cycles: beginning-of-life (IC0) and middle-of-life (IC1). To simulate a biased environment, the training set was constructed with a 100% correlation between the Initial Condition and the fault (e.g., IC0 always MF0, IC1 always MF1). The test set inverted this relationship to evaluate if the model could generalize beyond the training bias.

To evaluate the model's ability to mitigate shortcut learning, the dataset was specifically structured to introduce a strong environmental bias during training, which is then inverted during the evaluation phase. The detailed configuration of the training and evaluation datasets is summarized in Table 1.

Table 1. Configuration of Training and Evaluation Datasets.

Dataset	Initial Condition (IC)	Fault Label (MF)	Correlation Type
Training set	IC0 (BOL)	MF0 (Normal)	Biased correlation
	IC1 (MOL)	MF1 (Fault)	
Evaluation set	IC0 (BOL)	MF1 (Fault)	Inverted bias
	IC1 (MOL)	MF0 (Normal)	

3.2. *Results*

The experiments were conducted across several comparative case studies to validate the effectiveness of the debiasing framework.

3.2.1. *Comparison with Adversarial Debiasing*

The classifier used in this section adopts a simple architecture with a single hidden layer, while the standard model used for comparison does not undergo any bias removal process. In the initial experiment without noise, the standard model achieved 84.4% accuracy. The debiased model achieved 79.0% accuracy, indicating that it successfully began to ignore the IC features, even though this resulted in a slight drop in accuracy on the training-aligned data.

In this case study, the trained model exhibited a high dependency on the non-causal features, which led to a scenario where the removal of bias related to ICs actually resulted in a decrease in overall performance. Although initial condition 1

(IC1) is not associated with the fault (MF1), it is provided with a training dataset whose structure is fully matched to the fault. However, since the model already relies heavily on key variables associated with the fault, the adversarial debiasing process may inadvertently suppress essential diagnostic features while attempting to eliminate IC-related bias.

- **Results of noise-added datasets:** To make the model depend on ICs information, 4% noise was added to the sensor data to weaken the fault-related signals. The standard model's performance dropped to 57.0%, while the debiased model maintained a higher accuracy of 61.1%, demonstrating superior robustness to initial status by focusing on core dynamics.
- **Results of sensitive feature swap:** When labels were swapped (predicting IC as the label while treating MF as the sensitive bias), the standard model failed significantly with only 17.0% accuracy. The debiased model improved this to 36.3%, successfully stripping away some of the unintended influence of the fault data from the IC classification task.

Table 2. Comparative Performance Analysis of Biased and Debiased Models across Experimental Cases.

Case Study	Classification	Model Accuracy	
		Biased	Debiased
Baseline	MF	84.4%	79.0%
With noise (4%)	MF	57.0%	61.1%
Sensitive feature swap	IC	17.0%	36.3%

The results in Table 2 are obtained on the bias-inverted test set, not on training-aligned data. This section uses a simplified model and various experimental conditions to examine the effect of the proposed algorithm. Since such settings do not fully reflect practical scenarios, where complex models may achieve high performance due to spurious correlations, robustness is evaluated in detail using a deeper architecture in Section 3.2.2, which serves as the main evaluation of the proposed approach.

3.2.2. *Comparison with a Deeper Structure*

We determined that the overly simplistic structure of the baseline model limited its ability to capture comprehensive data features, causing the influence of the fault to be disproportionately dominant. To address this and properly verify the effectiveness of the debiasing process, we enhanced the model's complexity, enabling it to learn more sophisticated representations and better distinguish between causal and non-causal correlations. A 3-layer multi-layer perceptron was utilized. The model was trained for 300 epochs with a batch size of 32, using HeNormal initialization and an adversary loss weight of 0.5.

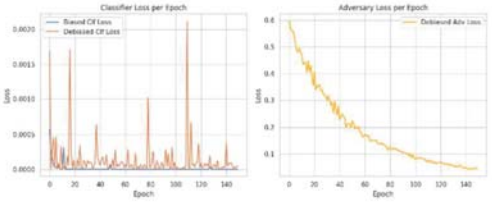

Fig. 1. This displays the classifier loss for both models and the adversary loss for the debiased model over training epochs. It shows how the adversarial component prevents the classifier from minimizing its loss (overfitting to the bias).

The finalized experiment demonstrated that standard models failed to generalize on the inverse test set, achieving low accuracy due to shortcut learning. The debiased model successfully ignored spurious IC signatures. Adversarial debiasing improved the test accuracy from 49.6% to 77.1% under biased conditions, yielding a significant performance gain of 27.5% points.

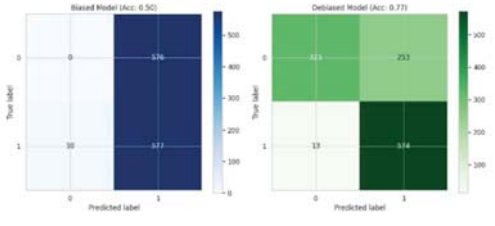

Fig. 2. This figure shows comparing the confusion matrices of the biased and debiased models side-by-side. the biased model effectively guessing,

while the debiased model achieves significantly better true positive and true negative rates.

4. Conclusions

This study demonstrates that adversarial debiasing is an effective tool for mitigating environmental biases in NPP diagnostic models. By constraining the learning process to ignore non-causal features like Initial Conditions, we significantly improved model reliability and generalization performance. Although accuracy and bias improvement can sometimes be inversely related, this framework successfully maintains high performance while ensuring the model relies on physical sensor patterns. However, it would be valuable to investigate the performance of the proposed framework under more moderate levels of bias, beyond the extreme setting considered in this study. Although, future work will utilize quantitative fairness metrics like equalized odds to further validate these trustworthy AI systems.

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

Batra, C. Integral Pressurized Water Reactor Simulator Manual. Vienna, Austria: International Atomic Energy Agency, 2018.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. "Domain-adversarial training of neural networks." Journal of machine learning research 17, no. 59 (2016): 1-35.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 335-340. 2018.