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

doi: 10.3850/ESREL2026061419_esrel26-p26559-cd

Conceptual Framework for Data Integrity Assessment for Artificial Intelligence-Based Abnormal Event Diagnosis in Nuclear Power Plants

Seung Gyu Cho

Nuclear Engineering, Ulsan National Institute of Science and Technology, Republic of Korea. E-mail: sgcho@unist.ac.kr

Seung Jun Lee*

Nuclear Engineering, Ulsan National Institute of Science and Technology, Republic of Korea. E-mail: sjlee420@unist.ac.kr
*(*Corresponding Author)*

Research on nuclear power plant state diagnosis has progressed across multiple directions, including diagnostic model development, performance optimization, uncertainty quantification, and explainable artificial intelligence. While these studies aim to improve diagnostic reliability, they commonly assume that the training data used to develop diagnostic models are inherently appropriate. In practice, training data for state diagnosis are generated according to different abnormal causes and their associated severity levels, which often leads to uneven coverage across scenarios, abnormal intensities, and event progression stages. Although these characteristics are frequently treated as secondary issues, they can fundamentally influence diagnostic behavior. Therefore, for artificial intelligence to be credibly deployed in safety-critical domains such as nuclear power plants, it is necessary to explicitly verify whether the training data themselves are appropriate for the intended diagnostic task, in addition to model-level validation. This study proposes a model-agnostic framework to quantitatively assess data integrity prior to diagnostic model development, specifically for nuclear power plant state diagnosis. Data integrity is quantified along three complementary dimensions: scenario integrity, which evaluates coverage and imbalance across target abnormal scenarios; severity integrity, which assesses the distribution of abnormal intensity within each scenario using a normalized multivariate deviation-based severity proxy; and phase integrity, which evaluates temporal coverage relative to event onset. These dimensions are integrated into a unified representation that quantifies discrepancies between the empirical data distribution and a diagnostically justified target distribution. The quantified integrity results are then directly used to guide under-sampling and data augmentation. The proposed framework is intended to serve as a principled basis for integrity-aware data preparation, with detailed validation and performance assessment to be addressed in future studies.

Keywords: Nuclear power plant, safety-critical systems, artificial intelligence, data integrity, abnormal event diagnosis, data augmentation, under-sampling.

1. Introduction

Research on nuclear power plant (NPP) state diagnosis has progressed across multiple directions, including diagnostic model development, performance optimization, uncertainty quantification, and explainable artificial intelligence (Chae et al. 2022; Cho, Shin, and Lee 2025; Mendoza and Tsvetkov 2023; Sethu et al. 2023; Shin et al. 2023; Wang et al. 2025). These efforts have significantly improved diagnostic capability under complex and transient conditions. However, most existing studies implicitly assume that the training data used to develop diagnostic models are inherently appropriate and representative of the intended diagnostic task. In practice, training data for NPP state diagnosis are generated according to different abnormal causes and their associated severity levels. This process often results in uneven coverage across abnormal scenarios, imbalanced distributions of abnormal intensities, and inconsistent representation of event progression stages. For example, some scenarios may be overrepresented due to their ease of simulation, while severe or early-phase conditions may be sparsely sampled. Although such characteristics are often treated as secondary issues, they can fundamentally influence diagnostic behavior, leading to biased decision boundaries, unstable predictions, and reduced trustworthiness (Chen et al. 2024). For example, when the abnormal cause is a pump-related malfunction, the resulting data imbalance becomes particularly evident. Pump-related abnormalities can manifest in diverse forms, such as complete pump trip, partial degradation, intermittent operation, or delayed startup failure. However, in many training datasets, pump abnormalities are represented predominantly by abrupt pump trip scenarios, since they are straightforward to define and simulate. As a result, intermediate severity conditions, such as gradual flow degradation or oscillatory behavior before trip, are often underrepresented. Moreover, data samples are frequently concentrated after the pump trip event, while the early-phase transient leading up to the failure is sparsely captured. This imbalance can cause diagnostic models to associate pump-related faults primarily with post-trip signatures, reducing sensitivity to early or mild pump degradation and leading to delayed or unstable diagnosis.

In this study, data integrity is defined not as the absence of missing values or sensor noise, but as the structural suitability of a dataset for a specific diagnostic objective. Data integrity refers to whether the training data provide diagnostically meaningful and sufficiently balanced coverage across abnormal scenarios, abnormal severity levels, and event progression phases. From this perspective, even a large and noise-free dataset may lack integrity if its distribution is misaligned with the intended diagnostic task. This issue is closely related to verification and validation (V&V) requirements in safety-critical systems (Park and Koo 2025). Regulatory frameworks emphasize not only correct implementation and acceptable performance, but also the validity of assumptions underlying system behavior. When data-driven diagnostic systems are trained on structurally inappropriate data, model-level validation alone cannot guarantee reliable behavior under unseen or safety-significant conditions. Therefore, data integrity assessment should be regarded as a prerequisite step in the V&V process for AI-based diagnostic systems in nuclear power plants.

Beyond the nuclear domain, the importance of training data quality has been systematically emphasized through the data-centric AI paradigm (Zha et al. 2025), which argues that systematic engineering of the dataset itself often yields greater reliability gains than further model-centric optimization. In parallel, the imbalanced learning community has extensively investigated class imbalance and its effects on deep learning (Johnson and Khoshgoftaar 2019), and assurance-oriented studies have identified dataset balance, completeness, and relevance as essential desiderata in the machine learning lifecycle for safety-critical systems (Ashmore, Calinescu, and Paterson 2021). Within the nuclear domain, several recent studies have also addressed imbalance through data augmentation techniques such as generative adversarial networks and oversampling (Ma et al. 2025). However, most of these efforts treat imbalance as a single-dimensional class-level problem and introduce corrections that are tightly coupled to a particular diagnostic model. A structured assessment that jointly considers the multi-dimensional structure of abnormal events in NPPs, namely the scenario, severity, and phase dimensions, and that operates independently of any specific diagnostic architecture, has not yet been explicitly formulated. The proposed framework aims to fill this gap.

To address this gap, this study proposes a model-agnostic framework to quantitatively assess data integrity prior to diagnostic model development for NPP state diagnosis. Data integrity is evaluated along three complementary dimensions.

(i) Scenario integrity evaluates whether the training data provide sufficient and balanced coverage across all target abnormal scenarios required for reliable state diagnosis.

(ii) Severity integrity quantifies the distribution of abnormal intensity levels within each scenario using a normalized multivariate deviation-based severity proxy and assesses the representativeness of mild, moderate, and severe conditions.

(iii) Phase integrity evaluates the adequacy of temporal data coverage relative to event onset, including early, intermediate, and post-event phases of abnormal progression.

These three integrity dimensions are integrated into a unified representation that quantifies discrepancies between the empirical data distribution and a diagnostically justified target distribution. The resulting integrity metrics are then directly used to guide under-sampling and data augmentation. By correcting data imbalance at the dataset level, the proposed approach improves diagnostic stability and trustworthiness without modifying the diagnostic model itself.

2. Methodology

2.1. *Overall Framework of Data Integrity Assessment*

This section presents the overall framework for assessing and correcting the integrity of training data prior to diagnostic model development. The proposed framework is designed to operate independently of the diagnostic model architecture, focusing instead on the structural suitability of the training dataset used for NPP state diagnosis. Figure 1 illustrates the overall workflow of the proposed data integrity assessment and correction framework. The framework starts from a raw training dataset composed of scenario, severity, and time dimensions, which inherently reflect the characteristics of abnormal events in NPP operations. Rather than directly using these data for model training, the framework first evaluates whether the dataset provides balanced and representative coverage of abnormal operating conditions.

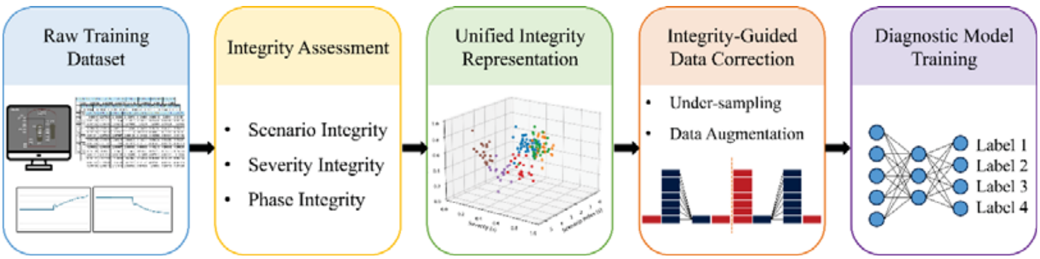

Fig. 1 Overview of the proposed data integrity assessment and correction framework for NPP state diagnosis

To this end, integrity assessment modules are applied to quantify data integrity along three complementary dimensions: scenario integrity, severity integrity, and phase integrity. Scenario integrity evaluates whether different abnormal event scenarios are sufficiently represented, severity integrity examines the distribution of event magnitudes within each scenario, and phase integrity assesses the coverage of different stages in accident progression. These integrity measures are independently quantified to avoid bias toward any specific diagnostic model or learning strategy. The results of the integrity assessments are then unified into an integrity representation by measuring the distance between the observed data distribution and a target balanced distribution. This unified representation enables the identification of structural imbalances in the training dataset that may adversely affect diagnostic reliability.

Based on the quantified integrity representation, an integrity-guided data correction process is performed. Sampling-based strategies, including under-sampling and data augmentation, are selectively applied to mitigate detected imbalances. Importantly, this correction process is driven by integrity metrics rather than by model performance, ensuring that data modification remains objective and model-agnostic. Through this pre-training evaluation and correction procedure, the proposed framework functions as a gate in the verification and validation (V&V) process of training data. As a result, the final training dataset is structurally balanced and suitable for reliable and robust NPP state diagnosis, regardless of the diagnostic model employed in subsequent stages.

2.2. *Integrity Quantification Metrics for Training Data*

In this study, data integrity is defined as the degree to which the pre-training dataset appropriately represents diagnostically relevant operating conditions for nuclear power plant state diagnosis, without systematic omission or distortion. Integrity is quantified by jointly evaluating three complementary aspects of data composition: scenario integrity, severity integrity, and phase integrity.

Let $s \in S$ denote an abnormal scenario and let i denote an individual data sample associated with scenario s.

First, scenario integrity evaluates whether each target abnormal scenario is sufficiently represented in the training dataset. For each scenario s, coverage conditions such as presence, minimum representation, and relative imbalance are examined to identify scenarios that are missing or disproportionately represented.

Second, severity integrity assesses whether different abnormal intensities within each scenario are adequately represented. For each sample i belonging to scenario s, a severity representation value $S_{i,s}$ is computed to reflect the magnitude of deviation from a normal operating reference. The severity representation is normalized within each scenario and evaluated based on its distribution across predefined severity levels.

Third, phase integrity evaluates temporal representation with respect to abnormal event progression. Let t denote the time index of a sample and let t_e denote the abnormal event onset time for scenario s. Each sample is assigned to an early, mid, or late phase based on the relative time difference $t - t_e$. This classification ensures that different stages of accident progression are sufficiently represented in the training dataset.

The three integrity components are jointly considered by mapping samples into a three-dimensional representation space defined by scenario, severity level, and phase (s, v, p). For each cell in this space, a representation level $R_{s,v,p}$ is defined based on the number of samples assigned to the corresponding condition. Deviations from a target balanced representation are quantified as

$$\Delta R_{s,v,p} = R^*_{s,v,p} - R_{s,v,p} \qquad (1)$$

Where $R^*_{s,v,p}$ denotes the desired representation level. The resulting representation disparity provides a quantitative measure of structural data integrity and forms the basis for integrity-guided data correction.

The representation disparity $\Delta R_{s,v,p}$ is explicitly computed for each representation cell and directly used to assess structural imbalance in the training dataset. When the magnitude of $\Delta R_{s,v,p}$ remains within a predefined tolerance range, the corresponding data distribution is considered acceptable and is used without modification. Integrity-guided data correction is applied only when the absolute value of $|\Delta R_{s,v,p}|$ exceeds the acceptable threshold. In such cases, the sign of $\Delta R_{s,v,p}$ indicates whether the representation is insufficient or excessive relative to the target, and appropriate sampling actions are selectively applied. This threshold-based interpretation prevents unnecessary data manipulation while ensuring that only structurally significant imbalances are corrected.

2.2.1. *Scenario Coverage Metric*

integrity evaluates whether each abnormal event scenario is sufficiently and consistently represented in the training dataset. Its primary objective is to ensure that the dataset does not omit target scenarios or disproportionately emphasize a limited subset of scenarios, which could bias subsequent diagnostic models. Scenario integrity evaluates whether each abnormal event scenario is sufficiently and consistently represented in the training dataset. Its primary objective is to ensure that the dataset does not omit target scenarios or disproportionately emphasize a limited subset of scenarios, which could bias subsequent diagnostic models. For each scenario, a scenario-level representation metric R_s is defined by aggregating the representation metrics of all representation cells associated with scenario s. This metric reflects the overall degree to which the scenario is represented in the training dataset, independent of severity level or phase.

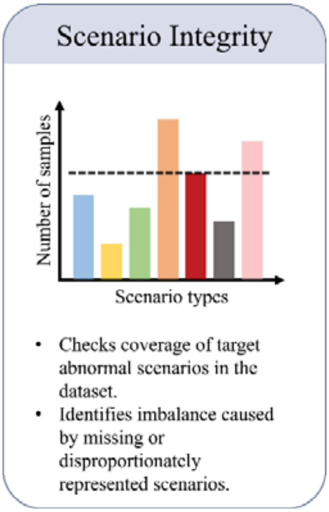

Fig. 2. Scenario Coverage Integrity across Abnormal Event Types

2.2.2. *Severity Distribution Metric*

Severity integrity evaluates whether different abnormal intensities within each scenario are adequately represented in the training dataset. Its objective is to prevent the dataset from being biased toward only mild or extreme conditions, which could limit diagnostic reliability across the full range of abnormal behavior. each sample is associated with a severity level that reflects the relative magnitude of abnormal deviation from a normal operating reference. Severity levels are normalized within each scenario to ensure comparability and are discretized into predefined severity bins v. These bins represent different ranges of abnormal intensity, independent of scenario identity.

Severity integrity is assessed by examining the representation metric across severity bins for each scenario. For each bin v, a severity-level representation metric R_v is computed by aggregating the representation metrics of all representation cells associated with the corresponding severity level. This metric indicates whether the training data sufficiently cover the range of abnormal intensities within the scenario. When the severity-level representation metric R_v remains within an acceptable tolerance range relative to the target representation, the corresponding severity range is considered adequately represented and is used without modification. Severity levels that exhibit significant under-representation or over-representation beyond the predefined threshold are identified as candidates for integrity-guided data correction. This evaluation ensures balanced learning across different abnormal intensities while avoiding unnecessary modification of acceptable data distributions.

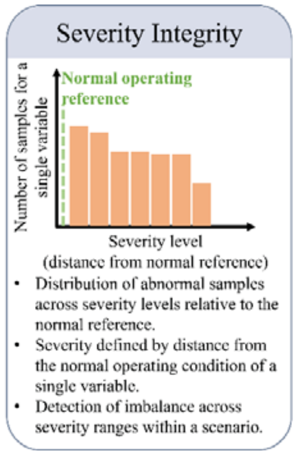

Fig. 3. Severity Distribution Integrity relative to Normal Condition

2.2.3. *Phase Representation Metric*

Phase integrity evaluates whether different stages of abnormal event progression are adequately represented in the training dataset. Its objective is to ensure that the dataset does not concentrate excessively on a specific temporal segment, such as early or late phases, while neglecting other stages that are equally important for reliable diagnosis.

For a given scenario s, each sample is associated with a temporal position relative to abnormal event onset. Let t denote the time index of a sample and t_e denote the time index of a sample and s. denote the onset time of the abnormal event for scenario $t - t_e$, samples are categorized into predefined phase levels p, such as early, mid, and late phases. These phase levels represent different stages of abnormal event evolution, independent of severity magnitude.

An important consideration in phase definition is that abnormal events in NPPs evolve on different characteristic time scales prior to reactor trip. Some abnormal events progress rapidly within seconds to minutes, such as a stuck-open pilot-operated safety relief valve (POSRV) that induces a LOCA-like pressure transient, whereas others develop gradually over tens of minutes to hours, such as a slow decrease in condensate storage tank (CST) level. Consequently, applying a single absolute time boundary for early, mid, and late phases would misrepresent the diagnostically meaningful stages across heterogeneous scenarios. To address this, phase boundaries in the proposed framework are defined relative to scenario-specific characteristic time scales rather than absolute time. Specifically, for each scenario s, a characteristic duration T_s is determined based on the time required for the dominant process variables to reach a post-event stabilized regime, and the normalized relative time τ=(t-t_e)/T_s is used to assign each sample to an early, mid, or late phase according to predefined thresholds on τ. This scenario-normalized definition ensures that phase integrity is evaluated on a consistent basis across scenarios with different evolution speeds, while preserving the diagnostic significance of each temporal stage. The specific threshold values and the determination of T_s for each scenario class will be detailed in subsequent case studies.

Phase integrity is assessed by examining the representation metric across phase levels for each scenario. For each phase level p, a phase-level representation metric R_p is defined by aggregating the representation metrics of all representation cells associated with the corresponding phase. This metric indicates whether the training dataset sufficiently covers the temporal evolution of abnormal events. When the phase-level representation metric R_p , remains within an acceptable tolerance range relative to the target representation, the corresponding phase is considered adequately represented and is used without modification. Phase levels exhibiting significant under-representation or over-representation beyond the predefined threshold are identified as candidates for integrity-guided data correction. This evaluation prevents temporal bias in training data while preserving meaningful temporal characteristics of abnormal event progression.

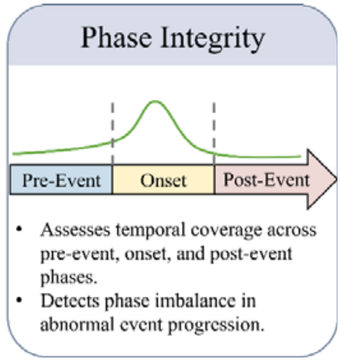

Fig. 4. Temporal Phase Integrity of Abnormal Event Progression

2.2.4. *Representation Disparity Metric*

This subsection integrates the scenario-, severity-, and phase-level integrity metrics into a unified representation metric that captures the overall structural balance of the training dataset. While the preceding metrics evaluate representation along individual dimensions, the integrated representation metric provides a joint assessment across scenario, severity level, and phase. As described in Section 2.2, samples are mapped into a three-dimensional representation space defined by (s, v, p), where each cell corresponds to a specific combination of scenario, severity level, and phase. For each representation cell, a representation metric $R_{s,v,p}$ is defined based on the number of samples assigned to the corresponding condition. A target representation metric $R^*_{s,v,p}$ is specified according to the desired balanced data distribution.

The deviation of the representation metric from its target is quantified as Eq. (1). The integrated representation metric is interpreted as a measure of structural data integrity. When the magnitude of $\Delta R_{s,v,p}$ remains within a predefined tolerance $|\Delta R_{s,v,p}|$, the representation of the corresponding condition is considered acceptable, and the data are used without modification. Conversely, representation cells exhibiting deviations beyond the acceptable range are identified as candidates for integrity-guided data correction.

Importantly, the integrated representation metric does not impose a deterministic sampling rule. Instead, it provides a quantitative criterion for identifying structurally significant imbalances, which are addressed selectively in the subsequent data correction stage. By jointly considering scenario, severity, and phase dimensions, this metric ensures that integrity assessment reflects the combined structure of abnormal event data rather than isolated marginal distributions.

2.3. *Sampling-Based Data Correction Strategy*

This section describes the data correction strategy guided by the representation metrics defined in Section 2.2. The objective of this strategy is not to maximize data quantity, but to mitigate structurally significant imbalances that may adversely affect diagnostic reliability. For each representation cell defined in the (s, v, p) space, the representation metric deviation $\Delta R_{s,v,p}$ is evaluated. When the magnitude of $\Delta R_{s,v,p}$ remains within a predefined tolerance range, the corresponding data are considered structurally acceptable and are used without modification. This prevents unnecessary data manipulation and preserves the original characteristics of the training dataset.

When $|\Delta R_{s,v,p}|$ exceeds the acceptable threshold, the representation cell is identified as a candidate for integrity-guided correction. The direction of the deviation indicates whether the cell is under-represented or over-represented relative to the target representation. For under-represented cells, data augmentation is selectively applied to improve coverage, whereas over-represented cells are addressed through under-sampling to reduce redundancy.

Importantly, data correction is not applied uniformly to all imbalanced cells. Practical constraints, such as physical plausibility of data generation, diagnostic relevance of the affected region, and the role of specific phases in early detection, are jointly considered before applying corrective actions. As a result, representation metrics serve as quantitative indicators for identifying where correction may be beneficial, rather than as deterministic rules that enforce mandatory sampling decisions. Through this integrity-guided correction strategy, the training dataset is adjusted to achieve improved structural balance across scenario, severity, and phase dimensions, while preserving diagnostically meaningful data characteristics. This process ensures that subsequent diagnostic model training is based on data distributions that are both representative and physically consistent.

3. Conclusion and Future Work

This study proposed a model-agnostic framework for assessing training data integrity in nuclear power plant state diagnosis. Unlike conventional approaches that primarily emphasize diagnostic model performance, this work focused on the structural suitability of training data prior to model development. Data integrity was formulated along three complementary dimensions—scenario integrity, severity integrity, and phase integrity—reflecting coverage across abnormal causes, abnormal intensity levels, and event progression stages, respectively. By integrating these dimensions into a unified representation, the proposed framework provides a systematic means to identify diagnostically meaningful imbalances in training datasets.

The primary contribution of this study lies in the conceptual formulation of data integrity as a quantifiable and verifiable property, independent of any specific diagnostic model. Rather than introducing a new diagnosis algorithm, this work establishes a principled basis for integrity-aware data preparation in safety-critical applications, where inappropriate training data can fundamentally distort diagnostic behavior even when model performance appears satisfactory.

Future work will focus on comprehensive validation of the proposed framework through plant-scale case studies and diagnostic model integration. This will include quantitative evaluation of how integrity-guided data handling influences diagnostic stability, robustness, and uncertainty under realistic operating conditions, as well as investigation of its applicability across different reactor types, abnormal scenarios, and diagnostic paradigms.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2022-00144042)

This research was supported by the National Research Council of Science & Technology(NST) grant by the Korea government (MSIT) (No. GTL24031-000)

References

Ashmore, Rob, Radu Calinescu, and Colin Paterson. 2021. 'Assuring the machine learning lifecycle: Desiderata, methods, and challenges', ACM computing surveys (CSUR), 54: 1-39.

Chae, Young Ho, Chanyoung Lee, Sang Min Han, and Poong Hyun Seong. 2022. 'Graph neural network based multiple accident diagnosis in nuclear power plants: Data optimization to represent the system configuration', Nuclear Engineering and Technology, 54: 2859-70.

Chen, Wuxing, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and CL Philip Chen. 2024. 'A survey on imbalanced learning: latest research, applications and future directions', Artificial Intelligence Review, 57: 137.

Cho, Seung Gyu, Ji Hyeon Shin, and Seung Jun Lee. 2025. 'Robust diagnosis of abnormal events using fuzzy-based feature extraction in nuclear power plants', Progress in Nuclear Energy, 185: 105765.

Johnson, Justin M, and Taghi M Khoshgoftaar. 2019. 'Survey on deep learning with class imbalance', Journal of big data, 6: 27.

Ma, Zhanguo, Jing Cui, Jing Zhang, Liang Zhang, Wenhao Jia, and Long Tian. 2025. 'Fault diagnosis study for nuclear power plants under imbalanced fault sample datasets based on deep learning', Annals of Nuclear Energy, 223: 111634.

Mendoza, Mario, and Pavel V Tsvetkov. 2023. 'An intelligent fault detection and diagnosis monitoring system for reactor operational resilience: power transient identification', Progress in Nuclear Energy, 156: 104529.

Park, JaeKwan, and SeoRyong Koo. 2025. 'Reliability verification methods for artificial intelligence components in instrumentation and control systems at nuclear power plants', Nuclear Engineering and Technology: 104096.

Sethu, Meenu, Bhavya Kotla, Darrell Russell, Mahboubeh Madadi, Nesar Ahmed Titu, Jamie Baalis Coble, Ronald L Boring, Klaus Blache, Vivek Agarwal, and Vaibhav Yadav. 2023. 'Application of artificial intelligence in detection and mitigation of human factor errors in nuclear power plants: a review', Nuclear Technology, 209: 276-94.

Shin, Ji Hyeon, Junyong Bae, Jae Min Kim, and Seung Jun Lee. 2023. 'An interpretable convolutional neural network for nuclear power plant abnormal events', Applied Soft Computing, 132: 109792.

Wang, Zixiao, Hongyang Wei, Ruifeng Tian, and Sichao Tan. 2025. 'A review of data-driven fault diagnosis method for nuclear power plant', Progress in Nuclear Energy, 186: 105785.

Zha, Daochen, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. 'Data-centric artificial intelligence: A survey', ACM Computing Surveys, 57: 1-42.