Uncertainty-Aware Diagnosis of Multi-Abnormal Events in Nuclear Power Plants Using Evidential Deep Learning

Seung Gyu Cho, Seung Jun Lee*
Ulsan National Institute of Science and Technology, 50, UNIST-gil, Ulsan, 44911
**Corresponding author: sjlee420@unist.ac.kr*

***Keywords : Evidential deep learning, abnormal event diagnosis, nuclear power plant**

1. Introduction

Abnormal event diagnosis in nuclear power plants (NPPs) plays a vital role in ensuring safe and stable operation. While deep learning techniques have demonstrated promising performance in this domain, conventional neural networks often suffer from overconfident predictions, especially under unfamiliar or out-of-distribution (OOD) conditions. This limitation poses significant challenges in safety-critical environments such as NPPs.

To address this issue, we propose a decoupled Evidential Deep Learning (EDL) framework that estimates predictive uncertainty in parallel with the diagnosis process. By modeling class probabilities as parameters of a Dirichlet distribution, the proposed method enables the model to output both class predictions and their associated confidence levels independently from the embedding training. This is particularly beneficial for diagnosing multi-abnormal events, where overlapping fault signatures and variable interactions often lead to ambiguous decision boundaries.

2. Methodology

2.1. Metric Learning-Based Multi-Abnormal Event Diagnosis

To effectively diagnose both single and multi-abnormal events in NPPs, we adopt a metric learning framework based on triplet loss. This approach encourages the model to learn a representation space where embeddings from the same abnormal class are close together, and embeddings from different classes are well-separated. This property is especially valuable for multi-abnormal events, where the interaction of multiple faults causes complex and overlapping patterns in the sensor data. Given an anchor input x_a, a positive sample x_p from the same class, and a negative sample x_n from a different class, the triplet loss is defined as [1]:

$$(1)\quad \mathcal{L}_{triplet} = \max(\left\|f(x_a) - f(x_p)\right\|_2^2 - \|f(x_a) - f(x_n)\|_2^2 + \alpha, 0)$$

where $f(\cdot)$ is the encoder network and α is a margin constant. The encoder maps time-series sensor data into a compact embedding vector, which captures the temporal evolution and class-discriminative features of abnormal events.

The overall training framework is illustrated in Fig. 1.

During training, the encoder network is optimized using the triplet loss to ensure that samples from the same abnormal class are embedded closely together, while those from different classes are well separated. After training, the encoder is frozen and used to compute representative vectors for each class, which serve as anchors for the classification process.

The classification process during inference is shown in Fig. 2. When a test sample is passed through the encoder, its embedding is compared to the representative class vectors using either Euclidean or Mahalanobis distance. The Mahalanobis distance is first used to determine whether the input likely corresponds to a multi-abnormal case. Based on this result, the system performs either a single-class (multi-class) diagnosis or a multi-label diagnosis. The final prediction is made based on the nearest class or pair of classes, depending on the inferred abnormality type.

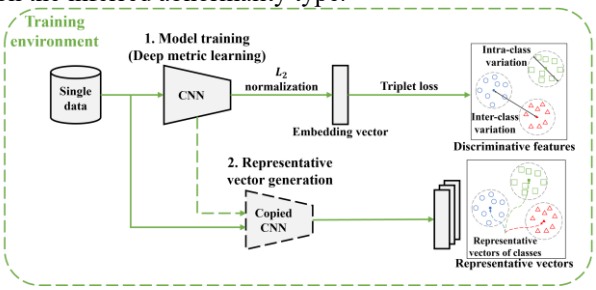

Fig. 1. Training architecture of multi-abnormal event diagnosis model using metric learning.

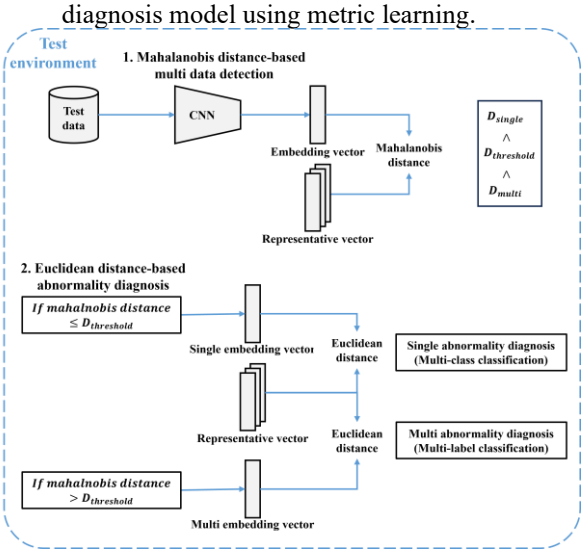

Fig. 2. Multi-abnormal event diagnosis framework.

2.2. Evidential Deep Learning for Uncertainty Estimation

To estimate uncertainty without interfering with the embedding learning, an evidential deep learning (EDL) head is applied in parallel to the metric learning branch.

Unlike conventional softmax-based classifiers that produce point estimates of class probabilities, EDL models the output as a Dirichlet distribution over class probability vectors. The Dirichlet distribution is a multivariate distribution commonly used as a conjugate prior for categorical distributions in Bayesian inference [2]. It is parameterized by non-negative values $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_K]$, where each α_k can be interpreted as evidence collected in support of class k. A concentrated Dirichlet distribution (i.e., large total evidence) leads to confident predictions, while a flat distribution with low evidence indicates uncertainty. This probabilistic framework allows the model to jointly capture class likelihoods and the epistemic uncertainty of its predictions.

This module outputs a vector of non-negative evidence values $\boldsymbol{e} = [e_1, \dots, e_K]$, which parameterizes a Dirichlet distribution over K predefined classes. The Dirichlet parameters are calculated by adding one to each evidence value:

$$(2) \quad \alpha_k = e_k + 1$$

Using these parameters, the expected probability for class k is given by:

$$(3) \quad \hat{p}_k = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j}$$

The total uncertainty, which reflects the degree of evidence insufficiency across all classes, is defined as:

$$(4) \quad u = \frac{K}{\sum_{j=1}^{K} \alpha_j}$$

Additionally, the belief mass, which can be interpreted as the model's confidence in each class, is computed by normalizing the evidence:

$$(5) \quad b_k = \frac{e_k}{\sum_{j=1}^{K} \alpha_j}$$

The belief mass b_k and uncertainty u together provide interpretable information about the model's predictive reliability. A low uncertainty and concentrated belief indicate a confident decision, while high uncertainty and dispersed beliefs suggest doubt or ambiguity in the classification. Importantly, to preserve the representation quality of the metric learning encoder, the EDL head is trained independently using a separate evidential loss function. No gradients from the EDL branch are propagated back into the encoder. This decoupled training strategy allows the model to learn class-separable embeddings while independently quantifying uncertainty from the same feature representation.

3. Experimental settings

This study utilizes time-series datasets generated using the 3KEYMASTER simulator, which models a two-loop 1400 MWe pressurized water reactor (PWR). All scenarios were simulated at 100% full power in the middle of the reactor's operational life. Each dataset contains 60 seconds of data sampled from approximately 2,800 plant variables, representing the signals displayed in the human–machine interface. Abnormal events were introduced by injecting faults over a 10-second period immediately after startup. For each abnormal event label, 50 datasets were generated with varying severity levels. In total, the dataset includes 25 classes, resulting in 1,250 unique scenarios and 75,000 seconds of simulated data. Table 1 provides the injection points and descriptions of all scenarios, including one normal condition and 24 single-abnormal events.

Table 1. Normal and single-abnormal event dataset description.

Index	Location of abnormality injection (Label)	Description	Index	Location of abnormality injection (Label)	Description
1	None (Normal)	Middle of life cycle at 100% power generation	14	Charging water system (CHRG[LK])	Charging line break upstream
2	Pilot-operated safety relief valve (POSRV[VO])	POSRV leak	15	Component cooling water (CCW[LK])	CCW service loop header leak to aux atm
3	Reactor vessel head flange leakage (RVHF[VO])	Reactor vessel head flange leak inside containment	16	Turbine control system (TCS[VC])	High-pressure turbine control valve positioner close failure
4	Steam generator tube leakage (SGTL[TL])	Steam generator tube leak	17	Main steam isolation valve (MSIV[VC])	MSIV positioner failure

5	Reactor coolant pump (RCP[LC])	Loss of reactor coolant pump component cooling water (CCW) to RCP	18	Steam bypass control system (SBCS[VO])	Steam bypass control valve stuck open
6	Reactor coolant pump (RCP[LS])	Reactor coolant pump injection seal loss	19	High-pressure feedwater heater (HFH[TL])	High feedwater heater tube break
7	Pressurizer (PZR[VO])	PZR spray valve positioner failure	20	Low-pressure feedwater heater (LFH[TL])	Low feedwater heater tube break
8	Volume and control tank level high (VCT[LL])	Volume and control tank level low	21	Condensate storage tank (CST[LL])	Condensate storage tank level low
9	Letdown water system (LTDN[LK])	Letdown line leak inside the containment	22	Condensate system (CDS[LV])	Loss of condenser vacuum
10	Letdown water system (LTDN[VC])	Loss of letdown line flow due to valve stuck close	23	Main feed water (MFW[VO])	MFW pump recirculating valve positioner open failure
11	Letdown water system (LTDN[LC])	Abnormal letdown temperature due to loss of CCW	24	Main feedwater isolation valve (MFIV[VC])	Main feedwater isolation valve stuck open
12	Charging water system (CHRG[PP])	Charging pump breaker trip	25	Circulating water system (CWS[TL])	Circulating water tube leak in low pressure condenser
13	Charging water system (CHRG[VC])	Charging line valve positioner failure			

4. Results

To analyze the impact of decision criteria on diagnostic performance, we varied the margin α and threshold k values used in the distance-based classification process. Table 2 shows the resulting performance across four metrics: single diagnosis accuracy, multi-abnormal event detection accuracy, multi diagnosis accuracy, and total diagnosis accuracy.

As the margin and threshold increase, multi diagnosis accuracy tends to improve, while single diagnosis accuracy remains relatively stable. Notably, total diagnostic performance is maximized at $\alpha = 1.25$ and $k = 4$, suggesting a balance between sensitivity and specificity in distinguishing overlapping events.

Table 2. Diagnostic performance (%) under different margin $\boldsymbol{\alpha}$ and threshold $\boldsymbol{k}$ values.

Margin (α)	Threshold (k)	Single diagnosis accuracy (%)	Multi detection accuracy (%)	Multi diagnosis accuracy (%)	Total diagnosis accuracy (%)
0.5	3	97.2	99.12	97.58	97.55
	4	99.2	99.09	97.42	97.59
	5	99.6	99.05	97.38	97.59
0.75	3	97.6	99.45	98.67	98.57
	4	99.2	99.52	98.67	98.72
	5	100	99.38	98.47	98.61
1.0	3	97.6	99.52	98.99	98.87
	4	98.8	99.49	98.87	98.87

	5	99.6	99.38	98.71	98.79
1.25	3	96.4	99.49	99.15	98.9
	4	99.2	99.49	98.99	99.01
	5	**100**	**99.45**	**98.91**	**99.01**
1.5	3	98	99.52	98.87	98.79
	4	98.8	99.52	98.83	98.83
	5	100	99.38	98.63	98.76

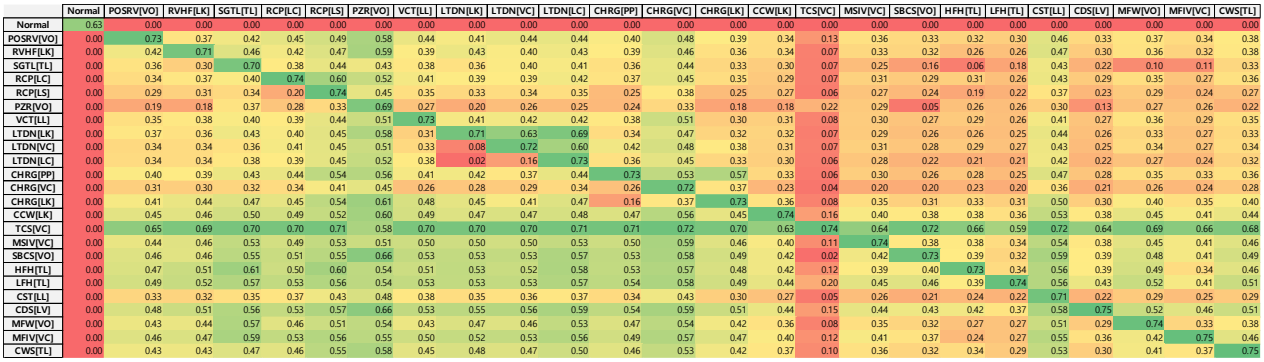

Fig. 3. EDL-based confidence distribution for single and multi-abnormal events.

The EDL-based confidence results are summarized in Fig. 3. In the confidence map, single abnormal events appear along the diagonal with consistently high values, indicating stable and reliable classification. In contrast, for multi-abnormal events, confidence values are distributed across multiple labels, reflecting the ambiguity introduced by overlapping event behaviors. This result demonstrates that EDL effectively captures the uncertainty structure of complex cases, providing an additional layer of interpretability beyond conventional accuracy metrics. These findings confirm that the proposed approach not only achieves high diagnostic accuracy but also offers valuable insights into the reliability of decisions in multi-abnormal event diagnosis.

An exception was observed for TCS[VC]. Despite its high diagnostic accuracy, the EDL results indicated patterns similar to single abnormal events even in multi-abnormal cases. This tendency may be attributed to the dominant and distinctive signature of TCS-related signals, which concentrate most of the evidential mass on the TCS label. While this characteristic enhances separability and detection power, it can obscure co-occurring events, suggesting the need for tailored criteria or training strategies to better capture multi-abnormal structures.

5. Conclusion

This study proposed a novel abnormal event diagnosis framework that combines metric learning and evidential deep learning (EDL) to address the challenges of diagnosing both single and multi-abnormal events in nuclear power plants (NPPs). The metric learning component, trained with triplet loss, enabled the model to learn a class-separable embedding space, while the EDL head provided uncertainty-aware predictions based on Dirichlet distributions. To prevent interference between embedding learning and uncertainty estimation, the two components were trained in parallel using decoupled losses.

The experimental results demonstrated that the proposed method achieves high diagnostic accuracy across both single and multi-abnormal scenarios, even under overlapping and ambiguous conditions. In particular, total diagnosis accuracy was maximized when the decision margin and threshold were appropriately tuned. Furthermore, the EDL component effectively quantified the predictive confidence of the model, offering interpretable uncertainty estimates that can support safer and more informed decision-making.

Future work will extend this framework to handle unknown abnormal events and integrate it into real-time operator support systems. The results suggest that the proposed approach can enhance the reliability and transparency of AI-assisted diagnosis in safety-critical domains such as nuclear power plant operation.

ACKNOWLEDGEMENTS

This work was supported by KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2024-Tech-09)
This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2022-00144042)

REFERENCES

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815-823.

[2] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems,* vol. 31, 2018.