

Signal Anomaly Classification in Piezoelectric Sensor Systems Using Tree-Based Models

Geonju Lee¹, Hee-Jae Lee², and Jonghyun Kim^{2*}

¹Dongguk University, Department of Electronics and Electrical Engineering, Seoul, Republic of Korea

²KAIST, Department of Nuclear and Quantum Engineering, Daejeon, Republic of Korea

*Corresponding author: jonghyun.kim@kaist.ac.kr

***Keywords:** piezoelectric sensor, loose parts monitoring system, signal classification, tree-based models, nuclear power plant safety

1. Introduction

The safety monitoring systems in nuclear power plants (NPPs) must detect mechanical impacts and structural anomalies at an early stage to prevent equipment failures and operational interruptions. Piezoelectric sensors, which convert mechanical impact into electrical signals, are therefore widely utilized in various systems such as vibration monitoring, seismic response analysis, and leakage detection [1]. Among these, the loose parts monitoring system (LPMS) is a representative system designed to detect metallic debris or structural fragments inside the reactor, typically relying on threshold-based logic to determine impact events [2].

However, in current industrial applications, piezoelectric sensor systems often generate high-amplitude signals in response to non-impact events such as cable vibrations or circuit anomalies, resulting in false alarms and operator confusion [3,4]. To address this issue, this study proposes a two-stage classification framework that integrates tree-based models such as eXtreme Gradient Boosting (XGBoost) and Random Forest [5,6] with feature extraction techniques like Short-Time Fourier Transform (STFT) and Continuous Wavelet Transform (CWT) [7,8]. The proposed method aims to distinguish between four signal classes: normal, impact, cable vibration, and circuit fault and to localize fault causes (charge amplifier vs. low-pass filter). The goal is to improve the accuracy of signal interpretation in LPMS and enhance the overall reliability of plant operations.

2. Data Collection

The reliability of machine learning models heavily depends on the quality of training data. However, collecting abnormal signal data, such as those caused by circuit faults, is challenging in real-world systems due to safety and operational constraints. To overcome this limitation, this study generates training data for fault signals using PSPICE circuit simulations, enabling the controlled acquisition of training data. Section 2.1 describes the PSPICE circuit modeling process, while Section 2.2 outlines the data collection methodology and characteristics of the generated signals.

2.1. Circuit Modeling

To replicate the signal processing of a piezoelectric sensor system, a nominal-state circuit was modeled using OrCAD Capture CIS. The circuit consists of a charge amplifier and a low-pass filter. In LPMS, the piezoelectric sensor system is typically composed of a charge amplifier that amplifies the measured charge and a low-pass filter that removes high-frequency noise. Considering this configuration, the circuit was modeled as shown in Figure 1.

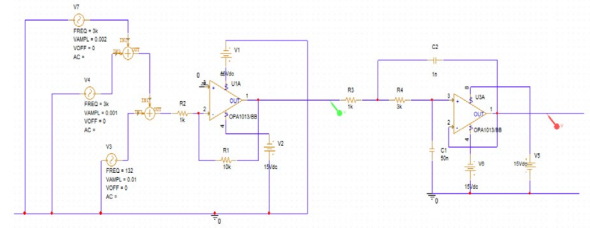


Figure 1. Design of normal circuit.

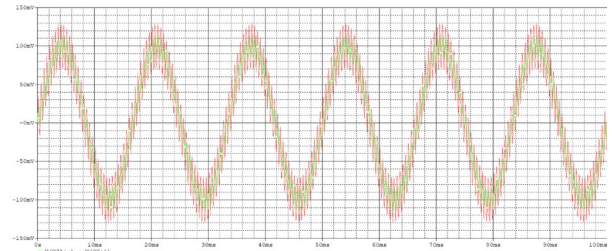


Figure 2. Data achieved from red and green probes (appear in Figure 1).

In the circuit, the charge output of the piezoelectric sensor was approximated as a voltage input, which is valid under high-impedance conditions since $q = Cv$. This approach avoids adding extra capacitance at the input stage, thereby simplifying the calculation of the filter's cutoff frequency and facilitating analysis. Noise injection was implemented by summing two voltage sources, and the filter was designed using a Sallen–Key topology.

Circuit parameters were adjusted to match the frequency, period, and amplitude of waveforms measured at the actual sensor terminal. As a result, the simulation generated a nominal waveform with a frequency of 60.10 Hz, a period of 16.638 ms, and an

amplitude of 0.1733 V. These parameters were tuned based on analytical equations for voltage gain and cutoff frequency.

As shown in Figures 3 and 4, the simulated signals closely resemble those collected from the actual device. To quantify similarity, signals were z-scored, coarsely aligned by cross-correlation, and finely phase-aligned at the dominant frequency, then evaluated with Dynamic Time Warping (DTW; path-length-normalized alignment cost under non-linear warps, lower is better) and Time-Delay Embedding (TDE; Takens embedding with $m = 4$, τ from the first non-positive autocorrelation; similarity via the symmetric Hausdorff distance, lower is better).

Using normal baselines, acceptance thresholds were set to the 90th percentile of the within-class distributions (P_{90} , $N = 32$): $\theta_{DTW} = 3.37 \times 10^{-3}$ and $\theta_{TDE} = 0.912$. On a representative pair, the normalized DTW cost was 9.11×10^{-4} and the TDE Hausdorff distance was 0.480, both below their respective thresholds (Table I), indicating that the PSPICE-generated waveforms are sufficiently close to real measurements and are appropriate for training and evaluation.

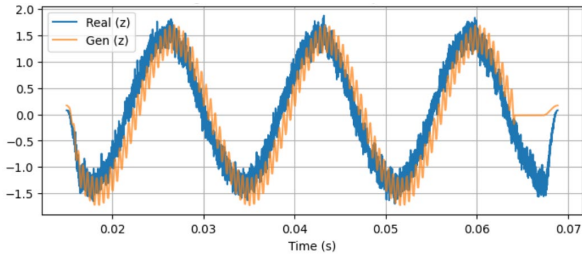


Figure 3. Voltage waveform overlap of real and gen signal.

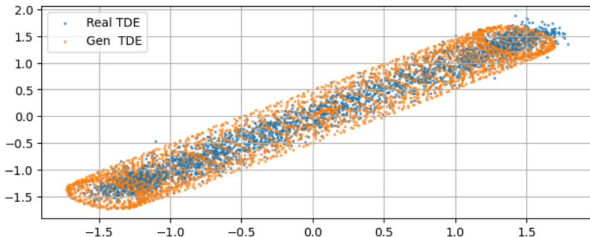


Figure 4. TDE scatter ($m=4$, $\tau=20$).

Table I. DTW/TDE similarity and acceptance thresholds.

Metric	Value	Acceptance criterion
Normalized DTW cost	9.11×10^{-4}	$\leq 3.37 \times 10^{-3}$ (P_{90} , $N=32$)
TDE symmetric Hausdorff ($m=4$, τ via ACF)	0.480	≤ 0.912 (P_{90} , $N=32$)

2.2. Waveform Generation

To train and evaluate the classification model, 200 samples of 300 ms waveform data were generated for

four scenarios: normal, impact, cable vibration, and circuit fault. Simulations were conducted using PSPICE and OrCAD Capture CIS, and some signals were synthesized through pattern-based post-processing.

Table II. 1st-Stage Class distribution.

1st Stage Class	Sample Count
Normal	10
Impact	10
Cable Vibration	10
Circuit Fault	170
Total	200

Normal signals were extracted from the analog circuit designed in Section 2.1, with $\pm 5\%$ noise added to simulate the variability observed in real measurements. Impact signals were generated by adding a VSIN source to the normal circuit output via a sum component. The VSIN source was configured with a 1 ms time delay (TD) to replicate the initial onset observed in actual impact events and had the following parameters: damping factor (DF) = 20, amplitude (VAMPL) = 0.618 V, frequency (FREQ) = 1000 Hz, and AC magnitude = 5. This configuration produced a pronounced high-amplitude transient (~ 600 mV) that is substantially higher than the nominal signal level, followed by a decaying oscillation characteristic of real impact-induced responses.

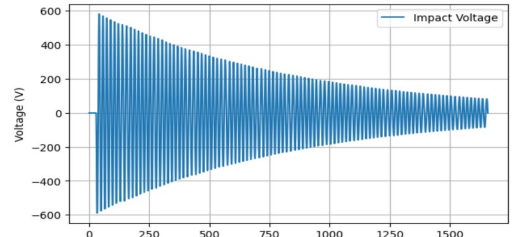


Figure 5. Impact signal voltage waveform.

Since cable-vibration waveforms are not readily reproducible in PSPICE, we emulated them by superimposing short, spike-like transients on normal signals. Each spike was randomly positioned and lasted for 30 samples, with amplitudes ranging from 0.02 to 0.07. The small amplitude was deliberately chosen to make class discrimination more challenging. The design reflects the characteristics commonly observed in real-world cable vibration scenarios.

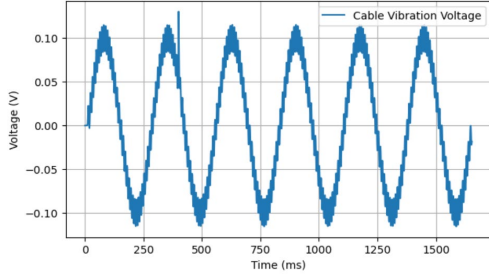


Figure 6. Cable vibration signal voltage waveform.

Circuit fault signals were generated via PSPICE simulations by defining 17 distinct fault classes. For each class, 10 representative waveforms were produced. Introducing 5% variations on noise and fault values, total 170 samples were collected. Depending on the fault location, the signals were categorized into two groups: charge amplifier-related and filter-related faults. A detailed summary of these fault types is provided in Table III.

Table III. 2nd-stage fault classes and parameters.

Class	Sub Label	Sample Counts	Baseline value	Fault value
Charge Amplifier	R1-high	10	10 k Ω	30 k Ω
	R1-low	10	10 k Ω	100 Ω
	R1-short	10	10 k Ω	—
	R1-open	10	10 k Ω	—
	R2-short	10	1 k Ω	—
Noise Filter	C1-high	10	50 nF	500 nF
	C1-low	10	50 nF	5 nF
	C1-short	10	50 nF	—
	C1-open	10	50 nF	—
	C2-high	10	1 nF	100 nF
	C2-low	10	1 nF	0.1 nF
	C2-short	10	1 nF	—
	C2-open	10	1 nF	—
	R4-high	10	3 k Ω	30 k Ω
	R4-low	10	3 k Ω	100 Ω
	R4-short	10	3 k Ω	—
	R4-open	10	3 k Ω	—

3. Signal Classification Framework

3.1. Overview of Two-Stage Signal Classification Framework

A two-stage hierarchical classification framework was developed using tree-based machine learning methods to process sensor signals. In each stage, the one-dimensional time-series signals are transformed into two-dimensional time–frequency representations and then flattened into fixed-length vectors for input to the classifier.

In the first stage, STFT-derived features are fed into an XGBoost model to classify the signals into four categories: normal, cable vibration, impact, and circuit

fault. Only signals identified as circuit faults proceed to the second stage.

In the second stage, CWT is applied to extract more fine-grained features, which are then classified by a Random Forest model into either charge amplifier faults or noise filter faults. This two-stage design reduces false alarms in the first stage by improving discrimination between impact and non-impact events, and specifies the cause of circuit faults in the second stage, thereby enhancing the reliability of signal interpretation in piezoelectric sensor systems. XGBoost and Random Forest—both tree-based learning models—offer robust performance even with relatively small training datasets.

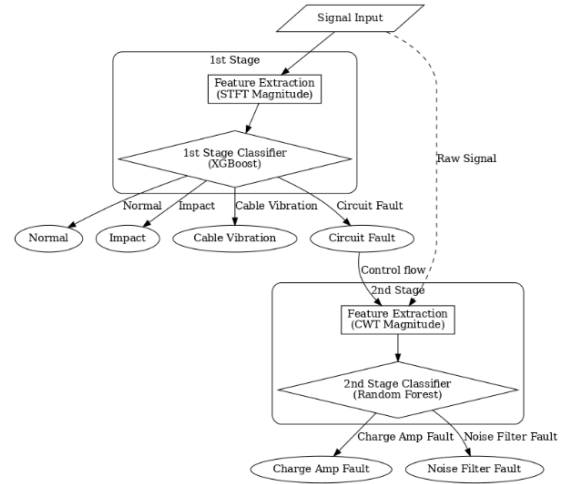


Figure 7. Classification schema.

3.2. First-stage Classification

In the first stage, the raw time-series signals collected from the sensors are processed using STFT to extract time–frequency domain features. STFT is well-suited for capturing changes in frequency content over short time windows, enabling precise alignment of transient and stationary components.

The STFT parameters are summarized in Table IV. The frequency range was limited to components below 300 Hz to focus on the dominant energy band of the signals. The resulting spectra were log-scaled and normalized, and for each frequency bin, statistical features—mean, standard deviation, maximum, minimum, skewness, and kurtosis—were calculated. The resulting feature vector was zero-padded to a fixed length of 3,000 before being passed to the classifier.

The extracted features were classified using XGBoost, a gradient boosting–based ensemble method known for its strong generalization ability and capacity to model complex, nonlinear decision boundaries. The hyperparameters used for XGBoost are listed in Table V. This stage outputs one of four classes: normal, cable vibration, impact, or circuit fault. Only signals labeled as circuit faults are passed to the second stage.

Table IV. STFT parameter settings.

Parameter	Value	Description
Sampling frequency	1600 Hz	Acquisition rate of raw signals
Window length	256	Number of samples per window
Overlap	128	Overlapping samples between windows
Frequency range	≤ 300 Hz	Retained frequency components

Table V. XGBoost parameter settings.

Parameter	Value	Note
use_label_encoder	False	Disable deprecated label encoder
eval_metric	mlogloss	Multiclass log loss evaluation
Random_state	42	Reproducibility
Others	Default	XGBoost default settings

3.3. Second-stage Classification

The second stage further analyzes the signals classified as circuit faults in order to identify the underlying cause. CWT is applied to these signals to extract time–frequency features with variable resolution, enabling precise characterization of both transient and noise-related patterns. This makes CWT effective in distinguishing between faults originating from the charge amplifier and those from the noise filter.

The CWT parameters are summarized in Table VI. The wavelet transform was performed using the Morlet mother wavelet with integer scales ranging from 1 to 30, providing coverage of both fine and coarse resolutions. For each scale, the mean, standard deviation, maximum, and minimum of the wavelet coefficients were calculated, and the results were flattened into a one-dimensional feature vector.

These features were then classified using a Random Forest model, an ensemble of decision trees that improves classification accuracy through majority voting. Random Forest is robust to noise and performs well even on relatively small datasets, making it suitable for the conditions of this study. The hyperparameters used for the Random Forest classifier are listed in Table VII.

Table VI. CWT parameter settings.

Parameter	Value	Description
Mother wavelet	Morlet (morl)	Wavelet type used for decomposition
Scale range	1–30	Integer scales for multi-resolution analysis
Statistical features	Mean, Std, Max, Min	Extracted from wavelet coefficients

Table VII. Randomforest parameter settings.

Parameter	Value	Note
Random_state	42	Reproducibility
Others	Default	scikit-learn default settings

4. Training & Test Results

4.1. Training

To assess the learning behavior of the proposed classifiers, this study utilized loss curves and t-SNE embeddings. The dataset was stratified and split into 80% training and 20% testing. As shown in Figure 8, the validation loss of the XGBoost model converged to below 0.01 around boosting round 60, indicating stable training. For qualitative visualization, the feature vectors used by the Random Forest stage were projected using t-SNE (Figure 9). The clusters may appear diffuse due to the aggregation of 17 fault subclasses into two top-level categories, but are well separated at the subclass level, indicating effective model learning.

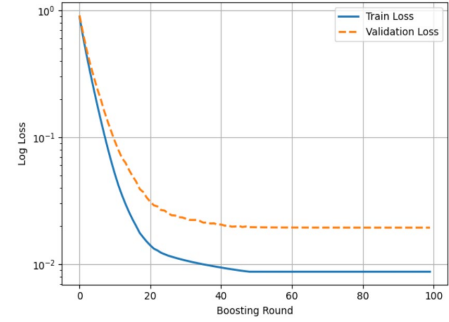


Figure 8. XGBoost training and validation loss.

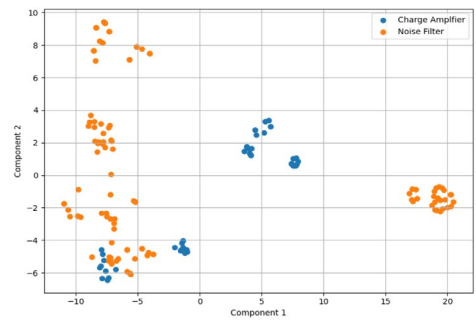


Figure 9. Randomforest t-SNE projection of features (aggregated into charge-amplifier vs. noise-filter groups).

4.2. Test Results

The two-stage classification framework was evaluated with STFT features and an XGBoost classifier in the first stage, and CWT features and a Random Forest classifier in the second stage. Because the held-out test set is class-imbalanced (e.g., some first-stage classes have only two samples), performance is summarized using overall

accuracy together with macro- and weighted-averaged precision/recall/F1.

At the First stage (STFT + XGBoost), the classifier correctly classified 40 out of 40 samples (100.0%) on the test set, with macro F1 = 1.00 and weighted F1 = 1.00 (Table VIII). Notably, impact was cleanly separated from circuit fault, a pair often confounded in threshold-based LPMS pipelines. On the Second stage (CWT + Random Forest), the classifier correctly classified 34 out of 34 samples (100.0%), with macro F1 = 1.00 and weighted F1 = 1.00 (Table IX), distinguishing charge-amplifier from noise-filter faults without error.

To compare classification performance, a 1D-CNN was also trained on raw waveforms (no explicit time–frequency features). The model achieved 90% accuracy at the first stage (macro F1 = 0.49, weighted F1 = 0.85) and 94% accuracy at the second stage (macro F1 = 0.93, weighted F1 = 0.94) (Tables XI–XII). For validation, Stage-2 performance of the CNN baseline was computed on a fixed set of 24 ground-truth circuit-fault samples, bypassing the usual routing of only Stage-1 “circuit-fault” predictions. The first-stage degradation is concentrated in small-support classes (Normal, Cable Vibration; $n = 2$ each). These results show that a two-stage classification framework with STFT/CWT features and tree-based models outperform an end-to-end CNN under limited and imbalanced data—improving accuracy by +10 percentage points in Stage 1 (100% vs 90%) and +6 points in Stage 2 (100% vs 94%), with large gains in macro F1 for small-support classes.

Table VIII. 1st-stage classifier results.

	Precision	Recall	F1-score	support
Normal	1.00	1.00	1.00	2
Impact	1.00	1.00	1.00	2
Cable	1.00	1.00	1.00	2
Vibration				
Circuit	1.00	1.00	1.00	34
Fault				
Accuracy			1.00	40
Macro	1.00	1.00	1.00	40
avg				
Weighted	1.00	1.00	1.00	40
avg				

Table IX. 2nd-stage classifier results.

	Precision	Recall	F1-score	support
Charge	1.00	1.00	1.00	10
Amplifier				
Noise	1.00	1.00	1.00	24
Filter				
Accuracy			1.00	34
Macro	1.00	1.00	1.00	34
avg				
Weighted	1.00	1.00	1.00	34
avg				

Table X. 1D-CNN architecture used (both stages).

Layer	Details	Output channels / size (for input length 3000)
Conv1d + ReLU	in=1, out=16, kernel=64, stride=16, pad=0	16 × 184
MaxPool1d	kernel=4, stride=4	16 × 46
Conv1d + ReLU	in=16, out=32, kernel=16, stride=4, pad=0	32 × 8
MaxPool1d	kernel=2, stride=2	32 × 4
Conv1d + ReLU	in=32, out=64, kernel=4, stride=1, pad=0	64 × 1
Flatten	—	64
Linear + ReLU	64 → 128	128
Dropout	p = 0.5	128
Linear (output)	Stage-1: 128 → 4 classes; Stage-2: 128 → 2 classes	—

Table XI. 1st-stage 1D-CNN classification report.

	Precision	Recall	F1-score	support
Normal	0.00	0.00	0.00	2
Impact	1.00	1.00	1.00	2
Cable	0.00	0.00	0.00	2
Vibration				
Circuit	0.89	1.00	0.94	34
Fault				
Accuracy			0.90	40
Macro	0.47	0.50	0.49	40
avg				
Weighted	0.81	0.90	0.85	40
avg				

Table XII. 2nd-stage 1D-CNN classification report.

	Precision	Recall	F1-score	support
Charge	0.83	1.00	0.91	10
Amplifier				
Noise	1.00	0.92	0.96	24
Filter				
Accuracy			0.94	34
Macro	0.92	0.96	0.93	34
avg				
Weighted	0.95	0.94	0.94	34
avg				

5. Conclusion

This work presents a two-stage classification framework to improve piezoelectric signal interpretation accuracy in nuclear power-plant monitoring, with a focus on LPMS. A PSPICE-based reconstruction of the analog front-end enabled high-quality data generation for both normal and abnormal scenarios. STFT and CWT time–frequency features were extracted and fed to XGBoost (first stage) and Random Forest (second stage), respectively. On a held-out test set, the framework classified 40/40 samples (100%) in Stage 1 and 34/34 (100%) within the circuit-fault subset in Stage 2, while a 1D-CNN trained on raw waveforms reached 90% and 94%, respectively. These findings clearly indicate the superiority of tree-based models with engineered time–frequency features over an end-to-end CNN under small and imbalanced data, reducing false alarms (impact vs non-impact) and improving diagnostic specificity (charge-amplifier vs noise-filter faults). This work is applicable to safety-critical piezoelectric sensor systems, including integrated small modular reactors (i-SMRs), where accurate fault classification supports safer operation. Expanding this study, the authors plan to integrate field measurements, explore hybrid pipelines (tree-based + deep backbones), and broaden robustness studies across noise/operating conditions.

Acknowledgements

This work was supported by the Innovative Small Modular Reactor Development Agency grant funded by the Korea Government (MIST) (No. RS-2023-00258052).

REFERENCES

- [1] M. A. Mangi, H. Elahi, A. Ali, H. Jabbar, A. B. Aqeel, A. Farrukh, S. Bibi, W. A. Altabey, S. A. Kouritem, and M. Noori, Applications of Piezoelectric-Based Sensors, Actuators, and Energy Harvesters, Sensors and Actuators Reports, Vol. 9, 2025.
- [2] V. Kannan, D. V. Dao, and H. Li, Detection of Signal Integrity Issues in Vibration Monitoring Using One-Class Support Vector Machine, Journal of Vibration Engineering & Technologies, Vol. 12, 2024.
- [3] W. D. Rhodes and D. W. Langenberg, Export Control Guide: Loose Parts Monitoring Systems for Nuclear Power Plants, ORNL/TM-2012/601, Oak Ridge National Laboratory, 2012, pp. 13, 16–19, 25.
- [4] International Atomic Energy Agency, On-line Monitoring for Improving Performance of Nuclear Power Plants, Part 2: Process and Component Condition Monitoring and Diagnostics, IAEA Nuclear Energy Series No. NP-T-1.2, Vienna, 2008, pp. 24–26.
- [5] T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, Proc. 22nd ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining, pp. 785–794, 2016.
- [6] L. Breiman, Random Forests, Machine Learning, Vol. 45, No. 1, pp. 5–32, 2001.
- [7] J. B. Allen, Short-Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform, IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 25, No. 3, pp. 235–238, 1977.
- [8] C. Torrence and G. P. Compo, A Practical Guide to Wavelet Analysis, Bulletin of the American Meteorological Society, Vol. 79, No. 1, pp. 61–78, 1998.