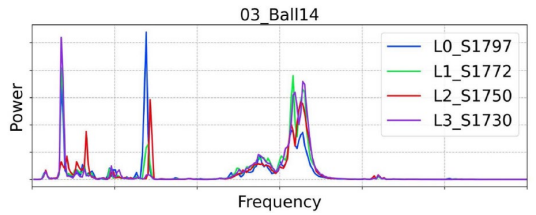
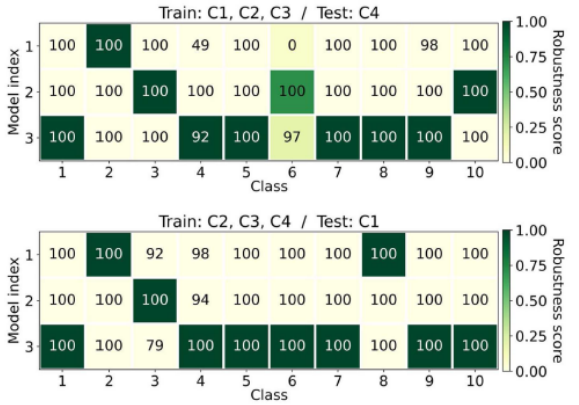
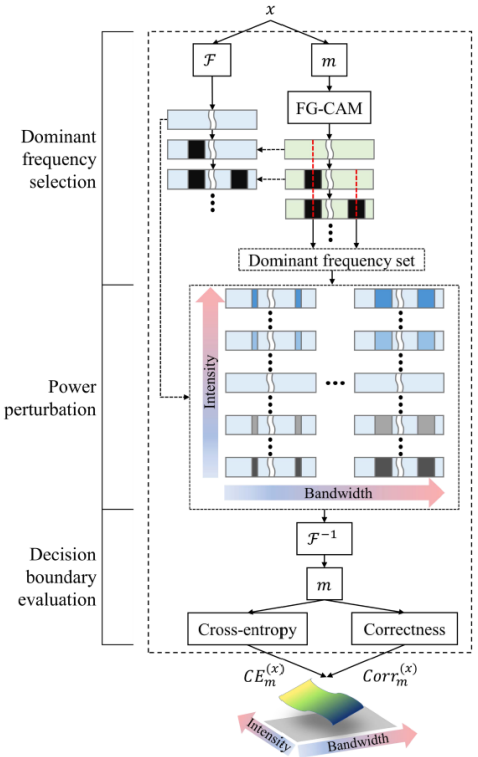
Robust and Explainable Fault Diagnosis with Power-Perturbation-Based Decision Boundary Analysis of Deep Learning Models

Background and Objective

Faults in rotating machines or their components can be diagnosed through vibration data analysis. This research aims **to explain the decision criteria, decision boundaries, and robustness of black-box 1D-CNNs**. The proposed method goes beyond simple classification by providing reasoning behind the model's decisions and insights into its generalization capability under unseen working conditions.

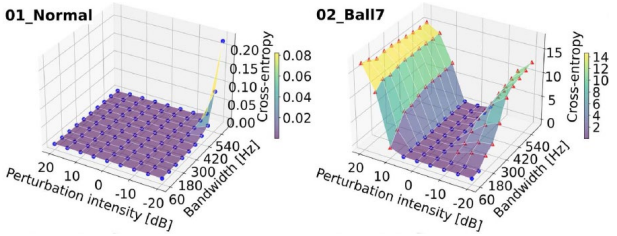






Methods

- Robustness evaluation based on power perturbations around the dominant frequencies and corresponding changes in cross-entropy



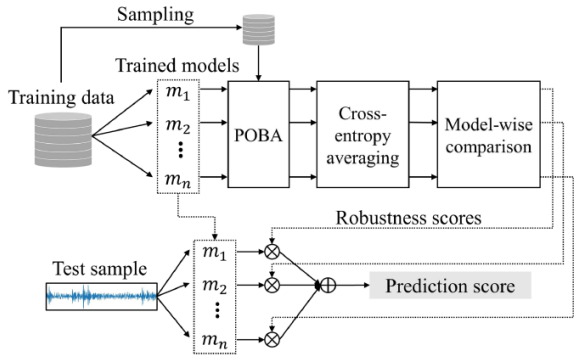
$$A_c^{(x)}(f) = \mathcal{F}\{a_c^{(x)}\}$$

$$a_c^{(x)} := \sum_{k=1}^{N} \max(\hat{\gamma}_c^{(k)}(x), 0) F^{(k)}(x)$$

$$\gamma_c^{(k)}(x) = \frac{1}{D} \sum_{i=1}^{D} \frac{\partial y_c(x)}{\partial F_i^{(k)}(x)}$$

A frequency-domain-based gradient-weighted class activation mapping method was used to analyze dominant frequencies.

- Robustness-based ensemble model, where each model is weighted by class-wise robustness scores



Results

1. Intuitive visual explanations of model's robustness
2. Validated effectiveness of the robustness scores and ensemble strategy

(Upper figure: Accuracy under unseen working conditions aligns with the calculated robustness scores)

Minseon Gwak