2025 Global Reliability and Prognostics and Health Management Conference (PHM-Xian) | 979-8-3315-2675-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/PHM-Xian66756.2025.11427542

Neurons-State-Attention-Driven Trustworthy Framework: An Interpretable Networks for Wind Turbine Gearbox Fault Diagnosis

Chun Zhang
College of Engineering, Shantou University, Shantou, Guangdong, China
chunzhang.st@foxmail.com

Peng Chen*
College of Engineering, Shantou University, Shantou, Guangdong, China
Key Laboratory of Intelligent Manufacturing Technology, Ministry of Education of China, Shantou, Guangdong, China
pengchen@alu.uestc.edu.cn

Yuhao Wu
College of Engineering, Shantou University, Shantou, Guangdong, China
507663149@qq.com

Meng Rao
Department of Mechanical Engineering, Tsinghua University, Beijing, China
Qingdao Mingserve Technology Ltd., Qingdao, Shandong, China
mrao@tsinghua.edu.cn

***Abstract*—In industrial wind turbine systems, gearbox fault diagnosis faces significant challenges due to complex operational conditions and the limitations of traditional diagnostic approaches, particularly in reliability assessment of deep learning models. Current methodologies often rely on black-box predictions, leading to uncertainty in fault detection accuracy. To address these challenges, this study proposes a novel Neurons-State-Attention-Driven (NSAD) framework for enhanced gearbox fault diagnosis in wind turbine systems. The framework introduces an innovative approach to quantifying model prediction confidence through systematic analysis of neuronal activation states and patterns, incorporating a sophisticated multi-component loss function that combines cross-entropy, multi-center, contrastive, and variance losses. This comprehensive diagnostic methodology demonstrates significant potential in advancing wind turbine gearbox diagnostics by improving both out-of-distribution detection rates and reducing in-distribution false positives, while providing interpretable confidence metrics for reliable fault detection.**

Keywords-wind turbine; gearbox; fault diagnosis; interpretable neural networks; trustworthy AI; vibration signal analysis.

I. INTRODUCTION

Wind turbine gearboxes, serving as the fundamental power transmission mechanism in wind power systems, are pivotal in determining operational reliability and energy conversion efficiency [1]. These critical components, operating under sustained complex mechanical stresses, exhibit the highest failure rates and maintenance costs within the turbine assembly, potentially leading to substantial economic losses and safety hazards. Consequently, the implementation of sophisticated fault diagnosis systems, enhanced by modern sensor technology and prognostic health management frameworks, has become instrumental in facilitating real-time monitoring and preemptive fault detection, thereby mitigating catastrophic failure risks and optimizing operational continuity [2], [3], [4].

In the field of fault diagnosis, researchers have advanced both traditional vibration signal analysis and deep learning approaches[5]. Chen et al. [6] proposed an adaptive signal processing regime with three innovations—real-time transient shift identification, a Dynamic Markov Transition Frequency with Adaptive Peak Rates (DMTF-APR) model, and a Multi-Period Weighted Average Framework (MPWAF)—to enhance wind turbine fault diagnosis under variable operating conditions and noise interference. Feng et al. [7] proposed an encoder-free Vold-Kalman filter combined with higher-order energy separation (HOES) for accurate time-varying fault frequency extraction in planetary gearboxes under nonstationary conditions, enabling effective detection of both distributed and localized gear faults. Chen et al. [8] proposed a Markov-based signal transition modeling method combined with wavelet transforms and an amplitude interference-limiting mechanism to enhance demodulation band selection and improve fault diagnosis accuracy for rolling bearings under transient noise interference. Despite their success in wind turbine fault diagnosis, these techniques are constrained by both the

* Corresponding author: Peng Chen (pengchen@alu.uestc.edu.cn or dr.pengchen@foxmail.com)

This research was supported by the National Natural Science Foundation of China (Grant 52105111), the Guangdong Basic and Applied Basic Research Foundation (Grant 2025A1515012256), the Shantou University (STU) Scientific Research Initiation Grant (NTF21029), and the Industry-Academia Cooperation Project from the Guangdong Institute of Special Equipment Inspection and Research Shunde Branch (XTJ-KY01-202503-030).

complexity of their feature extraction methodology and their heavy dependence on expert intervention. With breakthrough advancements in deep learning and artificial intelligence, researchers have revolutionized this field by developing automated, data-driven approaches for fault diagnosis and health condition monitoring. Jiang et al. [9] proposed a multiscale convolutional neural network (MSCNN) for end-to-end fault diagnosis of wind turbine gearboxes, which simultaneously extracts multiscale features from raw vibration signals and classify faults, outperforming traditional CNNs and multiscale feature extraction methods. Chen et al. [10] proposed a Metric-Guided Graph Contrastive Learning (MGCL) framework with feature-decoupled pre-training, hybrid distance metrics, and two-stage training to overcome scarcity and domain shift challenges in planetary gearbox fault diagnosis, improving robustness and diagnostic accuracy. While these methods successfully eliminate the need for labor-intensive feature engineering, they are often criticized for their reduced model interpretability, which can undermine the transparency and trustworthiness of prediction outcomes, particularly in high-stakes applications. To enhance the reliability of deep learning, extensive research focuses on making models "know when they do not know," thereby improving robustness. A key approach is distinguishing between in-distribution (ID) and out-of-distribution (OOD) inputs to evaluate the trustworthiness of model predictions. Gal et al. [11], [12] proposed a framework linking dropout training in deep neural networks to approximate Bayesian inference in Gaussian processes (BNN), enabling practical uncertainty estimation. Lakshminarayanan et al. [13] proposed a simple and scalable method using deep ensemble (DE) for estimating predictive uncertainty in deep neural networks. However, both BNN and DE require multiple forward passes, leading to relatively low computational efficiency. Sensoy et al. [14] proposed a novel uncertainty-aware neural network framework named Evidential Deep Learning (EDL) based on subjective logic and Dirichlet distributions that directly models prediction confidence without requiring Bayesian inference. However, since the evidence in EDL is entirely generated by black-box models, it still lacks reliability.

To overcome the limitations of existing methodologies, we propose the Neurons-State-Attention-Driven (NSAD) framework, which offers a novel mechanism to quantify the model's prediction confidence based on the activation states of its neurons. By identifying key neurons according to the activation intensity of samples, the framework constructs activation paradigms through the statistical aggregation of activation patterns across all classes. During the validation phase, prediction confidence is assessed by comparing the activation states of these key neurons within a given sample against the corresponding activation paradigms. To further enhance the discriminability of these paradigms, we introduce a subclass partitioning strategy alongside a composite loss function that integrates cross-entropy loss, multi-center loss, contrastive loss, and variance loss. This design effectively improves intra-class compactness and inter-class separability, resulting in a robust neural network with a high OOD detection rate and a significantly reduced ID false positive rate.

II. Related Work

Contemporary research on neural network confidence assessment predominantly focuses on three key methodological approaches: Bayesian Neural Networks (BNN), Deep Ensembles (DE), and Evidential Deep Learning (EDL), which collectively analyze model outputs to evaluate prediction reliability and uncertainty quantification in artificial intelligence systems. These methodologies have substantially enhanced our understanding of model behavior and decision-making processes, contributing to more reliable machine learning applications.

A. BNN

The core characteristic of BNN lies in the representation of its model weights as probability distributions rather than fixed values. Each forward pass requires sampling from these weight distributions to determine the specific weight values used for that inference. Consequently, passing the same input through the same BNN multiple times typically yields different output results. BNN quantifies epistemic uncertainty by computing the variance of multiple predictions for the same sample, while aleatoric uncertainty is estimated via the entropy of the prediction results:

$$U_e = \frac{1}{T}\sum_{t=1}^{T}(p_t - \frac{1}{T}\sum_{t=1}^{T}p_t)^2 \quad (1)$$

$$U_a = \frac{1}{T}\sum_{t=1}^{T}H(p_t) \quad (2)$$

where U_e denotes epistemic uncertainty, U_a denotes aleatoric uncertainty, T denotes the number of forward passes, p_t denotes the outputs of i^{th} forward pass.

B. DE

The core of this method lies in independent training multiple distinct deterministic neural networks. Each model converges to a different local minimum of the loss function during training, thereby forming diverse predictive perspectives. For the same input sample, all member models within the ensemble make independent predictions. The mechanism by which DE quantifies uncertainty is fundamentally identical to that of BNN: both leverage multiple predictions for a single input sample to estimate uncertainty.

C. EDL

The core innovation of this method lies in directly modeling prediction uncertainty by modifying the neural network's output from probability vectors to higher-level evidence parameters $\{e_i\}_{i=1}^{K}$ that characterize subjective logic. These evidence parameters are used to construct a Dirichlet distribution. This distribution enables the calculation of the probability that a

sample belongs to each class, as well as the output of the model's uncertainty regarding this result:

$$p = \frac{e}{\sum_{i=1}^{K}(e_i + 1)} = \frac{e}{K + \sum_{i=1}^{K} e_i} \tag{3}$$

$$U = \frac{K}{K + \sum_{i=1}^{K} e_i} \tag{4}$$

where p denotes the distribution of the sample, U denotes uncertainty, K denotes the number of classes.

These methods achieve uncertainty quantification in neural networks to varying degrees but still face notable limitations. Both BNN and DE methods require multiple forward passes per input to estimate uncertainty, resulting in significantly higher computational overhead compared to conventional neural networks. Although EDL demonstrates a significant advantage in computational efficiency, its approach lacks full rigor. This is because the evidence used for quantifying uncertainty is entirely dependent on the neural networks, rendering the evidence inherently unreliable.

Based on these traditional approaches, several recent studies have been proposed. Ren et al. [15] proposed a unified uncertainty-aware deep learning framework (UU-DLF) that achieves OOD detection by modeling the outputs as multivariate Gaussian distributions and employing deep ensemble techniques. UU-DLF defines aleatoric uncertainty and epistemic uncertainty in the following formula and discriminates OOD samples using uncertainty thresholds.

$$U_a = \frac{1}{K}\sum_{k=1}^{K}\prod_{c=1}^{C}\hat{\sigma}^{i,c^2} \tag{5}$$

$$U_e = \frac{1}{K}\sum_{k=1}^{K}(\hat{\boldsymbol{\mu}}_k - \bar{\boldsymbol{\mu}})^2 = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\boldsymbol{\mu}}_k - \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\mu}}_k\right)^2 \tag{6}$$

where K is the number of models in the deep ensemble, C is the number of classes, $\hat{\sigma}^{i,c^2}$ denotes the diagonal elements of the covariance matrix in the multivariate Gaussian distribution.

However, UU-DLF still suffers from low computational efficiency. Zhang et al. [16] proposed MM-DE build upon deep ensembles by learning a set of sub-models that produce highly consistent predictions for ID samples while maximizing disagreement on OOD samples. MM-DE defines the consistency level between two models in the following formula:

$$a_{ij} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\left(p_{i,k,1} - p_{j,k,1}\right)^2 + \cdots + \left(p_{i,k,L} - p_{j,k,L}\right)^2} \tag{7}$$

where K is the number of samples and L is the number of classes.

However, this method introduces OOD samples during training, which is often infeasible in real-world applications. Wei et al. [17] improves upon EDL by generating pseudo-OOD samples and incorporating an abstention class, but its mechanism for generating pseudo-OOD samples is limited and can only cover a small portion of OOD instances.

III. Proposed Method

The proposed Neurons-State-Attention-Driven Trustworthy Framework (NSAD) integrates cognitive neuroscience principles, particularly Hebbian Theory and Neuronal Activation Paradigm, to establish a robust foundation for trustworthy machine learning. This comprehensive framework encompasses sub-class classification, subclass center updates, and confidence quantification methodologies, unified through a specialized loss function design. By leveraging neuron activation patterns for confidence assessment, NSAD achieves an optimal balance between computational efficiency and explanatory transparency, enabling more reliable prediction confidence evaluations through internal activation state analysis, thereby advancing the development of accountable and transparent neural network systems.

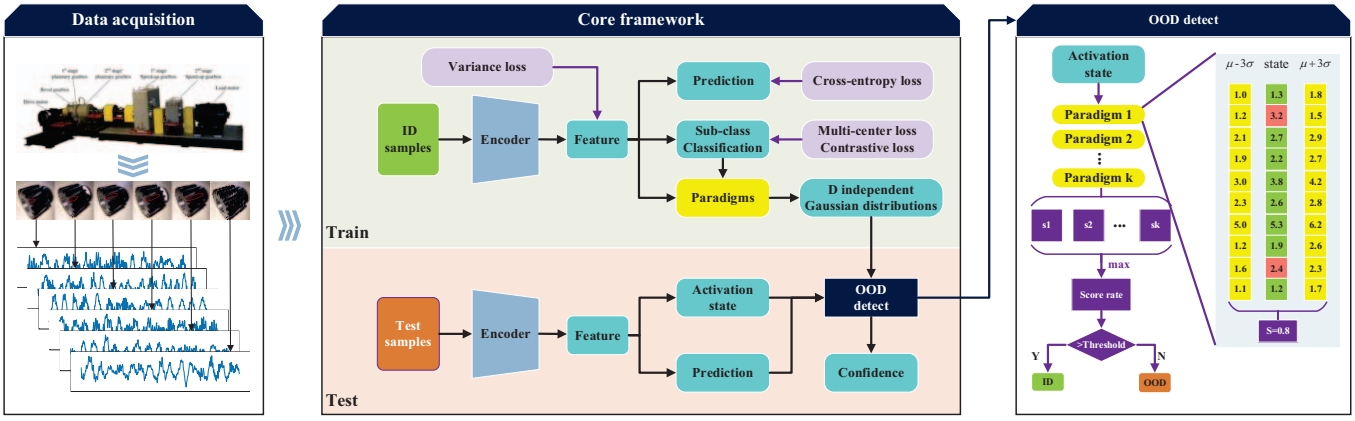

Figure 1. The proposed Neurons-State-Attention-Driven trustworthy framework.

A. Hebbian Theory and Neuronal Activation Paradigm

The foundational premise of our research stems from Hebbian theory [18], [19] in neuroscience, which posits that neural pathways are strengthened through repeated activation, colloquially expressed as "neurons that fire together, wire together." This neurobiological principal manifests in cognitive processes where successful problem-solving iterations reinforce specific neural circuits, subsequently facilitating more

efficient processing of analogous challenges through established pathways.

Building upon this neurological principle, we extend the concept to artificial neural networks, proposing that training samples within a specific category consistently activate predetermined neuronal patterns. Conversely, novel or out-of-distribution samples tend to generate irregular activation sequences that deviate significantly from these established categorical paradigms. This theoretical proposition has been substantively validated through our experiments utilizing ResNet-18 architecture as the primary model backbone.

The architecture of the proposed method is illustrated in Figure 1, with ResNet-18 serving as the backbone of the model. The $Feature^{(d)}$ represent high-dimensional feature vectors extracted by the model, where the values within these vectors correspond to neuronal activation states. To model the activation characteristics of neurons for samples belonging to the same class, we represent them as d independent Gaussian distributions:

$$\mu_i = \frac{1}{n_i}\sum_{k=1}^{n_i} Feature(x_k) \tag{8}$$

$$\sigma_i{}^2 = \frac{1}{n_i}\sum_{k=1}^{n_i} (Feature(x_k) - \mu_i)^2 \tag{9}$$

where μ_i denotes the average activation intensity of the i^{th} class, $\sigma_i{}^2$ denotes the variance of the activation value of the i^{th} class, n_i is the number of samples belonging to the i^{th} class. $\mu \pm 3\sigma$ represents the standard activation paradigm for the type of sample.

Empirical evidence demonstrates that neural network activations exhibit a distinctive pattern wherein samples from identical classes predominantly engage a limited subset of neurons, while the majority of neuronal units display minimal activation patterns[20], [21]. The inclusion of these minimally activated neurons not only introduces computational inefficiencies but may also compromise the model's predictive accuracy by potentially incorporating spurious feature representations. To address this limitation, this methodology implements a selective approach by identifying and retaining the top $1/p$ neurons demonstrating the highest significance for each class, where significance is quantified through mean activation intensity metrics. This optimization ultimately preserves d/p neurons for OOD detection purposes, with the activation characteristics of this refined neuronal subset being represented by μ_{i-key} for mean intensity and σ_{i-key} for standard deviation, respectively.

B. Subclass Classification

It is a common challenge in classification tasks that a single broad class can be subdivided into finer subclasses. Samples belonging to different subclasses will exhibit distinct activation patterns, too. To further enhance model performance, it is necessary to adopt a finer-grained subclass structure. To achieve the goal of sub-class classification, this paper embeds a 3-dimensional tensor $centers$ within the model. The $centers$ stores $n_{class} * n_{sub}$ subclass center vectors. Each sample will be assigned to the nearest subclass center within its broad class.

According to the recommendation from SimCLR [22], [23], we do not directly use the $Feature^{(d)}$ for subclass partitioning and contrastive learning, as this could significantly interfere with the training of the main model. Therefore, the $Feature^{(d)}$ is passed through a $Projector$ (as shown in Figure 2) and yielding a $Projection^{(d')}$. This $Projection^{(d')}$ is then used for both subclass partitioning and contrastive learning. Due to the introduction of sub-class classification, the class activation mean and standard deviation are correspondingly refined. They are now denoted as $u_{i-j-key}$ and $\sigma_{i-j-key}$, representing the activation mean and standard deviation for the j^{th} subclass within the i^{th} broad class respectively.

Figure 2. Projector module: projects features for subclass partitioning and contrastive learning.

C. Subclass Center Update

During the model's training process, the continuous evolution of the feature space mapping necessitates corresponding adjustments to the subclass center positions. However, to maintain stable convergence dynamics and prevent potential disruptions to the learning trajectory, these positional updates must be implemented with careful consideration of their frequency. Thus, we establish a systematic approach by executing subclass center updates at predetermined intervals of $epoch_u$, striking a balance between adaptive representation and training stability.

The new subclass centers are determined using the K-means clustering algorithm [24]. To ensure smooth updates, this paper employs a momentum update strategy:

$$centers_{new} = \beta centers_{new} + (1-\beta) centers_{new} \tag{10}$$

To promote stability in the subclass center update process and mitigate oscillations induced by stochastic mini-batch fluctuations, we employ a conservative update coefficient, thereby constraining the step size and reducing the risk of divergence during training; specifically, we set this coefficient to 0.1, which offered a consistent balance between responsiveness and robustness across our experiments.

However, clustering does not assign fixed class labels, introducing uncertainty in the correspondence between new and old cluster centers. As shown in Figure 3, this ambiguity can lead to suboptimal update outcomes. To address this cluster center matching problem, this paper employs the Hungarian algorithm [25]. This ensures the rationality and accuracy of the centers update process, effectively preserving the consistency of cluster centers across iterations.

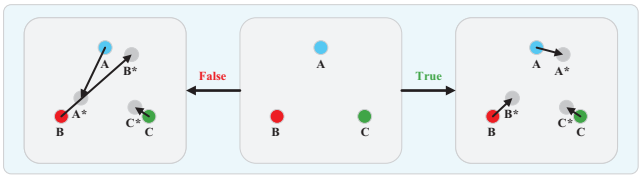

Figure 3. Subclass center matching problem.

D. Confidence Quantification for Prediction Results

In the testing phase, the methodology encompasses a comprehensive evaluation process wherein a sample undergoes forward propagation to generate its corresponding $Feature^{(d)}$ and logit, with the latter determining the initial prediction. To address potential prediction unreliability, this study implements a novel confidence quantification mechanism utilizing $Feature^{(d)}$, specifically focusing on the d/p most significant neurons unique to each sample. The refined representation, denoted as $Feature^{(d/p)}$, is derived by extracting activation values from these crucial neurons, and the confidence score rate is subsequently calculated by analyzing the proportion of elements within $Feature^{(d/p)}$ that fall within the statistical bounds of $\mu \pm 3\sigma$, wherein samples exhibiting confidence scores below a predetermined threshold are classified as Out-of-Distribution (OOD). Notably, this framework demonstrates robustness despite not requiring high precision in model subclass partitioning, as the confidence score computation involves evaluating $Feature^{(d/p)}$ against each subclass within the predicted broad class, ultimately assigning the maximum obtained confidence score as the definitive metric for the sample under examination.

$$score = max_{k=1}^{n_{sub}}\left(\sum_{t=1}^{d/p}\mathbf{1}(m_t \le F_t \le M_t)\right) \tag{11}$$

where $\mathbf{1}(\cdot)$ is an indicator function that takes the value 1 if and only if a specified condition is satisfied, M and m denotes $u_{i-j-key} \pm 3\sigma_{i-j-key}$ respectively, F denotes $Features^{(d/p)}$. Finally, calculating confidence score rate by dividing the score by d/p.

E. Loss Function Design

The proposed architecture aims to achieve accurate predictions while utilizing a projector to cluster similar category samples proximally. By minimizing $\sigma_{i-j-key}$ in the neuronal activation pattern, we establish stricter 3σ intervals, enabling effective discrimination of Out-of-Distribution samples through lower confidence scores without compromising In-Distribution sample performance, thus informing our loss function design.

1) Cross-entropy Loss

The implementation of cross-entropy loss functions serves a dual objective in the learning framework: primarily, it enables the model to achieve precise discrimination among established class categories, while simultaneously facilitating the optimization of feature space by ensuring that instances belonging to identical classes maintain proximal geometric relationships in their latent representations. This mathematical construction thereby promotes both classificatory accuracy and intra-class cohesion.

$$p = \frac{e^{logits}}{\sum_B e^{logits}} \tag{12}$$

$$L_{ce} = -\frac{1}{B}\sum_{k=1}\sum ylog(p) \tag{13}$$

where y is the label of the sample, B represents the number of samples in a batch.

2) Multi-center Loss

The model embeds $centers$ to achieve the classification of subclasses. The multi-center loss ensures accurate subclass partitioning by guiding the model to move samples of the j^{th} subclass closer to their corresponding center.

$$L_{mc} = \frac{1}{B}\sum_{i=1}^{n_{class}}\sum_{j=1}^{n_{sub}}\sum_{k=1}^{n_{ij}}\frac{1}{n_{ij}}\left\|pro_{i-j-k} - center_{i-j}\right\| \tag{14}$$

where $\|\cdot\|$ denotes Euclidean distance, pro_{i-j-k} denotes the result where a sample belong to the j^{th} subclass within the i^{th} broad class is projected by the $Projector$, n_{ij} is the number of samples belong to the j^{th} subclass within the i^{th} broad class in a batch.

3) Contrastive Loss

The contrastive learning loss facilitates intra-subclass cohesion by minimizing sample distances within subgroups. While analogous to multi-center loss, this mechanism operates distinctly. Implementation requires three key components: positive sample masking, target probability distribution, and similarity matrix computation.

$$M_{ij} = \begin{cases} 1, & (y_i = y_j)\ and\ (sub_i = sub_j)\ and\ (i \ne j) \\ 0, & other \end{cases} \tag{15}$$

where y_i is the label of the sample i and sub_i is the subclass label of the sample i.

$$T_{i\cdot} = \frac{M_{i\cdot}}{\sum_{k=j}^{B} M_{ij} + \epsilon} \tag{16}$$

where $M_{i\cdot}$ denote the i^{th} row of the M and $T_{i\cdot}$ denote the i^{th} row of the T, ϵ is a small value to avoid division by zero.

$$sim_{ij} = \frac{z_i^T z_j}{\tau} \tag{17}$$

$$z = \frac{pro(x)}{||pro(x)||} \tag{18}$$

where τ is a hyper-parameter.

$$L_{ct} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} T_{ij} log\frac{\exp(sim_{ij})}{\sum_{k=1}^{B}\exp(sim_{ik})} \tag{19}$$

4) Variance Loss Function

To make the 3σ range as small as possible and thereby make it easier to obtain low scores for the OOD samples, we introduced the variance loss:

$$L_{\sigma} = \frac{1}{n_{class} * n_{sub}}\sum_{i=1}^{n_{class}}\sum_{j=1}^{n_{sub}}\sigma_{i-j-key}^2 \tag{20}$$

Finally, the overall loss function is:

$$L = L_{ce} + \lambda_{mc}L_{mc} + \lambda_{ct}L_{ct} + \lambda_{\sigma}L_{\sigma} \tag{21}$$

where λ_{mc}, λ_{ct}, and λ_{σ} are all hyperparameters.

IV. Experimental Validation

A. Test rig and Dataset Description

To validate our proposed methodology, we employed the University of Alberta wind turbine gearbox dataset [26]. For experimental consistency, we selected a subset operating under constant conditions: 20 Hz rotational speed at 13% load capacity. The experimental setup and data acquisition system are depicted in Figure 4. The dataset encompasses six distinct conditions (one healthy state and five fault states, as shown in Figure 5), with vibration signals recorded through four sensor channels. These signals were preprocessed by segmenting into non-overlapping sequences, each containing 1024 data points, yielding 750 samples per condition class. To emulate real-world scenarios where anomalous conditions may emerge unexpectedly, we designated Faulty1 as OOD samples. The remaining classes were partitioned into training and testing sets using an 80-20 split ratio. The final test set incorporated both OOD samples and ID test data, enabling comprehensive evaluation of the model's performance under various operational conditions.

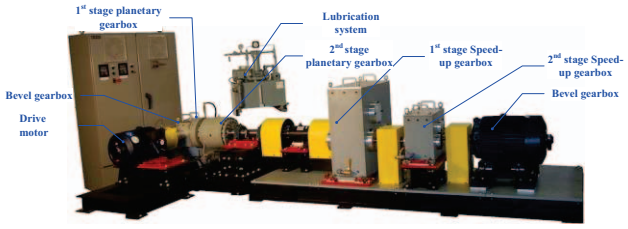

Figure 4. Alberta dataset acquisition equipment.

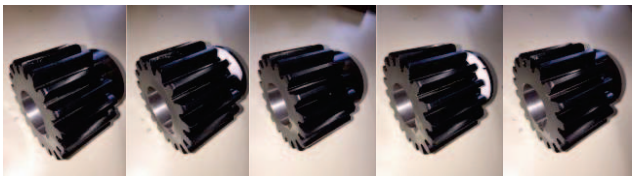

Figure 5. Physical faulty gears, gears from left to right are with fault level 1~5.

B. Experimental Configurations

In this experimental framework, we implement ResNet-18 as the foundational backbone architecture, which extracts high-dimensional features ($Feature^{(d)}$) with a dimensionality of 512. These features undergo subsequent projection through the Projector module, resulting in 128-dimensional representation vectors. The model's key hyperparameters are precisely calibrated: the neuron control parameter p is set to 8 (designating 1/8 of neurons per class as key neurons), the temperature coefficient τ maintains a value of 0.1, and the momentum factor β is established at 0.1. The loss function incorporates three essential hyperparameters: λ_{mc}, λ_{ct}, and λ_{σ}, set at 0.2, 0.2, and 0.05 respectively, with each broad classification category subdivided into 5 distinct subclasses.

The model's training methodology follows a sophisticated three-stage progressive approach. Stage I encompasses 100 epochs of training exclusively utilizing cross-entropy loss, establishing foundational learning patterns. Stage II augments the training process by incorporating both multi-center loss and contrastive learning loss components for an additional 100 epochs, enhancing feature discrimination. The final Stage III implements the complete loss function for 200 epochs, allowing for comprehensive model optimization. Throughout this process, center updates occur at 5-epoch intervals, and the OOD sample detection threshold is established at 0.8, meaning samples exhibiting confidence levels below 80% are classified as OOD instances.

The model's performance evaluation encompasses three critical dimensions: classification accuracy (ACC), OOD sample detection accuracy (ODA), and false alarm rate (FAR), complemented by detailed confusion matrix analysis. The ACC metric has been specifically modified to accommodate our experimental requirements - a test sample is deemed correctly classified only if it successfully passes both initial classification and subsequent OOD detection verification. This stringent criterion means that even if a sample is initially assigned to the correct class, it is considered misclassified if the model identifies it as an OOD instance during verification. Furthermore, the ACC computation incorporates OOD samples, with correct OOD detection positively contributing to the overall ACC score, ensuring a comprehensive evaluation of the model's discriminative capabilities.

$$ACC = \frac{ID_{correct} + OOD_{correct}}{ID_{all} + OOD_{all}} \tag{22}$$

where $ID_{correct}$ and $OOD_{correct}$ are the number of correctly classified ID and OOD samples, respectively. ID_{all} and OOD_{all} are the number of ID samples and OOD samples respectively.

ODA metric is employed to quantify the proportion OOD samples that are successfully detected by the model:

$$ODA = \frac{OOD_{correct}}{OOD_{all}} \tag{23}$$

FAR quantifies the proportion of ID samples that are incorrectly classified as OOD samples:

$$FAR = \frac{ID_{OOD}}{ID_{all}} \tag{24}$$

where ID_{OOD} is the number of ID samples that are incorrectly classified as OOD samples.

For the experimental setup, we extend the confusion matrix with an additional row and column specifically designed to represent OOD sample classification outcomes and detection status.

C. Experimental Results and Comparative Analysis

The feasibility of the proposed method hinges on identifying significant differences in neuron activation patterns across diverse sample types. Figure 6 suggest that the method's foundational mechanisms are valid. A critical element within

this framework is the design of the loss function, specifically variance loss. This function minimizes fluctuations in neuron activation, ensuring consistent activation intensities for identical sample categories. Consequently, a narrower ($\mu \pm 3\sigma$) range emerges, significantly enhancing the ability to assign low confidence scores to OOD samples, thus improving detection reliability.

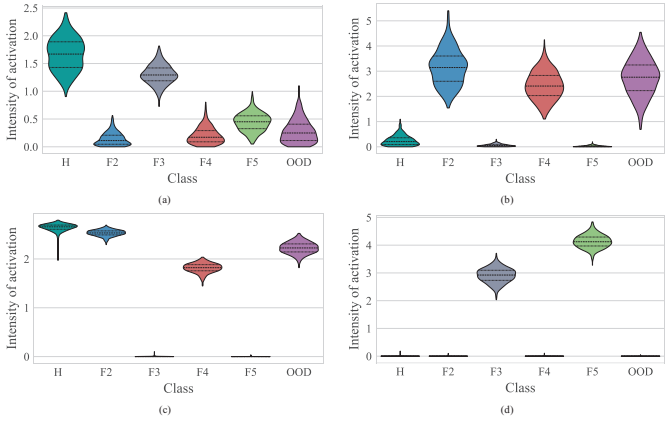

Figure 6. The activation states of different types of samples on specific neurons. (a) and (b) are the activation states in the case where the variance loss function is not used. (c) and (d) are the activation states in the case of using the variance loss function.

Two crucial insights arise from Figure 6: (1) Key neurons particularly respond to specific classes while displaying negligible activation for unrelated classes. (2) The application of variance loss fosters consistency in activation intensity for samples within the same class. These findings substantiate the proposed method's robustness by indicating that activation patterns align closely with theoretical expectations, reinforcing the validity of the variance loss mechanism.

According to the 3σ rule of normal distribution, each ID sample has a 99.74% probability of falling within the $\mu \pm 3\sigma$ range of its corresponding key neuron activation value. As such, ID samples can achieve a score rate exceeding 99.74% under their specific discrimination criteria. Setting the OOD detection threshold at 0.9974 represents a theoretically reasonable value. Specifically, when classifying a sample into the j^{th} subclass within the i^{th} broad class, the $\mu \pm 3\sigma$ statistical test requires the sample's confidence score rate to exceed 99.74% to be recognized as an ID sample. However, experimental observations reveal that OOD samples typically yield extremely low scores under the discrimination criteria of known classes, setting excessively high thresholds unnecessary. Overly stringent thresholds may lead to misclassification of ID samples as OOD instances.

As shown in Figure 7, detection thresholds above 0.6 yield superior performance, ensuring both high classification accuracy and robust OOD detection. Only a small fraction of ID samples is misclassified as OOD, reflecting the method's robustness. Moreover, the flexibility of threshold settings allows practical adjustments for specific application needs. For scenarios emphasizing minimal false alarms, lower thresholds suffice, while stringent thresholds cater to environments prioritizing high sensitivity and zero tolerance for missed OOD detections.

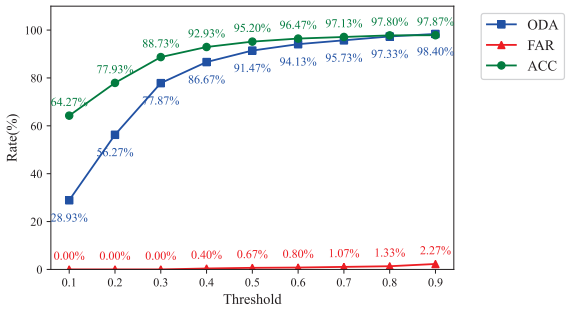

Figure 7. The performance of the model under different detection thresholds.

The method's efficacy was benchmarked against BNN, DE, EDL and UU-DLF models using metrics such as Accuracy (ACC), Out-of-Distribution Detection Accuracy (ODA), False Alarm Rate (FAR), and confusion matrix analysis.

The experimental configurations NSAD-0.8 and NSAD-0.9 represent implementations of the NSAD model with detection thresholds set at 0.8 and 0.9, respectively. As demonstrated in Table 1, the proposed NSAD methodology consistently outperforms the baseline models (BNN, DE, EDL and UU-DLF) across all evaluation metrics: ACC, ODA, and FAR. A critical observation emerges from the comparative analysis presented in the final two rows of Table 1: while elevated detection thresholds enhance OOD sample detection capabilities, they simultaneously increase the likelihood of incorrectly categorizing ID samples as OOD instances. Consequently, the practical implementation of NSAD necessitates careful threshold calibration based on application-specific requirements to optimize the balance between detection sensitivity and false alarm frequency.

Table 1. Performance evaluation across different methods

	ACC	ODA	FAR
BNN	85.53%	83.07%	12.0%
DE	89.22%	96.00%	12.13%
EDL	91.40%	93.60%	10.80%
UU-DLF	86.80%	94.67%	21.07%
NSAD-0.8	**97.80%**	**97.33%**	**1.33%**
NSAD-0.9	**97.87%**	**98.40%**	**2.27%**

Analysis of the experimental results depicted in Figure 8 reveals that under standard classification conditions (without OOD detection constraints), all evaluated models achieve exceptional classification accuracy approaching theoretical limits. However, the benchmark models (BNN, DE, EDL and UU-DLF) demonstrate a notably higher propensity for erroneously classifying In-Distribution samples as Out-of-Distribution cases. In contrast, the proposed NSAD framework exhibits superior discriminative capabilities, successfully identifying approximately 100% of OOD samples while maintaining remarkably low false positive rates in ID sample classification.

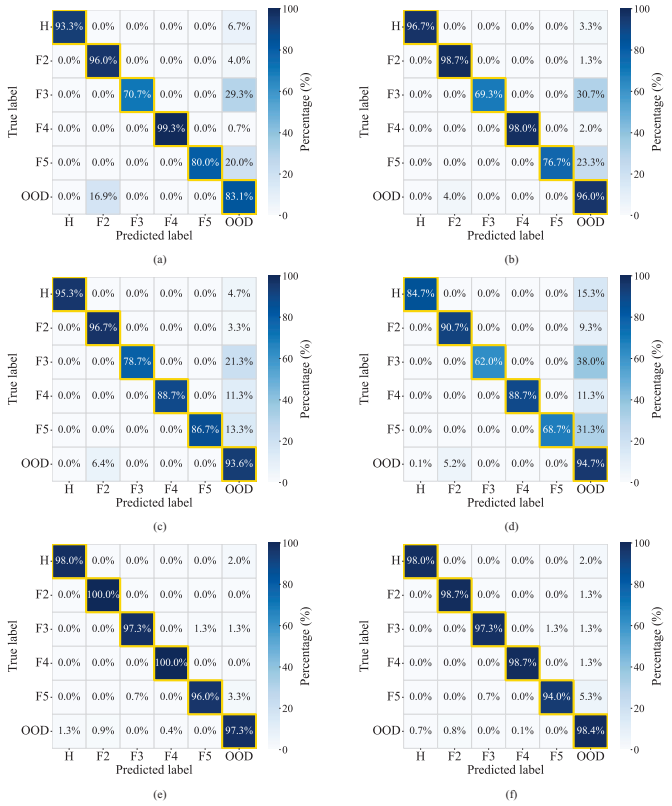

Figure 8. Confusion matrix. (a) BNN; (b) DE; (c) EDL; (d) UU-DLF; (e) NSAD-0.8; (f) NSAD-0.9.

V. Conclusion

The proposed NSAD framework represents a significant advancement in wind turbine gearbox fault diagnosis, effectively addressing the reliability limitations of conventional black-box models through innovative neuron state analysis and multi-component optimization. By successfully combining activation pattern analysis with sophisticated loss function design, the framework achieves enhanced diagnostic accuracy while providing interpretable confidence metrics. This approach not only improves the robustness of fault detection but also establishes a new paradigm for reliable diagnostic systems in wind power applications, potentially reducing maintenance costs and improving operational safety.

References

[1] M. Rao, "Development of deep learning-based methods for rotating machinery fault diagnosis under varying speed conditions," 2023.

[2] A. K. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical systems and signal processing*, vol. 20, no. 7, pp. 1483–1510, 2006.

[3] K. L. Tsui, N. Chen, Q. Zhou, Y. Hai, and W. Wang, "Prognostics and health management: A review on data driven approaches," *Mathematical Problems in Engineering*, vol. 2015, no. 1, p. 793161, 2015.

[4] H. Gu, W. Y. Liu, Q. W. Gao, and Y. Zhang, "A review on wind turbines gearbox fault diagnosis methods," *J. vibroeng.*, vol. 23, no. 1, pp. 26–43, Feb. 2021.

[5] P. Chen, Y. Wu, C. Xu, C.-G. Huang, M. Zhang, and J. Yuan, "Interference Suppression of Nonstationary Signals for Bearing Diagnosis Under Transient Noise Measurements," *IEEE Transactions on Reliability*, 2025.

[6] P. Chen *et al.*, "Adaptive signal regime for identifying transient shifts: A novel approach toward fault diagnosis in wind turbine systems," *Ocean engineering*, vol. 325, p. 120798, 2025.

[7] Z. Feng, S. Qin, and M. Liang, "Time–frequency analysis based on Vold-Kalman filter and higher order energy separation for fault diagnosis of wind turbine planetary gearbox under nonstationary conditions," *Renewable Energy*, vol. 85, pp. 45–56, 2016.

[8] P. Chen, Y. Wu, C. Xu, Y. Jin, and C. Zhou, "Markov modeling of signal condition transitions for bearing diagnostics under external interference conditions," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[9] G. Jiang, H. He, J. Yan, and P. Xie, "Multiscale convolutional neural networks for fault diagnosis of wind turbine gearbox," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3196–3207, 2018.

[10] P. Chen *et al.*, "Metric-guided graph contrastive learning: An unsupervised approach for few-shot gearbox fault diagnosis," *Measurement Science and Technology*, 2025.

[11] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," Oct. 04, 2016, *arXiv*: arXiv:1506.02142.

[12] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?," Oct. 05, 2017, *arXiv*: arXiv:1703.04977.

[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," Nov. 04, 2017, *arXiv*: arXiv:1612.01474.

[14] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential Deep Learning to Quantify Classification Uncertainty," Oct. 31, 2018, *arXiv*: arXiv:1806.01768.

[15] J. Ren, J. Wen, Z. Zhao, R. Yan, X. Chen, and A. K. Nandi, "Uncertainty-Aware Deep Learning: A Promising Tool for Trustworthy Fault Diagnosis," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 6, pp. 1317–1330, Jun. 2024.

[16] X. Zhang, C. Wang, W. Zhou, J. Xu, and T. Han, "Trustworthy Diagnostics With Out-of-Distribution Detection: A Novel Max-Consistency and Min-Similarity Guided Deep Ensembles for Uncertainty Estimation," *IEEE Internet Things J.*, vol. 11, no. 13, pp. 23055–23067, Jul. 2024.

[17] D. Wei, M. Zuo, and Z. Tian, "Open-Set Fault Diagnosis for Industrial Rotating Machines Based on Trustworthy Deep Learning," *Trans. Ind. Cyb-Phy. Sys.*, vol. 3, pp. 181–189, 2025.

[18] D. O. Hebb, The organization of behavior: A neuropsychological theory. Psychology press, 2005.

[19] T. H. Brown, E. W. Kairiss, and C. L. Keenan, "Hebbian synapses: biophysical mechanisms and algorithms," *Annual review of neuroscience*, vol. 13, no. 1, pp. 475–511, 1990.

[20] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in neural information processing systems*, vol. 28, 2015.

[21] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2736–2744.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PmLR, 2020, pp. 1597–1607.

[23] P. Chen *et al.*, "Progressive contrastive representation learning for defect diagnosis in aluminum disk substrates with a bio-inspired vision sensor," *Expert Systems with Applications*, p. 128305, 2025.

[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, University of California press, 1967, pp. 281–298.

[25] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955.

[26] Y. Chen, M. Rao, X. Chen, X. Liang, L. Liu, and M. Zuo, "Experiment design and data collection on the fixed-axis gearbox under time-varying operation conditions technical report," *Reliability Research Lab, Department of Mechanical Engineering, University of Alberta*, 2018.