Diagnostics for Mechanical Systems with Unknown Fault Modes: A Novel Open Set Recognition Approach

Jiaxuan Song, Juseong Lee, Claudia Fecarotti, and Geert-Jan van Houtum

Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

j.song1@tue.nl

j.lee@tue.nl

c.fecarotti@tue.nl

G.J.v.Houtum@tue.nl

ABSTRACT

A common challenge in condition-based maintenance is that not all fault modes of the system are known from historical data, particularly in systems with evolving operating conditions, or are newly developed. Conventional data-driven diagnostic methods typically rely on a closed-set assumption, where all possible fault modes are represented during training. As a result, previously unseen fault modes are often incorrectly assigned to known ones with high confidence, potentially leading to ineffective or even risky maintenance decisions. To address this limitation, this paper proposes an open-set diagnostic approach that integrates supervised contrastive learning with a simplified Hopfield energy score. An encoder is trained using a supervised contrastive loss function to obtain well-separated embeddings of known system states. During inference, the alignment between a test observation and the learned state prototypes is quantified using the simplified Hopfield energy score. Observations with low similarity to known states are identified as unknown through thresholding. Experimental results on a benchmark dataset demonstrate that the proposed method effectively distinguishes unknown states while maintaining an accurate classification of known states, achieving competitive performance compared to established baselines. By explicitly identifying unknown states, the proposed approach enables more reliable and risk-aware maintenance decisions, particularly in safety-critical applications.

1. INTRODUCTION

Data-driven diagnostics is a key enabling technology to ensure the availability, productivity and reliability of complex systems while minimizing operational costs (Fink et al., 2020; Hu, Miao, Si, Pan, & Zio, 2022). Advances in IoT and sensor technologies allow system owners to continuously monitor machine conditions through dense sensor networks, often complemented by periodic inspections. These data streams support condition assessment, failure prediction, and eventually the development of optimal maintenance strategies.

Due to the complexity of modern mechanical systems, degradation can occur through multiple mechanisms, resulting in various fault modes. However, in real-world diagnostic scenarios, training data rarely cover all possible fault modes. This situation commonly arises in several cases: (i) newly developed systems that have not yet operated long enough to accumulate sufficient fault data; (ii) legacy systems operating under previously unseen or significantly changed conditions; and (iii) critical systems of safety, for which failure data are inherently scarce due to stringent safety requirements (Zonta et al., 2020). Such circumstances highlight an important challenge in prognostics and health management: how to achieve reliable diagnostic performance when only limited and incomplete fault data are available.

In recent years, intelligent diagnostic methods have already been adopted to address various data-related challenges, including data sparsity, class imbalance, and noise contamination (J. Li et al., 2024; Y. Zhang, Ding, Li, Ren, & Feng, 2024; Lai, Baraldi, & Zio, 2024). However, most existing approaches improve diagnostic performance under the assumption of a fixed and predefined set of fault modes. These methods either generate samples that resemble a predefined set of faults (Rombach, Michau, & Fink, 2023) or transfer knowledge from related datasets to improve recognition within the same fixed label space (X. Li, Hu, Li, & Zheng, 2020; X. Li, Hu, Zheng, Li, & Ma, 2021). This

closed-set assumption inherently restricts the model's capacity to accommodate previously unseen fault modes that were not observed during prior operation, referred to as unknown fault modes. When such unknown fault modes emerge in practice, diagnostic models are typically forced to assign them to one of the predefined known classes. This overconfident mis-classification undermines diagnostic reliability, compromises system trustworthiness, and may ultimately lead to inappropriate maintenance decisions.

In response to the limitations imposed by the closed-set assumption, several recent studies have explicitly addressed the diagnosis of unknown fault modes (Xie, Han, Pei, & Xie, 2023; Yoon, Kim, & Kim, 2022). In this context, a diagnostic model is expected not only to accurately classify observations belonging to known fault modes but also to identify observations that correspond to unknown fault modes. Techniques designed for this purpose are collectively referred to as Open Set Recognition (OSR) (Yang, Zhou, Li, & Liu, 2024). Existing approaches to OSR can generally be categorized into two groups: *logit-based* methods and *feature-based* methods (Gavarini, Stucchi, Ruospo, Boracchi, & Sanchez, 2022). Logit-based methods determine whether a sample belongs to a known class based on output confidence scores or logits, whereas feature-based methods assess the similarity between the extracted feature embeddings of the new observation and the prototypes of known fault modes. Between these two categories, feature-based methods are often preferred for diagnostic applications, as they tend to produce more compact fault-mode clusters and demonstrate greater robustness to variations in operating conditions (Lee, Lee, Lee, & Shin, 2018; Sun, Ming, Zhu, & Li, 2022).

In feature-based OSR methods, extracting informative embeddings from raw monitoring data plays a crucial role in determining diagnostic performance (Peng et al., 2022). High-quality embeddings that are compact within the same fault mode and well separated between different fault modes facilitate reliable fault detection and improved generalization. Motivated by this observation, supervised contrastive learning (Khosla et al., 2020) can be incorporated to impose explicit structural or semantic constraints associated with known fault modes. This approach enables the learning of structured fault representations that enhance discrimination among known fault modes and provide a robust foundation for open-set fault detection.

In this paper, we propose a novel OSR approach for fault diagnostics in mechanical systems with unknown fault modes. Following the general framework of feature-based OSR, our method uses supervised contrastive learning to extract fault-aware prototypes of known fault modes by maximizing intra-class similarity and minimizing inter-class similarity. A simplified Hopfield energy–based score (J. Zhang et al., 2022) is then used to quantify the similarity between a new observation and the learned prototypes of known fault modes. This combination enables reliable diagnostics by raising alerts when unknown fault modes are encountered, while maintaining accurate diagnostic performance for known fault modes.

The main contributions of this paper are summarized as follows:

1. We introduce an open-set fault diagnostic framework that addresses the limitations imposed by the "closed-set" assumption commonly adopted in existing diagnostic models.
2. We propose a novel approach that integrates supervised contrastive learning with the simplified Hopfield energy score to enable reliable identification of unknown fault modes.
3. We demonstrate the effectiveness of the proposed approach on a benchmark dataset and discuss its relevance in the context of maintenance applications.

2. PROBLEM DESCRIPTION

In this study, we consider a system whose health condition evolves through a set of discrete states: healthy, faulty, and failed. Within the faulty state, the system can exhibit M distinct fault modes that differ in terms of location, severity, or underlying root causes. Let $s \in \mathcal{S}$ denote the state of the system, where the state space is defined as

$$\mathcal{S} = \{H, FM_1, \ldots, FM_M\}, \tag{1}$$

where H represents the healthy state, FM_m denotes the faulty state associated with the fault mode $m = 1, 2, \ldots, M$. Although the system can eventually reach a failed state, the failed state is not explicitly included in $\mathcal{S}$, as it is assumed to be self-revealing and does not require diagnostic inference.

We assume that the system initially operates in the healthy state H, transitions to a faulty state FM_m associated with the fault mode $m = 1, 2, \ldots, M$, and remains in this faulty state until it eventually reaches the failed state. In other words, the fault modes are assumed to be mutually exclusive, and transitions between different faulty states are not allowed. Although the failed condition can be directly observed (e.g., through system shutdown), both the healthy and faulty states are not directly observable and must be inferred from condition monitoring data.

During operation, the system is monitored by condition monitoring observations. Let o denote the observation obtained from the monitoring system. The relationship between the hidden system state and the observation can be described as a mapping of the system state s to the observed condition monitoring data o, subject to observation noise. Since mapping is generally unknown and is not explicitly

available for maintenance decision-making, diagnostics are performed by inferring the hidden state of the system from the observed condition monitoring data o. In practice, the diagnostic model can be implemented using a neural network parameterized by θ_f,

$$\hat{s} = f(o; \theta_f) \tag{2}$$

where $f(\cdot)$ denotes the diagnostic model that estimates the underlying state of the system based on observation o. If the inferred state $\hat{s}$ corresponds to one of the faulty states FM_m, an appropriate maintenance action is initiated to restore the system to the healthy state. Furthermore, since we assume that the system stops operating immediately when the failed state is reached, a replacement action is initiated without delay.

Suppose that historical condition monitoring data with true system state are known. A training sample is represented as a pair (o_i, s_i), where o_i denotes the observation and s_i denotes the corresponding state of the system. The historical dataset is defined as $\mathcal{D}_{\text{known}} = \{(o_i, s_i)\}_{i=1}^{N}$, where N is the total number of samples in the dataset. Let $\mathcal{S}_{\text{known}}$ denote the set of states that appear in the historical dataset, $\mathcal{S}_{\text{known}} = \{s | (o, s) \in \mathcal{D}_{\text{known}}\}$. We assume that the historical dataset contains only a subset of all possible states of the system. Therefore, $\mathcal{S}_{\text{known}} \subsetneq \mathcal{S}$. The states included in $\mathcal{D}_{\text{known}}$ are referred to as known states, whereas the states that do not appear in the dataset are referred to as unknown fault states. We assume that healthy state is always known. The set of unknown states is defined as $\mathcal{S}_{\text{unknown}} = \mathcal{S} \setminus \mathcal{S}_{\text{known}}$.

Given the training dataset $\mathcal{D}_{\text{known}}$, the objective of this study is to develop a diagnostic model that can infer the state of a newly observed sample based on its observation o. In particular, the model should be able to determine whether the sample belongs to one of the known states in $\mathcal{S}_{\text{known}}$ or corresponds to an unknown state in $\mathcal{S}_{\text{unknown}}$. This problem can be formulated as an open-set fault diagnosis task in which the diagnostic model must accurately classify samples from known states while simultaneously identifying samples originating from previously unseen fault states.

3. METHODOLOGY

To address the open-set fault diagnosis task defined in the previous section, the proposed method consists of two main components. First, monitoring observations are mapped into an embedding space using supervised contrastive learning to learn discriminative representations of system states. Second, a confidence score is computed in the embedding space using simplified Hopfield energy to determine whether a new observation corresponds to a known state or an unknown state.

3.1. Supervised Contrastive Learning

Supervised contrastive learning is employed to learn representative embeddings of observations. To achieve this objective, data augmentation is applied to the historical dataset $\mathcal{D}_{\text{known}}$. Specifically, for each observation o_i, multiple augmented views are generated by introducing random perturbations. These perturbations are designed to preserve the underlying state characteristics while introducing variability into the input data.

Suppose that K augmented views are generated for each observation o_i. The resulting augmented observations are denoted by $\tilde{\mathcal{O}}_i = \{\tilde{o}_i^{(1)}, \tilde{o}_i^{(2)}, \ldots, \tilde{o}_i^{(K)}\}$. Let $\tilde{s}_i^{(k)}$ denote the state of the system associated with the augmented observation $\tilde{o}_i^{(k)}$. Since augmentation does not alter the underlying state of the system, the state remains unchanged, i.e., $\tilde{s}_i^{(k)} = s_i, \ \forall\, k = 1, 2, \ldots, K$.

Let $J \equiv \{1, 2, \ldots, KN\}$ denote the index set of all samples of the augmented observations, where K augmented views are generated for each original observation $i = 1, 2, \ldots, N$. For a given sample with index $j \in J$, we define $A(j) = J \setminus \{j\}$ as the set of indexes of all samples excluding sample j. Within the supervised contrastive learning framework, the positive index set $P(j)$ associated with the sample j consists of all samples that share the same state of the system, that is, $P(j) = \{p \in A(j) : \tilde{s}_p = \tilde{s}_j\}$. All remaining samples in $A(j)$ that do not belong to $P(j)$ are defined as negative samples.

Based on these definitions, a neural network encoder $h(\cdot; \theta_h)$ is trained to map the monitoring observations to an embedding space. To learn state-discriminative embeddings, the encoder is trained using supervised contrastive loss:

$$\mathcal{L}^{\text{sup}} = \sum_{j \in J} \frac{-1}{|P(j)|} \sum_{p \in P(j)} \log \frac{\exp(z_j \cdot z_p / \tau)}{\sum_{a \in A(j)} \exp(z_j \cdot z_a / \tau)} \tag{3}$$

where z_j, z_p, z_a denotes the embedding extracted from the augmented observations, $|P(j)|$ represents the number of positive samples associated with the index $j \in J$, and τ is a temperature hyperparameter that controls the concentration of the embedding distribution by scaling similarity scores: larger values of τ lead to tighter clusters, while smaller values produce more dispersed embeddings.

After training, the learned encoder $h(\cdot; \theta_h)$ is applied to the original dataset $\mathcal{D}_{\text{known}}$ to obtain embeddings of monitoring observations, given by

$$z_i = h(o_i; \theta_h). \tag{4}$$

Using these embeddings, a representative *prototype* is

constructed for each known state. Let $c \in \mathcal{S}_{\text{known}}$ denote a known state. The prototype of state c is defined as the mean embedding of all samples belonging to that state,

$$p_c = \frac{1}{n_c} \sum_{i=1}^{N} z_i \cdot \mathbb{I}(s_i = c) \tag{5}$$

where $n_c = \sum_{i=1}^{N} \mathbb{I}(s_i = c)$ denotes the number of samples associated with state c, and $\mathbb{I}(\cdot)$ is the indicator function.

3.2. Simplified Hopfield Energy

During deployment, a new monitoring observation o_ξ is provided to the trained diagnostic model. Using the encoder learned through supervised contrastive learning, the observation is mapped into the embedding space as

$$z_\xi = h(o_\xi; \theta_h) \tag{6}$$

where $h(\cdot; \theta_h)$ denotes the trained encoder and z_ξ is the embedding of the test observation.

A classifier can be finetuned based on the encoder to predict the most likely system state among the known states. Let q_c denote the predicted probability that the observation o_ξ belongs to state $c \in \mathcal{S}_{\text{known}}$. The predicted state is therefore obtained as

$$\hat{s}_\xi = \underset{c \in \mathcal{S}_{\text{known}}}{\operatorname{argmax}}\, q_c. \tag{7}$$

However, if the true state s_ξ corresponds to an unknown state in $\mathcal{S}_{\text{unknown}}$, the predicted label $\hat{s}_\xi$ becomes unreliable because the classifier is trained only on the states contained in the historical dataset $\mathcal{D}_{\text{known}}$. Therefore, an additional scoring function is required to measure the confidence in assigning the observation to the known state space.

To address this issue, we adopt the Simplified Hopfield Energy (SHE) score (Ramsauer et al., 2020; J. Zhang et al., 2022). The Hopfield network is a classical associative memory model that stores and retrieves continuous patterns by minimizing a predefined energy function. Inspired by this principle, the SHE score quantifies the discrepancy between a test embedding and stored class representations.

In this work, the stored representations correspond to the prototypes learned for each known state $c \in \mathcal{S}_{\text{known}}$, denoted by p_c. Given the predicted state $\hat{s}_\xi$, the simplified Hopfield energy score is defined as

$$\text{SHE}(z_\xi) = z_\xi \cdot p_{\hat{s}_\xi}, \tag{8}$$

where $p_{\hat{s}_\xi}$ denotes the prototype corresponding to the predicted state $\hat{s}_\xi$.

The inner product $z_\xi \cdot p_{\hat{s}_\xi}$ measures the similarity between the test embedding and the prototype of the predicted state. A higher energy score indicates a stronger similarity to the predicted known state, whereas a lower energy score suggests weaker alignment and potential deviation from all known system states.

Based on this interpretation, a threshold δ is introduced to distinguish between known and unknown states. The final diagnostic decision is defined as

$$\hat{s}_\xi = \begin{cases} \hat{s}_\xi, & \text{if SHE}(z_\xi) > \delta, \\ \text{unknown}, & \text{otherwise.} \end{cases} \tag{9}$$

Observations whose energy score falls to meet the threshold are regarded as unknown samples, indicating the potential presence of previously unseen system states.

4. EXPERIMENTAL STUDY

In this section, we evaluate the performance of the proposed method and compare it with several representative baseline approaches on a widely adopted benchmark dataset.

4.1. Experimental Setup

4.1.1. Dataset

We use the Case Western Reserve University (CWRU) bearing dataset (Smith & Randall, 2015), which is a widely adopted benchmark to evaluate machine fault diagnosis methods. The experiments were conducted on a 2 hp Reliance Electric motor, where vibration acceleration signals were collected at both the drive end and the fan end of the motor. Bearing faults were artificially introduced using electro-discharge machining to simulate different fault conditions.

In this study, we utilized vibration signals collected at the drive end with a sampling rate of 12 kHz, under a load of 2 hp and an approximate speed of 1750 rpm. Among all available recordings, 9 operating states were selected to ensure consistent operating conditions and bearing configurations while maintaining a representative diversity of fault modes. These states consist of 1 healthy condition and 8 distinct fault modes. Detailed descriptions of the selected operating states are provided in Table 1.

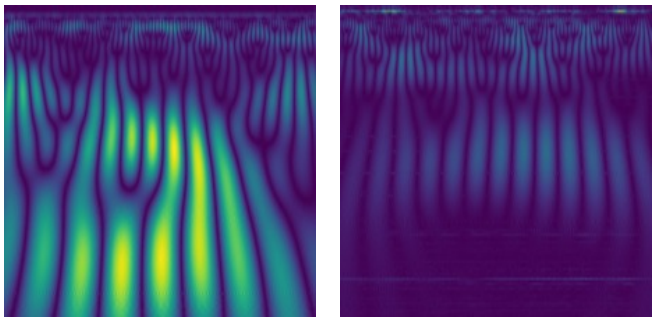
4.1.2. Pipeline

The selected vibration signals are pre-processed using a sliding window with 0.042 seconds to generate signal segments. Each segment covers more than one round of bearing rotation. Subsequently, the segments are transformed into the time–frequency domain using the

Table 1. Description of the states in CWRU dataset

State	Description of the state
H	Healthy state
FM_1	Rolling element fault with diameters of 7 mils
FM_2	Inner raceway fault with diameters of 14 mils
FM_3	Outer raceway fault at 6 o'clock with diameters of 14 mils
FM_4	Rolling element fault with diameters of 21 mils
FM_5	Outer raceway fault at 6 o'clock with diameters of 7 mils
FM_6	Rolling element fault with diameters of 14 mils
FM_7	Inner raceway fault with diameters of 21 mils
FM_8	Outer raceway fault at 3 o'clock with diameters of 21 mils

continuous wavelet transform with the complex Morlet wavelet (Zhu, Chen, & Peng, 2018; Łuczak, 2024). The resulting scalograms are converted to a logarithmic scale to enhance the visibility of features across different frequency bands. Fig. 1 shows two examples of the scalograms of states H and FM_8. The generated images are then resized to $224 \times 224 \times 3$ to match the input requirements of the network. Each processed image is treated as an individual sample for model training and evaluation.



(a) (b)

Figure 1. (a) Scalogram of healthy state; (b) Scalogram of state FM_8.

We consider an open-set diagnosis scenario in which only a subset of system states is available during training, while previously unseen states may appear during testing. To simulate this setting, four parallel tests are designed to evaluate the performance of the proposed diagnostic model. Specifically, FM_5, FM_6, FM_7, and FM_8 are selected as unknown system states, while the remaining system states are regarded as known states. In each test, one of these states is designated as unknown and is excluded from training but introduced during testing. The known system states are used for training, remaining the same across all four tests.

For model implementation, a ResNet-18 architecture (He, Zhang, Ren, & Sun, 2016) is adopted as the encoder to extract embeddings from monitoring observations. The network is first trained using supervised contrastive learning to learn discriminative embeddings for the known system states. Subsequently, a classification layer is appended, and the entire model is fine-tuned using the cross-entropy loss on samples from the known states. After training, feature embeddings are extracted from the penultimate layer to construct state-wise prototypes p_c for each known state $c \in \mathcal{S}_{\text{known}}$. During inference, test samples are projected into the learned embedding space, and the simplified Hopfield energy score is computed with respect to the most likely prototype.

4.1.3. Benchmarks

To assess the effectiveness of the proposed approach in open-set fault diagnosis, a comparative experiment is conducted against four representative open-set recognition methods. The selected baselines include two logit-based approaches, Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2016; Lin, Ren, Zhao, Zhang, & Chen, 2025; Wu, Gao, & Zhang, 2026) and ODIN (Liang, Li, & Srikant, 2017; Kemmerzell, Schreiner, Khalid, Schalk, & Bordoli, 2025), as well as two feature-based approaches, Mahalanobis distance (Maha) (Lee et al., 2018; Fang, Easwaran, Genest, & Suganthan, 2025) and Simplified Hopfield Energy (SHE) standalone without supervised contrastive learning (J. Zhang et al., 2022; Zohrabi et al., 2026).

Despite their methodological differences, these benchmark approaches follow a common paradigm. Specifically, feature embeddings z are first extracted from the input data, after which a scoring function $\gamma(z)$ is defined to quantify either classification confidence (for logit-based methods) or feature similarity (for feature-based methods). Without loss of generality, the score is defined in a way that higher values indicate a higher likelihood that an observation o belongs to an unknown state. Given a decision threshold τ, a sample is classified as unknown if $\gamma(z) > \tau$ and as known otherwise.

Based on this unified framework, three widely used metrics are adopted to evaluate open-set diagnostic performance: the area under the receiver operating characteristic curve (AUROC), the false positive rate at 95% true positive rate (FPR@95%TPR), and the true positive rate at 5% false positive rate (TPR@5%FPR). The true positive rate $TPR(\tau)$ and false positive rate $FPR(\tau)$ are computed by varying the decision threshold τ. The receiver operating characteristic curve plots $TPR(\tau)$ against $FPR(\tau)$ across all threshold values, while AUROC provides a threshold-independent measure of the overall separability between known and unknown samples. The metric FPR@95%TPR reports the false positive rate when the detection rate of unknown samples reaches 95%, reflecting the robustness of the method under high detection requirements. In contrast, TPR@5%FPR measures the detection rate of unknown samples when the false alarm rate on known samples is constrained to 5%.

4.2. Experimental Results

4.2.1. Performance of the proposed method

The resulting energy score distributions for known and unknown system states are shown in Fig. 2. The energy score distribution of the known states is illustrated in blue, while that of the unknown states is shown in red. The degree of separation between the two distributions is quantified using the Kolmogorov–Smirnov (KS) statistic, which measures the maximum distance between their cumulative distribution functions. A larger KS statistic indicates less overlap between the distributions and, consequently, stronger separability between known and unknown system states. Across all four test cases, the proposed method yields only limited overlap between the energy score distributions of known and unknown states. In particular, the KS statistics for Test 1 and Test 3 in Fig. 2 are 0.9868 and 0.9748, respectively, indicating a high degree of separation and strong discriminative capability of the proposed approach.

These results indicate that the simplified Hopfield energy score provides an effective criterion for distinguishing previously unseen states from known system states. The clear separation between the corresponding energy score distributions enables reliable threshold-based decision making, as defined in Eq. (9). Specifically, the decision threshold δ is set to the 5-th percentile of the SHE scores for the known states in the training set. Samples with SHE scores below this threshold are classified as unknown states, while those exceeding the threshold are classified as known states. Using this thresholding strategy, the resulting confusion matrices for the test set are presented in Fig. 3.

In open-set diagnostics, misdiagnosing errors can be categorized into three types: (i) misdiagnosing one known system state as another known system state; (ii) misdiagnosing unknown system states as known system states; and (iii) misdiagnosing known system states as unknown system states. As shown in Fig. 3, the proposed method demonstrates strong diagnostic reliability among known system states, with zero error of the first type. Furthermore, in Tests 1–3, errors of the second type occur only four times in total, with a slight increase observed in Test 4. By contrast, the third type of error constitutes the majority of misdiagnosing, indicating that the model occasionally assigns samples from known system states to unknown states. This behavior suggests that the method effectively mitigates the "overconfidence" issue commonly observed in closed-set diagnostic models, where unknown system states are incorrectly assigned to known states.

From a maintenance perspective, these error types correspond to two distinct operational consequences: under-prepared maintenance and over-prepared maintenance. Errors of the first and second types can lead to

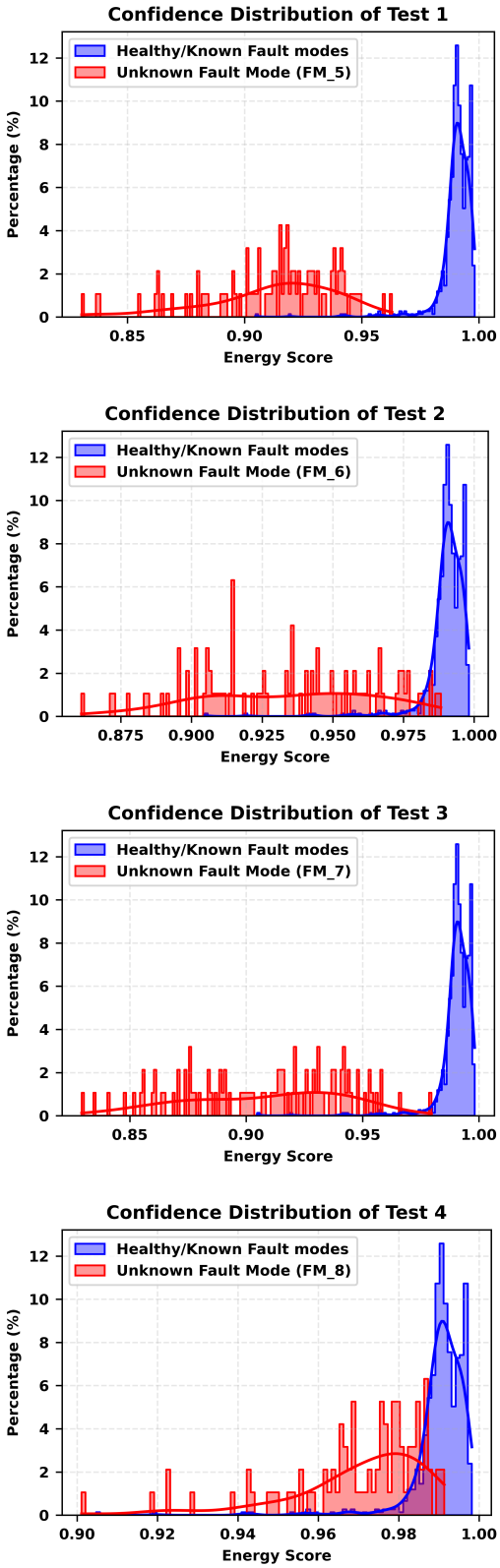


Figure 2. The confidence distribution of known and unknown system states.

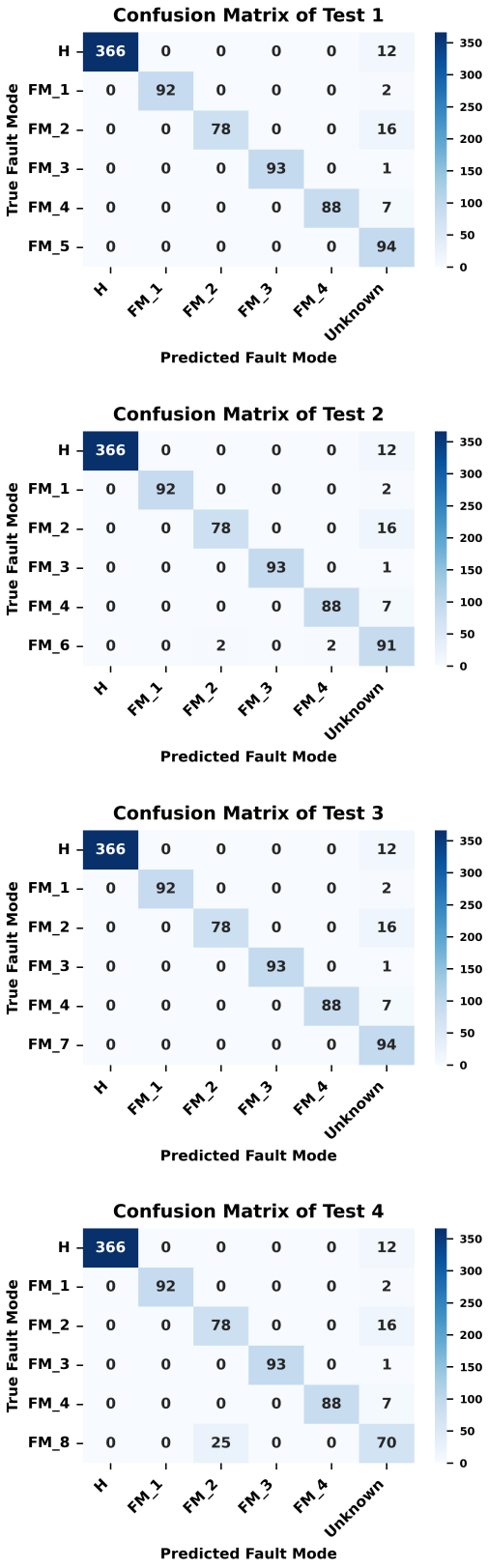


Figure 3. The confusion matrix of known/unknown system states.

under-prepared maintenance, as the true state of the system is either not recognized or expected, resulting in inadequate or incorrect preparation of required resources, such as spare parts, tools, and personnel. In contrast, errors of the third type lead to over-prepared maintenance, where additional resources may be allocated unnecessarily. In this interpretation, the observed error pattern indicates that the proposed method adopts a conservative diagnostic strategy in the presence of uncertainty. Such behavior is often desirable in predictive maintenance, particularly for newly developed or safety-critical systems, where the cost of incorrect maintenance actions can be substantial. In these scenarios, stakeholders typically prefer conservative decisions that trigger further inspection rather than relying on potentially overconfident diagnoses. Consequently, prioritizing the identification of unknown states and encouraging additional investigation can be considered a safer and more appropriate strategy for open-set fault diagnosis in practical applications.

4.2.2. Comparison with benchmark methods

Comparison experiments are conducted with four baseline methods and are evaluated using the three predefined metrics. The results are summarized in Table 2, where the best and second-best performances are highlighted.

Table 2. Comparison results of the open-set fault diagnostics

Evaluation Index	Method	Unknown Fault Modes				Ave	Std
		FM 3	FM 7	FM 9	FM 12		
AUROC (↑)	MSP	0.9992	0.8987	0.9883	0.9133	0.9499	0.0443
	ODIN	0.7968	0.4651	0.7652	0.4377	0.6162	0.1655
	Maha	1.0000	0.9983	1.0000	1.0000	0.9996*	0.0007*
	SHE	0.9999	0.9260	0.9938	0.9702	0.9725	0.0290
	SCL+SHE	0.9979	0.9875	0.9974	0.9436	0.9943**	0.0223**
FPR@ 95%TPR(↓)	MSP	0.0013	0.6887	0.0450	0.3775	0.2781	0.2781
	ODIN	0.5232	0.7868	0.5205	0.6305	0.6153	0.1085
	Maha	0.0000	0.0066	0.0000	0.0000	0.0017*	0.0029*
	SHE	0.0013	0.4159	0.0212	0.0781	0.1291	0.1680
	SCL+SHE	0.0066	0.0437	0.0079	0.1974	0.0639**	0.0785**
TPR@ 5%FPR(↑)	MSP	1.0000	0.6842	0.9681	0.5053	0.7894	0.2050
	ODIN	0.5745	0.1158	0.0340	0.0000	0.1811	0.2310
	Maha	1.0000	0.9895	1.0000	1.0000	0.9974*	0.0045*
	SHE	1.0000	0.6947	1.0000	0.8947	0.8974	0.1246
	SCL+SHE	1.0000	0.9579	1.0000	0.7368	0.9860**	0.1093**

Note: * and ** denote the best and second-best performance, respectively.

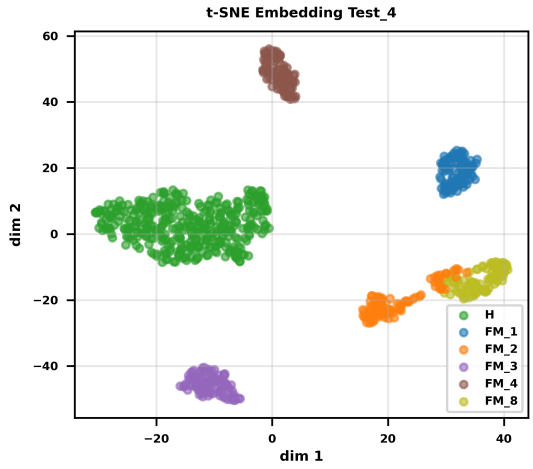
Compared with the logit-based approaches *MSP* and *ODIN*, the proposed method demonstrates clear advantages in both average detection accuracy and performance stability. In the

best-performing case, the feature-based approach improves the average TPR@5%FPR by 81.63% while reducing the average FPR@95%TPR by 27.65%. This improvement can be attributed to the inherent limitation of logit-based methods, which rely on the softmax operator to distribute the probability mass among the known state of the system. Although these methods attempt to extend closed-set classifiers to open-set diagnostics through probability-based heuristics, where unknown samples are first forced to be assigned to one of the known states before additional processing is applied, the model exhibits overconfidence when encountering previously unseen states.

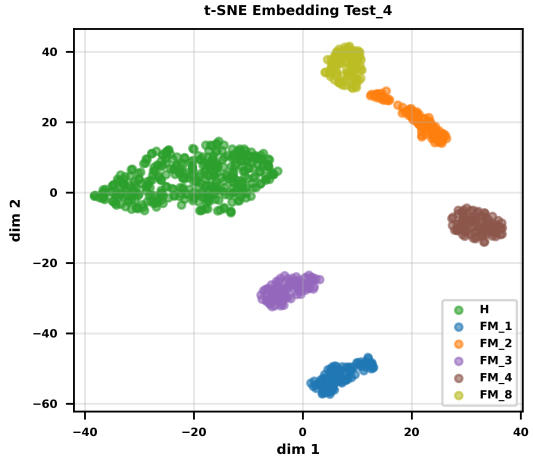
In general, feature-based methods, namely *Maha*, *SHE*, and *SHE+SCL*, exhibit strong performance, with only minor differences observed between them. *Maha* (Average AUROC 0.9996) performs slightly better than *SHE* (Average AUROC 0.9725) . One possible explanation is that *Maha* incorporates additional statistical modeling by representing each system state through both a mean embedding and a shared covariance matrix, thereby capturing not only the central tendency, but also the distributional characteristics of the feature space. This additional modeling capability may enhance its discriminative capability in open-set diagnostic tasks. Nevertheless, when supervised contrastive learning is integrated into *SHE*, the resulting *SHE + SCL* (Average AUROC 0.9943) achieves a performance comparable to that of *Maha*, demonstrating the contribution of supervised contrastive learning to open-set discrimination.

The advantage of incorporating supervised contrastive learning becomes more evident when examining the structure of the learned embedding space. As defined by the loss function in Eq. 3, supervised contrastive learning encourages samples belonging to the same state of the system to form compact clusters while simultaneously increasing the separation between samples from different states of the system. As a result, the embedding space becomes more structured and discriminative. Under such a representation, samples associated with unknown system states are more likely to be located in regions between the clusters of known system states, making them easier to identify using the simplified Hopfield energy score. An illustrative example that demonstrates the effect of supervised contrastive learning is presented in Fig. 4. In this example, FM_8 is the unknown state, and the embedding spaces learned with and without supervised contrastive training are compared.

As shown in Fig. 4a, when supervised contrastive training is not applied, the embeddings corresponding to the unknown state of the system (dots in yellow) overlap substantially with those of known states (dots in orange), making it difficult to distinguish it from the known state FM_2. In contrast, Fig. 4b shows that after supervised contrastive training, the embeddings of the unknown state of the system form a more clearly separated cluster in the latent space. This improved separability indicates that supervised contrastive learning encourages more discriminative embeddings and effectively pushes unknown system states away from known state clusters. Consequently, the ambiguity between known and unknown system states is reduced, facilitating more reliable identification of previously unseen states. Furthermore, the improved separability of the embedding space may provide additional benefits in terms of representation interpretability, as the distinctions between known and unknown system states become more apparent. This potential should be explored in further works.



(a)



(b)

Figure 4. (a). t-SNE embeddings of system states in Test 4 trained without supervised contrastive learning; (b). t-SNE embeddings of system states in Test 4 trained with supervised contrastive learning.

5. CONCLUSION

This paper proposes an open-set diagnosis approach that combines supervised contrastive learning with a simplified Hopfield energy-based detection mechanism. By learning discriminative embeddings of known system states and introducing an energy-based criterion for the detection of unknown system states, the method enables the reliable identification of previously unseen conditions.

Experimental results on a benchmark dataset demonstrate that the proposed approach effectively separates known and unknown states and achieves competitive performance compared to established baselines. The strong separability between known and unknown states supports more reliable and risk-aware maintenance decision-making. More importantly, in practical maintenance settings, the conservative misdiagnosing enables practitioners to trigger additional inspections or precautionary actions when unfamiliar conditions are detected, thereby reducing the risk of under-prepared operational responses. Consequently, the proposed method improves the robustness of diagnostic systems, particularly in scenarios with limited or incomplete data.

REFERENCES

Fang, X., Easwaran, A., Genest, B., & Suganthan, P. N. (2025). Your data is not perfect: Towards cross-domain out-of-distribution detection in class-imbalanced data. *Expert Systems with Applications*, *267*, 126031.

Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.-J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*, 103678.

Gavarini, G., Stucchi, D., Ruospo, A., Boracchi, G., & Sanchez, E. (2022). Open-set recognition: an inexpensive strategy to increase dnn reliability. In *2022 ieee 28th international symposium on on-line testing and robust system design (iolts)* (pp. 1–7).

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*.

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hu, Y., Miao, X., Si, Y., Pan, E., & Zio, E. (2022). Prognostics and health management: A review from the perspectives of design, development and decision. *Reliability Engineering & System Safety*, *217*, 108063.

Kemmerzell, N., Schreiner, A., Khalid, H., Schalk, M., & Bordoli, L. (2025). Towards a better understanding of evaluating trustworthiness in ai systems. *ACM Computing Surveys*, *57*(9), 1–38.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, *33*, 18661–18673.

Lai, C., Baraldi, P., & Zio, E. (2024). Physics-informed deep autoencoder for fault detection in new-design systems. *Mechanical Systems and Signal Processing*, *215*, 111420.

Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, *31*.

Li, J., Yue, K., Chen, Z., Xia, J., Li, W., & Zhang, X. (2024). An uncertainty-aware continual learning framework for fault diagnosis of rotating machinery with homogeneous-heterogeneous faults. *IEEE Transactions on Automation Science and Engineering*.

Li, X., Hu, Y., Li, M., & Zheng, J. (2020). Fault diagnostics between different type of components: A transfer learning approach. *Applied Soft Computing*, *86*, 105950.

Li, X., Hu, Y., Zheng, J., Li, M., & Ma, W. (2021). Central moment discrepancy based domain adaptation for intelligent bearing fault diagnosis. *Neurocomputing*, *429*, 12–24.

Liang, S., Li, Y., & Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Lin, F., Ren, J., Zhao, Z., Zhang, X., & Chen, X. (2025). An uncertainty-guided contrastive learning method for ood detection in trustworthy fault diagnosis. *Reliability Engineering & System Safety*, 111967.

Łuczak, D. (2024). Machine fault diagnosis through vibration analysis: continuous wavelet transform with complex morlet wavelet and time–frequency rgb image recognition via convolutional neural network. *Electronics*, *13*(2), 452.

Peng, P., Lu, J., Xie, T., Tao, S., Wang, H., & Zhang, H. (2022). Open-set fault diagnosis via supervised contrastive learning with negative out-of-distribution data augmentation. *IEEE Transactions on Industrial Informatics*, *19*(3), 2463–2473.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., ... others (2020). Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Rombach, K., Michau, G., & Fink, O. (2023). Controlled generation of unseen faults for partial and open-partial domain adaptation. *Reliability Engineering & System Safety*, *230*, 108857.

Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the case western reserve university data: A benchmark study. *Mechanical systems and signal processing*, *64*, 100–131.

Sun, Y., Ming, Y., Zhu, X., & Li, Y. (2022). Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning* (pp. 20827–20840).

Wu, C., Gao, F., & Zhang, R. (2026). An out-of-distribution fault detection framework using deep global feature modeling and extended logit fusion for industrial processes. *Process Safety and Environmental Protection*, 108778.

Xie, W., Han, T., Pei, Z., & Xie, M. (2023). A unified out-of-distribution detection framework for trustworthy prognostics and health management in renewable energy systems. *Engineering Applications of Artificial Intelligence*, *125*, 106707.

Yang, J., Zhou, K., Li, Y., & Liu, Z. (2024). Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, *132*(12), 5635–5662.

Yoon, J., Kim, D., & Kim, D. (2022). Sdbosr: Separable decision boundary based open set recognition for manufacturing equipment fault classification. In *2022 ieee international conference on prognostics and health management (icphm)* (pp. 121–126).

Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., … others (2022). Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The eleventh international conference on learning representations.*

Zhang, Y., Ding, J., Li, Y., Ren, Z., & Feng, K. (2024). Multi-modal data cross-domain fusion network for gearbox fault diagnosis under variable operating conditions. *Engineering Applications of Artificial Intelligence*, *133*, 108236.

Zhu, J., Chen, N., & Peng, W. (2018). Estimation of bearing remaining useful life based on multiscale convolutional neural network. *IEEE Transactions on Industrial Electronics*, *66*(4), 3208–3216.

Zohrabi, R., Hasani, H., Baghshah, M., Rohrbach, A., Rohrbach, M., & Rohban, M. H. (2026). Spurious-aware prototype refinement for reliable out-of-distribution detection. *Advances in Neural Information Processing Systems*, *38*, 44545–44589.

Zonta, T., Da Costa, C. A., da Rosa Righi, R., De Lima, M. J., Da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the industry 4.0: A systematic literature review. *Computers & industrial engineering*, *150*, 106889.