Out-of-distribution Detection for Power System Text Data by Enhanced Mahalanobis Distance with Calibration

Yixiang Zhang, *Student Member*, *IEEE*, Huifang Wang, *Member*, *IEEE*, Yuzhen Zheng, Zhengming Fei, Hui Zhou, and Huafeng Luo

***Abstract*—The increasing significance of text data in power system intelligence has highlighted the out-of-distribution (OOD) problem as a critical challenge, hindering the deployment of artificial intelligence (AI) models. In a closed-world setting, most AI models cannot detect and reject unexpected data, which exacerbates the harmful impact of the OOD problem. The high similarity between OOD and in-distribution (IND) samples in the power system presents challenges for existing OOD detection methods in achieving effective results. This study aims to elucidate and address the OOD problem in power systems through a text classification task. First, the underlying causes of OOD sample generation are analyzed, highlighting the inherent nature of the OOD problem in the power system. Second, a novel method integrating the enhanced Mahalanobis distance with calibration strategies is introduced to improve OOD detection for text data in power system applications. Finally, the case study utilizing the actual text data from power system field operation (PSFO) is conducted, demonstrating the effectiveness of the proposed OOD detection method. Experimental results indicate that the proposed method outperformed existing methods in text OOD detection tasks within the power system, achieving a remarkable 21.03% enhancement of metric in the false positive rate at 95% true positive recall (FPR95) and a 12.97% enhancement in classification accuracy for the mixed IND-OOD scenarios.**

***Index Terms*—Out-of-distribution detection, text classification, text data applications in power grid, machine learning, natural language processing.**

I. Introduction

With the ongoing advancement of intelligence and digital transformation of power systems, the demand for artificial intelligence (AI) models has significantly increased [1]–[3]. Among these innovations, natural language processing (NLP) techniques have emerged as transformative tools with the potential to revolutionize engineering by addressing increasingly complex challenges [4]–[6]. However, deploying these models in real-world applications remains challenging, particularly concerning their reliability [7]. While most AI models prioritize performance metrics such as accuracy, it is imperative to acknowledge that ensuring model reliability takes precedence over marginal accuracy improvements [8]. A fundamental challenge in achieving this end is that these models are inherently trained under a closed-world setting—presuming that the distributions of training and test data are identical. Although this assumption holds in controlled experimental settings, it does not consistently apply to real-world scenarios and is not always valid [9]. The absence of mechanisms to discern unforeseen data introduces a critical yet frequently overlooked problem: the out-of-distribution (OOD) problem.

The OOD problem in classification tasks [10] arises when a classifier encounters input samples that differ from its training data distribution, resulting in unreliable predictions. Without safeguards, the classifier assigns labels to an OOD sample to produce an output, often with unwarranted confidence. This issue is particularly exacerbated in deep neural network-based probabilistic classifiers, which frequently overestimate their certainty, even for OOD samples. Consequently, the model generates unreliable predictions without additional scrutiny.

In power systems, model security demands heighten

Received: January 21, 2025
Accepted: October 10, 2025
Published Online: January 1, 2026

Yixiang Zhang, Huifang Wang (corresponding author), and Yuzhen Zheng are with the College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 12110023@zju.edu.cn; huifangwang@zju.edu.cn; 12410085@zju.edu.cn).

Zhengming Fei is with the State Grid East China Branch, Shanghai 200120, China (email: fei_zm@ec.sgcc.com.cn).

Hui Zhou is with the State Grid Zhejiang Electric Power Co., Ltd., Hangzhou 310007, China (e-mail: 7150452@qq.com).

Huafeng Luo is with the Electric Power Research Institute of State Grid Zhejiang Electric Power Co., Ltd., Hangzhou 310014, China (e-mail: luohuafeng2014@126.com).

DOI: 10.23919/PCMP.2024.000406

the implications of such outputs, potentially leading to significant misinterpretations. This study presents a case study based on the risk assessment task outlined in [5]. The classification errors induced by OOD samples may be more critical than those originating from the model, highlighting inconsistencies between training and test dataset definitions. Consequently, the trained model may be unfit for deployment, often without the knowledge of operators.

Recent studies on text OOD detection aim to improve performance using arbitrary public in-distribution (IND) and OOD dataset pairs, such as applying a sentiment analysis dataset as IND and a natural language inference dataset as OOD [11], [12]. However, OOD samples in power systems are often linguistically similar to IND samples, referred to as near-OOD, which hampers the effectiveness of existing methods.

In summary, there exists a notable research gap regarding the OOD problem in data-driven tasks, particularly within the power system domain. Previous studies have largely neglected this problem, hindering the practical application of findings in complex real-world environments. Unlike OOD research in the general domain, power system text data often exhibit a high degree of similarity between IND and OOD samples, underscoring the need for specialized research to address the unique challenges posed by these complex scenarios.

To address this issue, this study proposes a novel method that combines Mahalanobis distance [13] with automatic principle component analysis (PCA) [14] for dimensionality reduction and probability calibration. The proposed method divides the OOD problem into two phases: model training and sample scoring. Utilizing the strengths of existing OOD detection techniques in both probability and feature space ensures superior OOD detection performance while maintaining interpretability. The enhancements, derived from linear modules, avoid introducing significant complexity. All experiments utilize a real power dataset from power system field operation (PSFO), focusing on "near-OOD" scenarios within the power domain. To simulate a real application scenario, all data available for model training are considered IND, with OOD samples only accessible during the testing process.

The remainder of this paper is structured as follows. Section II explicitly defines the OOD problem. Section III introduces the proposed method and discusses its similarities and differences with the methods discussed in related works. Section IV details the experimental setting, and Section V presents the results demonstrating the effectiveness of the proposed method. Finally, Section VI concludes this study's experimental findings and analysis.

II. Problem Formulation

The OOD problem is primarily encountered in real-world applications, rather than in controlled experimental settings, leading to its limited consideration in research. This section defines the OOD problem within the context of classical classification tasks and explores the challenges of implementing OOD detection in power systems.

A. Classical Classification Problem

In a supervised multi-class classification task, the input space of training data is represented as *X*, and the label space containing *K* classes is denoted as $Y=\{1,2,\cdots,K\}$. The training dataset D_{train} = $\{(x_i,y_i)\}_{i=1}^{N}$ consists of *N* training samples (x_i,y_i), where each $x_i \in X$ and each $y_i \in Y$. The distribution of *X* is indicated as P_{in}, implying the input distribution under the assumption of the IND for the classification task. The primary objective of the training process is to derive a classifier $f: x \rightarrow \mathbb{R}^K$ with learnable parameters denoted as θ, aiming to minimize the global expected risk loss. Given the difficulty of calculating global performance, a practical approach is to compute the empirical risk loss, which represents the local classification performance on the training dataset [15]. A standard classifier training method aims to minimize the following loss function:

$$R_{\text{emp}}=\frac{1}{N}\sum_{i=1}^{N}L\left(f(x_i;\theta),y_i\right) \tag{1}$$

where a typical loss function *L* refers to cross-entropy with Softmax $L_{\text{CE}}(f,y)$:

$$L_{\text{CE}}(f,y)=-\log p(y\mid x)=-\log\frac{\mathrm{e}^{f(x;\theta)|y}}{\sum_{k=1}^{K}\mathrm{e}^{f(x;\theta)|k}} \tag{2}$$

where $p(y\mid x)$ is the predicted probability for the ground truth; $f(x;\theta)\mid y$ signifies the *y*th element of $f(x;\theta)$ (*y* denotes the true label of *x*); and $f(x;\theta)\mid k$ signifies the *k*th element of $f(x;\theta)$.

After training, the model's final performance is assessed using the test dataset D_{test}.

B. Classical OOD Problem

As previously mentioned, the OOD problem occurs when samples follow a distribution different from P_{in}, which is denoted as P_{in}. Deep learning models generally assume that test and training data share the same distribution—a key assumption for evaluating the model's performance using the outcomes based on the test dataset. Consequently, the classification model cannot generate reliable outputs under an OOD scenario.

To address this issue, a scoring function is constructed to distinguish IND and OOD data. The score $S(x)$ is based on a fundamental principle:

$$\begin{cases} S(x) > \varepsilon,\ x \in P_{\text{in}} \\ S(x) < \varepsilon,\ x \notin P_{\text{in}} \end{cases} \tag{3}$$

where ε denotes the predefined threshold. The construction of $S(x)$ is inherently linked to the training process, making it theoretically difficult to obtain OOD samples for auxiliary purposes.

Researchers have examined multiple methods to construct scores [16]. However, these methods are primarily categorized into probability-based and feature-based methods.

Probability-based methods primarily analyze the model's output known as "logits" before the Softmax layer in deep learning. The first OOD detection method, maximum Softmax probability (MSP), is introduced in [17]. Based on MSP, out-of-distribution detector in neural networks (ODIN), a post-hoc technique that applies temperature scaling and input perturbation to differentiate IND and OOD samples, is proposed in [18]. G-ODIN [19] enhances ODIN by incorporating generative OOD techniques. In [20], overconfidence is mitigated by normalizing the logits during the training process. Additionally, in [21], an unconventional scoring mechanism that treats logits as vectors to utilize cosine similarity and magnitude is proposed.

Feature-based methods detect OOD samples by analyzing deviations in the feature space relative to IND samples. In [22], the minimum Mahalanobis distance to class centroids is employed for detection. In [23], a non-parametric nearest-neighbor distance approach for OOD detection is proposed. The utility of feature norms in the orthogonal complement space, alongside sample-to-class centroid distance, is highlighted in [24]. A reconstruction-based approach within the feature space is utilized in [25] for anomaly detection.

In summary, extensive research on OOD detection with general data has established theoretical foundations relevant to text-based OOD detection in power systems.

C. OOD Problem in Power System

To demonstrate the OOD problem in the power system, the risk assessment task of the PSFO dataset is employed, as outlined in Section IV [5]. Several factors contribute to the generation of OOD samples, notably an incomplete training dataset. For instance, a model trained on summer field operations may not accurately assess winter field operations. Furthermore, policy changes can introduce new operations or alter existing ones, resulting in OOD descriptions. For example, a model trained before 2019 would be ill-equipped to manage field operations related to COVID-19, as such scenarios were absent from its training dataset. Finally, the subjective nature of text input from operators in field operations increases sample variability, with some instances becoming OOD.

Most existing studies utilize various open-source datasets to construct IND-OOD pair, resulting in scenarios that are often far-OOD, which simplifies the OOD detection task. However, in real-world applications within the power system domain, it is typically uncommon for an operator to input data entirely unrelated to the PSFO. Instead, operators are more likely to input domain-relevant data derived from new regulations. Consequently, OOD samples often share terminology similar to that found in IND samples, complicating the differentiation process [26].

This observation suggests that in PSFO, the scenario is closer to near-OOD, where significantly different OOD samples are infrequent. Given the increased complexity of the task, the current methods face notable limitations in effectively detecting OOD samples within power systems, highlighting the urgent need for more robust methodologies.

III. Method Architecture

In this study, the OOD detection task involves two primary objectives: training an adept classifier for IND samples and constructing a scoring mechanism to distinguish IND and OOD samples. To achieve these dual objectives, a framework is proposed that decomposes the original task into two distinct subtasks, which are specifically defined below.

1) Classification subtask: The classification subtask focuses on accurately classifying samples while investigating a more effective feature representation method for OOD detection.

2) Sample-scoring subtask: The sample-scoring subtask involves utilizing the identified feature representation method to develop a scoring system that maximizes the dissimilarity between OOD and IND samples.

The overall structure, depicted in Fig. 1, comprises the two subtasks described above. The encoder is trained during the classification subtask and subsequently utilized in the sample-scoring subtask. This design enables the modular integration of probability-based and feature-based methods.

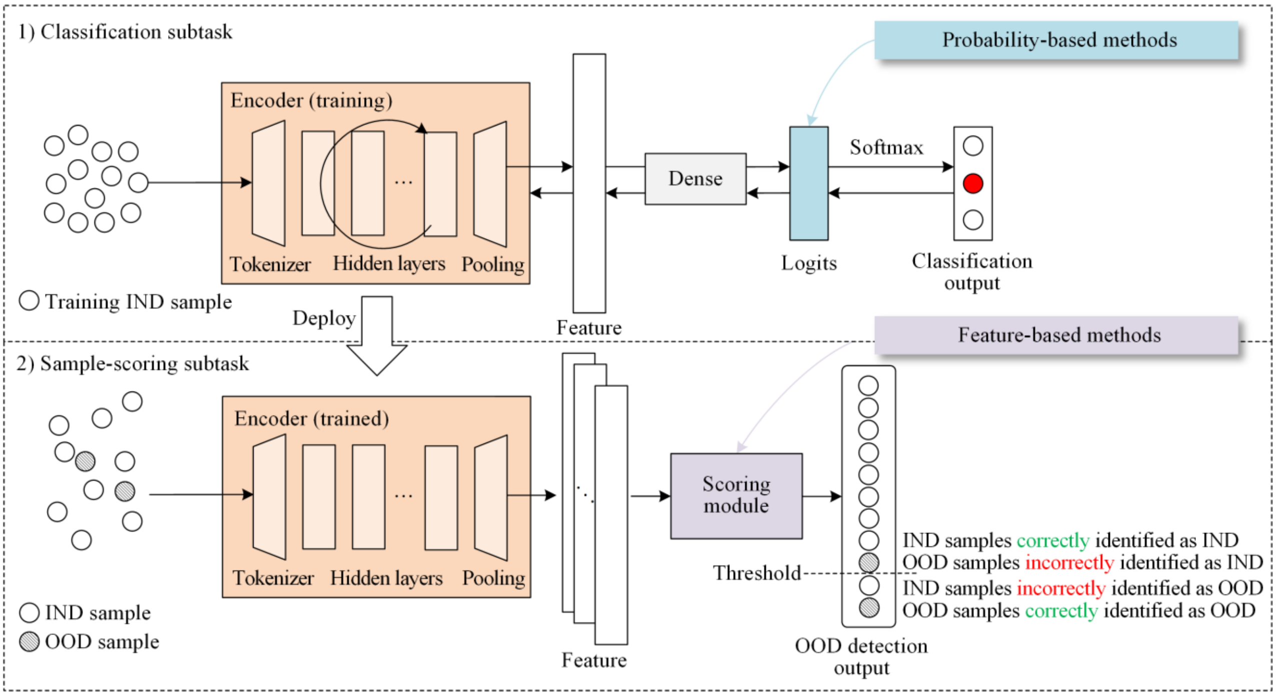

Fig. 1. Architecture of the proposed 2-subtask framework for OOD detection.

A. Temperature Scaling: Probability-based Method for Classification Subtask

Temperature scaling is a simple yet effective probability-based method. Reference [27] shows that it successfully calibrates deep neural networks, addressing the issue of overconfidence. Nevertheless, its sensitivity to parameters remains a challenge.

Temperature scaling builds upon MSP. As indicated in (4), MSP employs the highest probability assigned by the model to each class as its scoring function.

$$S_{\mathrm{MSP}}(x)=\max_i\left(\frac{\mathrm{e}^{f(x;\theta)|i}}{\sum_{k=1}^{K}\mathrm{e}^{f(x;\theta)|k}}\right),\ i\in Y \tag{4}$$

where $f(x;\theta)|i$ signifies the ith element of $f(x;\theta)$. Despite its simplicity and intuitive appeal, MSP demonstrates considerable overconfidence issues in deep neural networks, leading to nearly every output being assigned confidence values close to 1.0, irrespective of whether the samples are IND or OOD.

This overconfidence arises from the subtle disparity between the model's optimization through the loss function and its actual classification training objective. Although the model aims to classify samples accurately to improve accuracy, the loss function can still be minimized by adjusting the proportion of $\mathrm{e}^{f(x;\theta)|y}$ in $\sum_{k=1}^{K}\mathrm{e}^{f(x;\theta)|k}$, even when samples are correctly classified. The mode increases the magnitude of $f(x;\theta)$, which minimally improves classification accuracy but substantially elevates $S_{\mathrm{MSP}}(x)$. It can artificially inflate confidence scores using methods that do not necessarily enhance classification accuracy, thereby worsening the overconfidence issue.

As discussed earlier, overconfidence results in a discrepancy between accuracy and confidence. Reference [27] characterizes this phenomenon as a probabilistic calibration issue, suggesting that calibration can improve performance by controlling excessive growth in $S(x)$. Temperature scaling of cross-entropy loss is an effective calibration method, which involves appropriately rescaling $f(x;\theta)$ during training. The modified loss function is expressed as follows:

$$L_{\mathrm{CE\_T}}(f,y)=-\log\frac{\mathrm{e}^{f_y/T}}{\sum_{k=1}^{K}\mathrm{e}^{f_k/T}} \tag{5}$$

where e^{f_y} denotes $\mathrm{e}^{f(x;\theta)|y}$; e^{f_k} denotes $\mathrm{e}^{f(x;\theta)|k}$; and T represents the temperature parameter.

Figure 2 illustrates the OOD detection result based on the temperature scaling of PSFO. Temperature scaling simultaneously exhibits a high sensitivity to T. When $T<1$, the confidence of both IND and OOD samples decreases. A large T causes many OOD samples to maintain high confidence levels, whereas a small T causes many IND samples to enter low-confidence regions. Both scenarios result in poor detection performance. If T is excessively small, as illustrated in Fig. 2(a), the confidence scores of IND samples may even be lower than those of OOD samples, signaling model failure. Figures 2(b)–(e) illustrate how the results systematically vary with different values of T. In such a case, the model performs optimally only when $T=0.5$. However, obtaining an optimal T is nearly impossible, as the model lacks access to OOD samples during training. This challenge, compounded with the sensitivity of T, complicates the direct application of temperature scaling to the OOD problem.

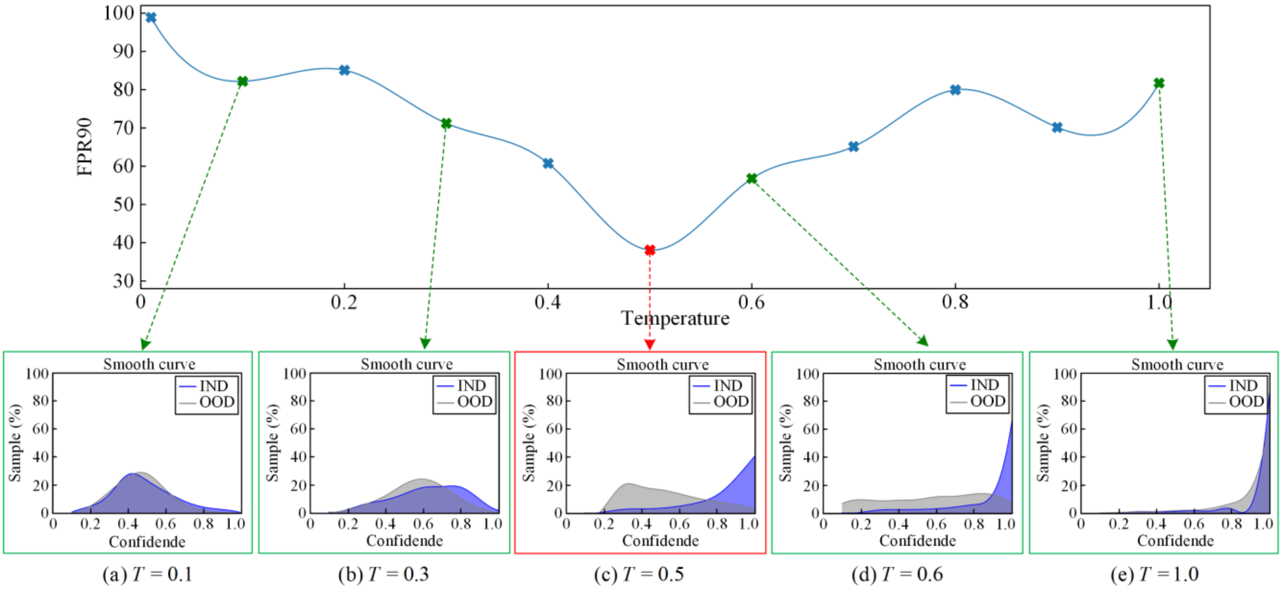

Fig. 2. Impact of T in temperature scaling in PSFO. (a)–(e) Confidence distribution of IND (blue area) and OOD (gray area) samples at various T values.

Regrettably, this issue persists when relying solely on probability-based methods. In the proposed framework, temperature scaling is applied exclusively in the classification subtask to calibrate the encoder's feature extractor, without affecting the scoring process. By preserving independence in the sample-scoring subtask, the proposed method achieves effective probabilistic calibration using temperature scaling, thereby alleviating the challenges posed by the high sensitivity of T in scoring processes.

B. Enhanced Mahalanobis Distance: Feature-based Method for Sample-scoring Subtask

In the proposed framework, OOD detection is conducted independently through a feature-based method within the sample-scoring subtask. Recent studies underscore the benefits of leveraging the high-dimensional feature space to enhance scoring in OOD detection [28]. Despite the benefits of the feature space, researchers, including [29], have identified two detrimental issues, that impede its effective utilization: anisotropy and sparsity.

To address these challenges, this study employed Mahalanobis distance to mitigate anisotropy and enhanced it with multi-scaled PCA to address sparsity.

1) Mahalanobis Distance for Anisotropy

Anisotropy refers to the tendency of word embeddings to form highly concentrated distributions within the feature space, often observed in pretrained transformers such as bidirectional encoder representations from transformers (BERT), generative pre-trained transformers (GPT), and sentence embedding [30]. Reference [31] indicates that training language models using a maximum likelihood leads to varying importance of word embeddings within high-dimensional space, leading to anisotropy due to differences in their effectiveness at capturing semantic meaning.

BERT-flow [28] is introduced to enhance sentence embeddings. More recently, BERT-whitening [32] is proposed, achieving comparable performance through whitening. Whitening is a linear transformation that applies a transformation matrix to the original matrix, resulting in whitened variables with an identity covariance matrix [33].

Building on BERT-whitening, this study utilizes Mahalanobis distance, which is a modified Euclidean distance metric under a whitening transformation to resolve dimensional correlations. By mitigating anisotropy, Mahalanobis distance effectively fulfills the role of whitening.

Within the proposed framework, the specific method comprises two steps.

1) The dataset D_{train} is partitioned into K label-based subsets, with the mean and covariance computed for each subset as follows:

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_i \tag{6}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k - 1}(\boldsymbol{X}_k - \mu_k)^{\mathrm{T}}(\boldsymbol{X}_k - \mu_k) \tag{7}$$

where N_k represents the number of samples labeled k; $\boldsymbol{X}_k$ indicates the matrix containing samples from the kth class in D_{train}; μ_k and $\boldsymbol{\Sigma}_k$ denote the mean and covariance matrix of $\boldsymbol{X}_k$, respectively.

2) The Mahalanobis distance of the test sample x is computed for each subset, and the minimum value is selected as the result. The corresponding equation is expressed as follows:

$$S_{\mathrm{MAHA}}(x)=\min_{k}\sqrt{(x-\mu_k)^T\boldsymbol{\Sigma}_k^{-1}(x-\mu_k)} \qquad (8)$$

where x is considered likely to be OOD when it exhibits a substantial Mahalanobis distance from any class. In practical experiments, the negative of $S_{\mathrm{MAHA}}(x)$ is computed and subsequently normalized to the range of (0, 1) to align with the previous notion of confidence, denoted as $S_{\mathrm{MSP}}(x)$.

2) Multi-scaled PCA for Sparsity

Sparsity, signifying redundant dimensions, is prevalent in neural networks. For instance, the widely used 768-d BERT embedding [34] for text data exhibits considerable sparsity in its feature space, as demonstrated in Fig. 3. PCA applied to PSFO data demonstrates that 5% of the feature values can capture over 95% of the features. Recent studies prove that constructing a low-dimensional manifold in the original feature space helps mitigate sparsity, with PCA-based dimensionality reduction proving effective [35].

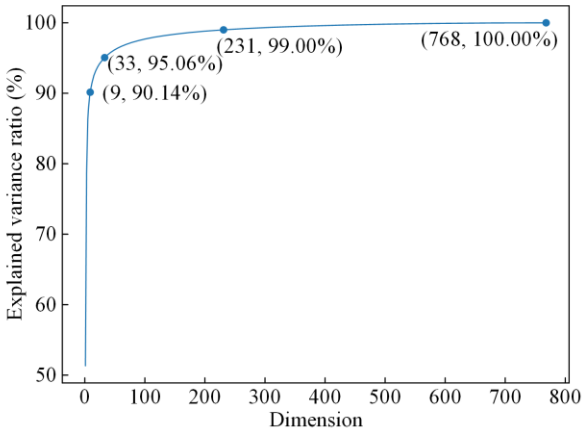

Fig. 3. PCA results of PSFO.

Mahalanobis distance standardizes data by transforming it into a space where all dimensions have zero mean, unit variance and zero correlation. However, this standardization may potentially overemphasize certain features while neglecting nuanced distinctions between IND and OOD samples. To address this limitation, combining Mahalanobis distance with PCA has proven to be an effective strategy, benefiting both sparsity and the enhancement of Mahalanobis distance.

However, selecting the optimal PCA dimension is often challenging and dataset-specific. To address this issue, this study proposes a multi-scaled PCA approach that eliminates the need for hyperparameter selection. The proposed scoring module comprises three steps: 1) extracting features from the encoder; 2) downscaling high-dimensional features into multiple low-dimensional features using PCA, including principal components that capture cumulative explained variances of 95%, 97.5%, and 99%; 3) calculating the Mahalanobis distance for each low-dimensional feature and then aggregating these distances. The structure of the proposed scoring module is depicted in Fig. 4.

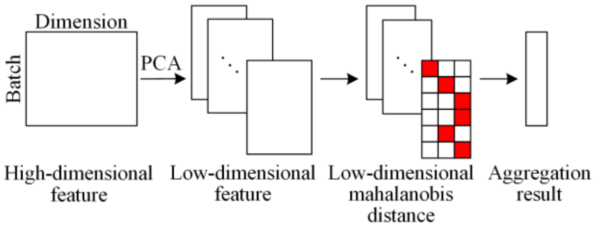

Fig. 4. Structure of the scoring module using multi-scaled PCA.

As depicted in Fig. 4, the low-dimensional Mahalanobis distance reflects the distance between the sample and the nearest class, as indicated by the red block. The final result is the aggregation of Mahalanobis distances across all scales.

C. Pooling Strategy

The encoder is a module that transforms text data into high-dimensional features. In this study, the BERT encoder is fully utilized, comprising a 768-d embedding with 12 hidden layers. The features are represented by the [CLS] token from the final layer, where the [CLS] token serves as a sentence vector for classification.

Recent studies have explored various pooling strategies for effective feature extraction. For instance, reference [36] has demonstrated that holistic pooling strategies, such as last-layer average pooling or first-last-layer average pooling, generally enhance performance in classification tasks. However, after a thorough comparison of several pooling strategies, it is concluded that pooling strategies do not significantly impact OOD detection within the proposed framework. Consequently, the simplest approach, using the [CLS] token from the final layer, is preferred. Detailed outcomes are provided in Section IV.

D. OOD Detection

The OOD detection process, illustrated in Fig. 5, begins by assessing whether a given test sample qualifies as OOD. If the sample is not classified as OOD, it advances to the classification task.

In the classification subtask, the BERT model is trained on all available IND data and subsequently calibrated using temperature scaling. For the sample-scoring subtask, features are extracted by the BERT encoder and trained on IND data, and the test sample's scores are computed using the Mahalanobis distance, enhanced by multi-scaled PCA. The mean and covariance matrix used in the Mahalanobis distance are derived from IND, thereby facilitating distinct feature representations for IND and OOD samples. Samples scoring below 95% of the IND samples are flagged as OOD and rejected.

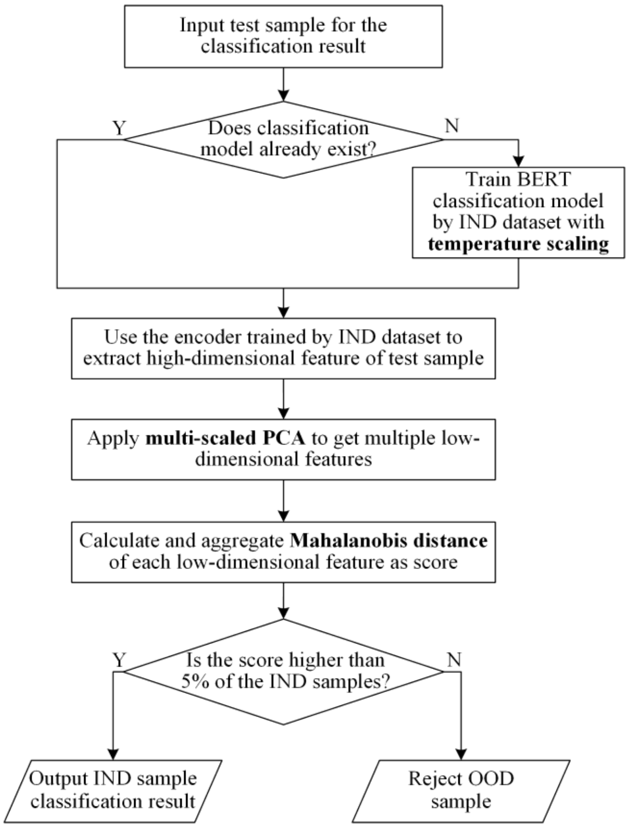
Input test sample for the classification result
Does classification model already exist?
Y
N
Train BERT classification model by IND dataset with **temperature scaling**
Use the encoder trained by IND dataset to extract high-dimensional feature of test sample
Apply **multi-scaled PCA** to get multiple low-dimensional features
Calculate and aggregate **Mahalanobis distance** of each low-dimensional feature as score
Is the score higher than 5% of the IND samples?
Y
N
Output IND sample classification result
Reject OOD sample

Fig. 5. OOD detection process of an input test sample.

IV. Experiment Settings

A. Dataset

This study examines the OOD challenges observed in the risk assessment task of PSFO. To enhance real-world applicability, the data for this case study are sourced from actual PSFO operations in a Chinese province, encompassing 101195 samples. Detailed information on the dataset is provided in Table I and Fig. 6.

For comparative analysis, the THUCNews dataset [37], consisting of Chinese news titles, is also included. This dataset contrasts with the PSFO data in terms of text structure, length, and other inherent characteristics.

TABLE I
Number and Percentage of Operations Based on Multiple Types of Operation

Type of operation	Number of samples	Percentage (%)
Manufacturing maintenance and modification	40 292	39.816
Meter installation	32 949	32.560
Rural grid project	9604	9.491
Independent operation	7220	7.135
Business expansion	3384	3.344
Relocation and modification	3110	3.073
Power grid construction	1684	1.664
External project	1497	1.479
Others	1455	1.438

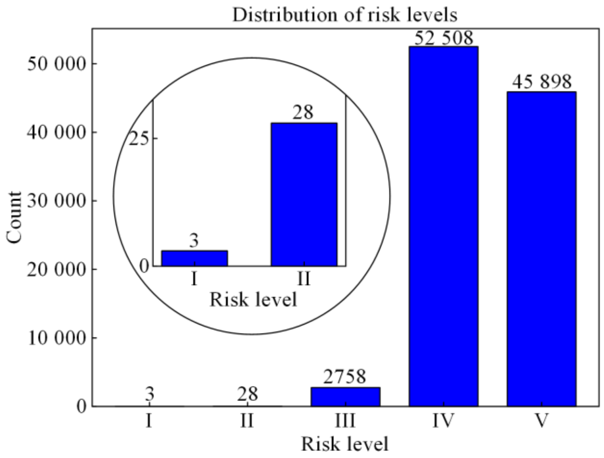

Fig. 6. Histogram of PSFO categorized with varying risk levels.

B. Training Details

The main framework employs the pre-trained Chinese BERT model sourced from Hugging Face, with 768 dimensions and 12 hidden layers. It is trained for two epochs using the AdamW optimizer, with a weight decay of 0.01, a batch size of 16, and a padding size of 256 tokens per sample. The learning rate is set to 2×10^{-5}. All models attain a minimum of 90% accuracy on the classification task of the IND dataset.

C. Evaluation Metrics

The performance of OOD detection is assessed using the following metrics: the false positive rate at 90% true positive recall (FPR90), the false positive rate at 95% true positive recall (FPR95), the area under the receiver operating characteristic curve (AUROC), and the area under the precision-recall curve (AUPR). These metrics collectively evaluate the model's ability to differentiate between OOD and IND samples. FPR90 and FPR95 directly measure the model's separation quality, while AUROC and AUPR provide an overall assessment of its performance. A lower FPR90 and FPR95, or a higher AUROC and AUPR, signify superior model performance.

D. Methods of Comparison

Unlike the OOD detection research on the visual domain, studies focusing on text data are limited. This study evaluates the proposed method against six techniques, most of which have been previously explored in OOD research on text data. The techniques include probability-based methods of MSP [17], LogitNorm [20], MaxLogit [21], and temperature scaling [27] and feature-based methods of Mahalanobis distance [22] and ViM [24]. Among these methods, ViM has achieved state-of-the-art results for text data [38]. However, previous studies do not include specialized datasets from the power systems. Therefore, a comprehensive comparison experiment is conducted using the PSFO to demonstrate the proposed method's effectiveness, particularly in the power system domain.

For temperature scaling, LogitNorm, and MaxLogit, parameter T is assigned various values between 0 and 1.0 and the calculation details for ViM are outlined in [24]. This study specifically compares the original Mahalanobis distance with its version downscaled to 200 dimensions. Additionally, the Euclidean distance, though not a conventional OOD detection method, is included solely for comparison with the Mahalanobis distance.

E. Experiment Environment

The proposed methods are implemented on the VScode platform using Anaconda software. The deep learning framework utilized in the experiment is Pytorch 1.11.0. The computing hardware includes an Intel i9-10900K CPU (3.70 GHz, 10 cores), 64 GB of RAM, and an RTX3090 GPU with 24 GB of memory.

V. Experiment Result and Analysis

A. OOD Detection Experiment

This study employs two different IND-OOD dataset pairs for the experiment to investigate both far-OOD and near-OOD detection challenges. For the far-OOD scenario, the entire PSFO dataset serves as the IND dataset, with 10 000 randomly sampled samples from THUCNews forming the OOD dataset. In the near-OOD scenario, both IND and OOD datasets are derived from PSFO, where the "meter installation" samples are designated as the OOD dataset, while the remaining samples are designated as the IND dataset. The selection of "meter installation" samples is based on two key criteria: sufficient sample size and a diverse range of internal representations. These samples are classified as the OOD dataset due to their association with risk levels I and II, as shown in Fig. 6.

In the experiment, particularly in the near-OOD scenario, PSFO serves as the source for both IND and OOD, removing any potential impact from text lengths and structures. This observation demonstrates that the performance achieved by the proposed method stems from understanding the content's meaning rather than mis-learned knowledge of text lengths or structures.

Table II presents the OOD detection result of different methods for PSFO, where bold numbers indicate strong performance and underlined numbers represent the best results. The "far-OOD" column represents the PSFO and THUCNews as the IND and OOD datasets. The "near-OOD" column designates PSFO, excluding "meter installation," as the IND dataset, with the remaining meter installation samples acting as the OOD dataset. The proposed method (Table II) leverages the strengths of probability-based and feature-based approaches, achieving the best experimental performance. After parameter optimization, probability-based methods demonstrate robust performance, outperforming feature-based methods in effectiveness. Feature-based methods such as ViM perform well without the need for parameter settings.

TABLE II
OOD Detection Outcomes for PSFO

Method		Parameter	Far-OOD				Near-OOD			
			FPR90↓	FPR95↓	AUROC↑	AUPR↑	FPR90↓	FPR95↓	AUROC↑	AUPR↑
Probability-based method	MSP		70.49	87.62	78.78	82.63	71.32	87.57	82.64	71.87
	Temperature scaling	$T = 0.75$	56.81	78.31	84.94	88.62	**67.37**	**85.38**	**84.29**	**74.88**
		$T = 0.5$	**38.09**	**61.99**	**87.96**	**89.63**	78.10	94.42	80.03	68.28
		$T = 0.25$	76.76	89.70	77.76	82.85	76.43	97.70	81.52	67.00
	LogitNorm	$T = 1.0$	32.98	75.30	89.27	91.99	50.68	**66.33**	85.61	65.66
		$T = 0.7$	**16.08**	**49.70**	**93.54**	**95.46**	50.55	72.94	85.08	62.60
		$T = 0.4$	60.74	81.65	83.99	87.68	**43.31**	67.60	**88.51**	**76.08**
		$T = 0.1$	73.26	90.68	84.28	89.55	84.90	98.42	76.43	62.48
	Max logit	$T = 1.0$	76.06	89.42	76.51	80.69	94.72	99.31	75.27	65.55
		$T = 0.7$	**15.58**	**42.22**	**94.02**	**95.69**	92.03	99.42	76.93	59.09
		$T = 0.4$	73.57	89.65	79.37	84.03	96.92	99.08	64.99	43.14
Feature-based method	Euclidean distance		50.96	78.51	84.38	87.16	96.28	98.38	54.81	31.46
	Mahalaonbis distance		35.94	53.94	89.90	92.02	66.75	85.17	79.57	63.15
	Mahalaonbis distance	PCA:200d	30.69	48.56	91.59	93.51	68.00	83.54	80.16	67.06
	Mahalaonbis distance	Multi-scaled PCA	**16.52**	**29.14**	**95.01**	**96.03**	**54.03**	**71.79**	**84.25**	**70.72**
	ViM		23.96	38.75	92.24	93.26	74.12	83.89	77.71	64.14
Composite method	Proposed method		**9.92**	**22.99**	**96.15**	**97.09**	**24.30**	**45.30**	**91.32**	**76.96**

The proposed method leverages these advantages to achieve top performance across all four metrics in both far-OOD and near-OOD scenarios. It surpasses existing methods, especially in near-OOD scenarios that better reflect real-world applications. Compared to the previous state-of-the-art method, ViM, the proposed method reduces FPR90 by 14.04% and 49.82% in the far-OOD and near-OOD scenarios, respectively.

B. Result Analysis

The detailed analysis of experimental outcomes highlights several critical aspects.

1) Distinction Between Near-OOD and Far-OOD Scenarios

In OOD detection, the near-OOD scenario presents a greater challenge, closely resembling the situations encountered in real-world power system applications. Table II reveals that when the parameters are properly tuned, most methods demonstrate superior performance in far-OOD compared to near-OOD scenarios. In contrast to conventional tasks, the near-OOD detection task presents challenges not only with reduced performance metrics but also with the potential failure of widely used methods.

Among probability-based methods, temperature scaling and LogitNorm show significant disparities between far-OOD and near-OOD scenarios. In particular, temperature scaling achieves a reduction in FPR90 from 70.49% to 38.09% in the far-OOD scenario with a suitable *T*, surpassing its performance in the near-OOD scenario, where FPR90 only drops from 71.32% to 67.37%. This disparity underscores the challenge of improving metrics in near-OOD scenarios.

Results from feature-based methods further validate that far-OOD scenarios are more conducive to high performance than near-OOD scenarios. For example, the Mahalanobis distance, known for its simplicity and effectiveness, reaches an FPR90 of 30.69% in far-OOD scenarios, in stark contrast to the 68.00% FPR90 in the near-OOD scenarios.

Typically, the related work based on generalized text datasets does not consider near scenarios. While ViM achieved state-of-the-art performance in previous studies, the scenario was too simplistic when compared to PSFO. The experimental results reveal that, despite ViM's exceptional performance with an FPR90 of 23.96% in far-OOD scenarios, it underperforms in near-OOD scenarios, failing to meet expectations.

A total failure of the model is indicated when FPR90 and FPR95 achieve 90% and 95%, respectively, as this finding suggests that the model cannot detect any OOD samples from the test dataset. In practical terms, methods that are effective in far-OOD scenarios often lose effectiveness in near-OOD scenarios, such as those represented by PSFO.

In conclusion, although current methods are effective for simpler far-OOD scenarios, novel approaches are essential for effectively addressing challenges in near-OOD scenarios. The proposed method shows promising potential in this regard.

2) Parameter Sensitivity of Probability-based Methods

Probability-based methods often exhibit strong dependence and sensitivity to the parameter *T*, characterized by two primary aspects. First, optimal model performance is typically limited to a very narrow range of *T* values. Second, this study observes significant disparities in the optimal *T* values across different datasets, such as 0.7 and 0.4 for LogitNorm in far-OOD and near-OOD scenarios, respectively. Given the impracticality of accessing OOD data in real-world applications, adjusting *T* based on OOD datasets is unfeasible. Consequently, this dependence and sensitivity pose significant challenges to the model's successful deployment.

In contrast, the proposed method incorporates probability-based approaches for the classification subtask while avoiding direct sample scoring by logits, thereby minimizing sensitivity to *T*. The method enables parameter optimization to improve the performance. Its advantage lies in expanding the parameter range within which the model performs effectively, reducing the risk of minor parameter adjustments negatively impacting overall model effectiveness.

Figure 7 reveals that the proposed method, using the same IND-OOD pair as "far-OOD" (Table II), exhibits significantly lower sensitivity to *T* than conventional methods such as temperature scaling. Temperature scaling is mostly effective only when $T = 0.5$. In contrast, while the proposed method exhibits some variability, it consistently maintains strong performance, except in specific situations, such as when $T = 0.1$.

Objectively, an optimal value of *T* exists that maximizes the performance. Importantly, the proposed method performs well even with an arbitrary *T*, as long as it is not excessively extreme. This observation is because the proposed method significantly reduces the model's dependence on the parameter *T*.

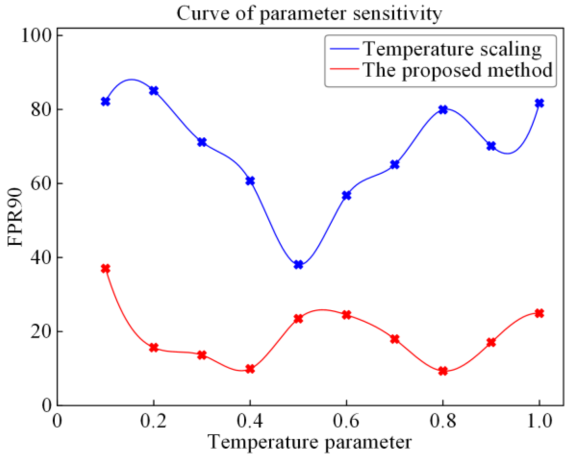

Fig. 7. FPR90 curve as a function of temperature parameter.

3) Ablation Study

The ablation study highlights the importance of each module within the proposed method. Table III presents the results of the ablation study for the OOD detection task of PSFO, using the same IND and OOD datasets as "near-OOD." The analysis examines the effects of various components and design choices on detection performance.

Table III reveals that the ablation study evaluates three modules of the proposed method: temperature scaling for calibration, Mahalanobis distance for scoring, and multi-scaled PCA for enhancement. PCA refers to the standard PCA, which has a target dimension of 200, whereas multi-PCA represents the abbreviation for multi-scaled PCA. All configurations, except for Choices 1 and 2, derive their scores from features, employing either MSP or temperature scaling during model training for the classification subtask. In the case of Choices 6–8, Mahalanobis distance should be used alongside PCA in this experiment. Choice 9 represents the proposed method. Table III highlights that all three modules can interact to significantly enhance the performance in OOD detection, underscoring the indispensability of these modules, particularly in challenging tasks near-OOD scenarios.

TABLE III
RESULTS OF THE ABLATION STUDY

Component	Choice								
	1	2	3	4	5	6	7	8	**9**
MSP	√		√	√		√		√	
Temperature scaling		√			√		√		√
Euclidean distance			√						
Mahalanobis distance				√	√	√	√	√	√
PCA						√	√		
Multi-PCA								√	√
FPR90↓	71.32	67.37	98.38	66.75	66.38	68.00	54.60	54.03	**24.30**
AUROC↑	82.64	84.29	54.81	79.57	79.42	80.16	84.18	84.25	**91.32**
AUPR↑	71.87	74.88	31.46	63.15	64.88	67.06	69.45	70.72	**76.96**

4) Choice of Pooling Strategy

The experiments demonstrate that selecting a pooling strategy is not a pivotal issue in tackling the OOD detection task within the power systems. Referring to [36], the study explores several pooling strategies, such as averaging the first layer and the last layer or averaging all layers from the BERT encoder.

Table IV shows that significant improvements are not observed by altering the pooling strategy within the proposed method. Some strategies even led to a decline in OOD detection performance. This outcome underscores that, for optimizing OOD detection performance in the scenario discussed, the choice of pooling strategy seems to be of little importance. Consequently, using the [CLS] token from the last layer, as is commonly practiced, remains a robust choice for the OOD detection task.

TABLE IV
RESULTS OF OOD DETECTION WITH VARIOUS POOLING STRATEGIES

Pooling strategy		FPR90↓	FPR95↓	AUROC↑	AUPR↑
Layers	Pooling				
Last	CLS	**24.30**	**45.30**	**91.32**	**76.96**
Last	AVG	78.72	86.52	73.10	57.29
Last	MAX	**30.79**	**50.96**	**90.07**	**75.58**
First & last	CLS	58.90	82.69	84.01	69.07
First & last	AVG	67.43	82.43	80.87	63.52
All	CLS	**29.96**	**56.04**	**90.98**	**79.14**
All	AVG	59.38	75.50	83.65	68.82

5) Real-world Application Improvement by OOD Detection

Reference [39] states that OOD detection should be recognized as a means rather than an end, emphasizing that evaluating performance on the original task is more important than focusing solely on OOD problems. In the risk assessment task of the PSFO, classification is the primary task. To simulate real-world application scenarios, 2000 IND and 2000 OOD samples are randomly selected from the IND-OOD dataset pair "near-OOD." These samples are mixed to form a dataset, and the classification accuracies are calculated under three scenarios: 1) calculating the 2000 IND samples to represent the experimental settings; 2) using the full 4000-sample mixed dataset to simulate the real-world application; 3) applying the proposed method to filter out OOD data from the mixed dataset to evaluate the effectiveness of OOD detection. The results are presented in Table V.

TABLE V
EXPERIMENTAL RESULTS ON REAL-WORLD APPLICATIONS

Scenario	Total number of samples	Number of IND samples	Number of OOD samples	Accuracy (%)
1)	2000	2000	0	91.70
2)	4000	2000	2000	73.45
3)	2372	1889	483	**86.42**

Table V reveals that the proposed OOD detection method successfully identified and excluded 1517 OOD samples while mistakenly filtering out only 111 IND samples. This result largely satisfies the requirement of filtering out OOD samples for the original classification task. The accuracy result of 91.70% achieved in scenario 1) reflects the classification model's expected performance in real-world applications. However, in scenario 2), there is a significant decline in classification performance, with accuracy dropping to 73.45% when all 4000 samples, including OOD samples, are considered. This outcome highlights the substantial impact of OOD samples on the model's reliability in real-world applications. In contrast, scenario 3) demonstrates that the proposed method effectively filters out OOD samples, thereby allowing the model to approach its anticipated performance and significantly enhancing the model's reliability.

Furthermore, the OOD problem can be viewed as the limitation of the training dataset. By labeling and incorporating the identified OOD samples, we can establish a continuous solution to mitigate the OOD problem.

VI. Conclusion

This study presents a simple framework for OOD detection, which decomposes the original task into two subtasks: classification and sample-scoring subtasks. This modular approach combines both probability-based and feature-based methods to improve OOD detection performance, especially in near-OOD scenarios within the power system domain. Temperature scaling is applied for calibration in the classification subtask, while multi-scaled PCA enhances Mahalanobis distance to compute the final score for each sample. Experiment results demonstrate that the proposed method achieves superior performance. In difficult far-OOD scenarios, where conventional methods often fail, this method reduces the FPR90 and FPR95 metrics by nearly 40%, demonstrating a marked improvement in the OOD detection performance. Another advantage of this framework is its modular design, which enables the easy replacement of individual modules with more advanced OOD detection methods—whether feature-based or probability-based—as they become available.

The proposed OOD detection method not only identifies OOD samples but also highlights potential limitations in the IND dataset, guiding updates to the dataset. However, despite utilizing real data from PSFO, the OOD scenarios evaluated in this study are simulated. To effectively tackle the OOD challenges in PSFO, it is crucial to have larger labeled datasets that cover diverse operational scenarios. Additionally, even when OOD samples are used to enhance the dataset quality, manual labeling and model retraining remain necessary. Overcoming these two challenges in future research could increase the practicality of OOD detection.

We firmly believe that the OOD problem will present shared characteristics and unique challenges across various situations. The scenario and the solution outlined in this study are not the sole possibilities, and we hope they act as a catalyst for further research that recognizes the OOD problem and proposes effective, application-specific solutions.

Acknowledgments

Not applicable.

Authors' Contributions

Yixiang Zhang: full-text writing, conceptualization, methodology, coding, and experiment design. Huifang Wang: methodology and writing guidance. Yuzhen Zheng: assisted coding. Zhengming Fei and Hui Zhou: real-world application guidance. Huafeng Luo: resource acquisition and data curation. All authors read and approved the final manuscript.

Funding

This work is supported in part by the Science and Technology Project of the State Grid East China Branch (No. 520800230008).

Availability of Data and Materials

Not applicable.

Declarations

Competing interests: The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this article.

Authors' Information

Yixiang Zhang received the B.Sc. degree from Zhejiang University, Hangzhou, China, in 2021. He is currently pursuing Ph.D. degree of electrical engineering in Zhejiang University, Hangzhou, China. His research mainly focuses on the application of AI techniques in power system, including data mining and natural language processing in smart grid.

Huifang Wang received the B.Sc. and M.Sc. degrees in power system relay protection from North China Electric Power University, Baoding, China, in 1995 and 1998, respectively, and the Ph.D. degree in power system and automation from Zhejiang University, Hangzhou, China, in 2006. She is currently an associate professor with the College of Electrical Engineering, Zhejiang University. Her current research interests include power system protection, condition-based maintenance, and the application of AI techniques in power system.

Yuzhen Zheng received the B.S. degree in electrical engineering from Zhejiang University, Zhejiang, China, in 2024, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interest includes artificial intelligence in modern power systems.

Zhengming Fei received the B.Sc. and M.Sc. degrees in electrical engineering from Shanghai University of Technology, Shanghai, China, in 1991 and 1994. His current research interests are safety management of power grid enterprises, emergency management of power grid enterprises.

Hui Zhou received Ph. D in power system and its automation from Huazhong University of Science and Technology, Wuhan, China in 2009. His current research interests are power system stability and smart grid data security.

Huafeng Luo received the B.Sc. in electrical engineering from Zhejiang University, Hangzhou, China, in 2011 and M.Sc. degrees in electrical engineering from Shanghai Jiao Tong University in 2014. His current research interests are power system security and AI applications of modern smart grid.

REFERENCES

[1] Z. Li, Z. Jiao, and A. He *et al.*, "A denoising-classification neural network for power transformer protection," *Protection and Control of Modern Power Systems*, vol. 7, no. 4, pp. 1-14, Oct. 2022.

[2] G. Lu, C. Tsang, and H. Yim *et al.*, "Interpretable fault diagnosis for overhead lines with covered conductors: a physics-informed deep learning approach," *Protection and Control of Modern Power Systems*, vol. 10, no. 2, pp. 25-39, Mar. 2025.

[3] N. Yang, J. Hao, and Z. Li *et al.*, "Data-driven decision-making for SCUC: an improved deep learning approach based on sample coding and seq2seq technique," *Protection and Control of Modern Power Systems*, vol. 10, no. 2, pp. 13-24, Mar. 2025.

[4] H. Wang, Z. Liu, and Y. Xu *et al.*, "Short text mining framework with specific design for operation and maintenance of power equipment," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 6, pp. 1267-1277, Nov. 2021.

[5] H. Wang, J. Cao, and D. Lin, "Deep analysis of power equipment defects based on semantic framework text mining technology," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 4, pp. 1157-1164, Jul. 2022.

[6] H. Wang, N. Zhou, and R. Huang *et al.*, "110 kV signal semantic analysis and situation awareness model based on deep learning theory for a power system monitoring system," *Power System Protection and Control*, vol. 51, no. 2, pp. 160-168, Jan. 2023. (in Chinese)

[7] J. Yu, H. Wang, and Y. Zhang *et al.*, "Automatic risk rating method for power grid field operation based on BERT," *Power System Technology*, vol. 47, no. 11, pp. 4746-4754, Jun. 2023. (in Chinese)

[8] T. Dragičević, P. Wheeler, and F. Blaabjerg, "Artificial intelligence aided automated design for reliability of power electronic systems," *IEEE Transactions on Power Electronics*, vol. 34, no. 8, pp. 7161-7171, Aug. 2019.

[9] G. Fei and B. Liu, "Breaking the closed world assumption in text classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, USA, Jun. 2016, pp. 506-514.

[10] Z. Liu, Z. Miao, and X. Zhan *et al.*, "Large-scale long-tailed recognition in an open world," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, USA, Jun. 2019, pp. 2532-2541.

[11] M. Tan, Y. Yu, and H. Wang *et al.*, "Out-of-domain detection for low-resource text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3566-3572.

[12] U. Arora, W. Huang, and H. He, "Types of out-of-distribution texts and how to detect them," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, Nov. 2021, pp. 10687-1070.

[13] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1-18, Jan. 2000.

[14] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37-52, Aug. 1987.

[15] S. Minaee, N. Kalchbrenner, and E. Cambria *et al.*, "Deep learning-based text classification: a comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, Apr. 2021.

[16] J. Yang, K. Zhou, and Y. Li *et al.*, "Generalized out-of-distribution detection: a survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635-5662, Jun. 2024.

[17] G. Shalev, G. Shalev, and J. Keshet, "A baseline for detecting out-of-distribution examples in image captioning," in *Proceedings of the 30th ACM International Conference on Multimedia*, New York, USA, Oct. 2022, pp. 4175-4184.

[18] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *the International Conference on Learning Representations*, Vancouver, Canada, Feb. 2018.

[19] Y. C. Hsu, Y. Shen, and H. Jin *et al.*, "Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, Jun. 2020, pp. 10948-10957.

[20] H. Wei, R. Xie, and H. Cheng *et al.*, "Mitigating neural network overconfidence with logit normalization," in *Proceedings of the 39th International Conference on Machine Learning*, Baltimore, USA, Jun. 2022, pp. 23631-23644.

[21] Z. Zhang and X. Xiang, "Decoupling maxlogit for out-of-distribution detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, Jun. 2023, pp. 3388-3397.

[22] K. Lee, K. Lee, and H. Lee *et al.*, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, New York, USA, Dec. 2018, pp. 7167-7177.

[23] Y. Sun, Y. Ming, and X. Zhu *et al.*, “Out-of-distribution detection with deep nearest neighbors,” in *Proceedings of the 39th International Conference on Machine Learning*, Baltimore, USA, Jun. 2022, pp. 20827-20840.

[24] H. Wang, Z. Li, and L. Feng *et al.*, “ViM: out-of-distribution with virtual-logit matching,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, Jun. 2022, pp. 4911-4920.

[25] J. An and S. Cho, (2015, Dec.) “Variational autoencoder based anomaly detection using reconstruction probability,” [Online]. Available: https://www.semanticscholar.org/paper/Variational-Autoencoder-based-Anomaly-Detection-An-Cho/061146b1d7938d7a8dae70e3531a00fceb3c78e8

[26] S. Fort, J. Ren, and B. Lakshminarayanan, “Exploring the limits of out-of-distribution detection,” in *35th Conference on Neural Information Processing Systems*, Online, Jul. 2021, pp. 7068-7081.

[27] C. Guo, G. Pleiss, and Y. Sun *et al.*, “On calibration of modern neural networks,” in *Proceedings of the 34th International Conference on Machine Learning*, Toulon, France, Jul. 2017, pp. 1321-1330.

[28] J. Ren, S. Fort, and J. Liu *et al.*, “A simple fix to mahalanobis distance for improving near-OOD detection,” *arXiv*: 2106.09022, Jun. 2021.

[29] B. Li, H. Zhou, and J. He *et al.*, “On the sentence embeddings from pre-trained language models,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 9119-9130.

[30] K. Ethayarajh, “How contextual are contextualized word representations comparing the geometry of BERT, ELMo, and GPT-2 embeddings,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 55-65.

[31] Z. Zhang, C. Gao, and C. Xu *et al.*, “Revisiting representation degeneration problem in language modeling,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, Nov. 2020, pp. 518-527.

[32] J. Su, J. Cao, and W. Liu *et al.*, “Whitening sentence representations for better semantics and faster retrieval,” *arXiv*: 2103.15316, Mar. 2021.

[33] A. C. Koivunen and A. B. Kostinski, “The feasibility of data whitening to improve performance of weather radar,” *The Journal of Applied Meteorology and Climatology*, vol. 38, no. 6, pp. 741-749, Jun. 1999.

[34] J. Devlin, M. W. Chang, and K. Lee *et al.*, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Jun. 2019, pp. 4171-4186.

[35] M. Meilă and H. Y. Zhang, “Manifold learning: what, how, and why,” *Annual Review of Statistics and Its Application*, vol. 11, no. 1, pp. 393-417, Mar. 2024.

[36] S. Chen, X. Bi, and R. Gao *et al.*, “Holistic sentence embeddings for better out-of-distribution detection,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 6676-6686.

[37] J. Li and M. Sun, “Scalable term selection for text categorization,” in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, Jun. 2007, pp. 774-782.

[38] J. Yu, H. Wang, and Y. Zhang *et al.*, “Virtual class matching based detection of out-of-distribution texts,” *Power System Technology*, vol. 49, no. 4, pp. 1681-1688, Apr. 2023. (in Chinese)

[39] J. Guerin, K. Delmas, and R. Ferreira *et al.*, “Out-of-distribution detection is not all you need,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, Jun. 2023, pp. 14829-14837.