Out-of-Distribution Detection of Unknown False Data Injection Attack With Logit-Normalized Bayesian ResNet

Guangxu Feng, *Graduate Student Member, IEEE*, Keng-Weng Lao, *Senior Member, IEEE*, and Ge Chen, *Member, IEEE*

***Abstract*—The progressive integration of cyber-physical systems in smart grids raises potential security concerns, exacerbating the risk of false data injection attack (FDIA) that leads to severe operational disruptions, especially when the FDIA displays profiles that deviate from known attack patterns. Current FDIA detection methods usually operate under the assumption of distributional consistency between training and testing data, thereby falling short in recognizing FDIA with such out-of-distribution (OOD) characteristics. To address this challenge, this paper proposes a novel logit-normalized Bayeisan ResNet (LNBRN) algorithm, a cutting-edge method to address the unexplored OOD FDIA issues efficiently. During offline training, the proposed LNBRN leverages dropout techniques to approximate Bayesian variational inference, thus reducing computational overhead. A key point is the introduction of logit normalization to the output layer, which significantly alleviate the model overconfidence and enhance the follow-up OOD detection performance. During online detection, LNBRN incorporates mutual information to quantify the epistemic uncertainty for incoming measurements, enabling accurate identification of high-risk OOD FDIA events. Comprehensive experimental evaluations on IEEE 14-bus and 118-bus test systems with real load data demonstrate the superiority of detecting OOD FDIA and validate the scalability in larger smart grids.**

***Index Terms*—Bayesian neural network, false data injection attack, out-of-distribution detection, epistemic uncertainty, uncertainty calibration.**

NOMENCLATURE

$\boldsymbol{a}$	Injected false data vector.
x, $\hat{x}$	Original and estimated state vector.
z, $\boldsymbol{z}_a$	Original and attacked measurement vector.
$\boldsymbol{h}(\cdot)$	Nonlinear system measurement mapping.
e	Measurement error vector.
P, Q	Active and reactive power.
V, θ	Voltage magnitude and phase angle.
g, b	Conductance and susceptance.
r	Measurement residual vector.
$\bar{\boldsymbol{z}}$	Unknown out-of-distribution FDIA vector.
$\bar{\boldsymbol{y}}$	Unknown FDIA type.
$\bar{\boldsymbol{c}}$	Unknown state deviation vector.
$\boldsymbol{z}_k$, $\boldsymbol{y}_k$	Data and label of kth known measurement.
D	Training measurement dataset.
$\mathbb{P}_{in}$, $\mathbb{P}_{ood}$	Probability distributions of in-distribution and out-of-distribution measurements.
$\boldsymbol{f}$, $\boldsymbol{w}$	Model and weight of the proposed LNBRN.
α	Temperature factor for output logits.
β	Dropout rate.
$\mathbb{I}$	Mutual information.
T	Times of Monte-Carlo estimation.
λ	Weight decay.
η	Learning rate.

Manuscript received 3 October 2023; revised 10 April 2024; accepted 5 June 2024. Date of publication 18 June 2024; date of current version 23 October 2024. This work was supported in part by the Science and Technology Development Fund, Macau, SAR, under Grant 0003/2020/AKP and Grant 001/2024/SKL, and in part by the Science and Technology Innovation Bureau, Zhuhai, China under Grant EF2023-00092-IOTSC and Grant 2220004002699. Paper no. TSG-01595-2023. *(Corresponding author: Keng-Weng Lao.)*

Guangxu Feng and Keng-Weng Lao are with the State Key Laboratory of Internet of Things for Smart City and the Department of Electrical and Computer Engineering, University of Macau, Macau, China (e-mail: yb97422@um.edu.mo; johnnylao@um.edu.mo).

Ge Chen is with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: chen4911@purdue.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2024.3416164.

Digital Object Identifier 10.1109/TSG.2024.3416164

I. INTRODUCTION

WITH the rapid development of advanced information and communication technologies, traditional power systems have evolved into tightly coupled cyber-physical systems (CPSs) [1]. However, CPSs are increasingly vulnerable to the threats of false data injection attack (FDIA) [2], which may cause serious consequences such as data loss [3], service disruption [4], and even power outages [5]. FDIA disrupts state estimation (SE) covertly by injecting elaborate false data into meter measurements [6]. Recently, massive FDIA detection methods, especially deep learning methods have been raised [7]. These methods generally adhere to the closed-world assumption that training and testing samples of FDIA are both under the same data distribution.

Particularly, there exist specific FDIA with unique patterns, such as intentionally-designed adversarial attacks [8] and FDIA with unforeseen patterns [9]. In real-world applications, it is more prevalent to encounter a new type of FDIA with unforeseen patterns deviating from the known ones. In the context of deep learning, such FDIA with unforeseen patterns exhibits a divergent data distribution from the training data distribution. And the distribution shift here is called out-of-distribution (OOD) issues [10]. OOD issues can be notably

common in regional FDIA owing to the presence of unforeseen target regions subject to attacks [11]. When encountering OOD FDIA event, the aforementioned FDIA deep learning methods mistakenly detect them as known FDIA classes, thus leading to substantial power system failures [12].

To date, there are only a few existing FDIA works handling the distribution change issues [9], [13]. These works aim to handle FDIA with concept shifts. Here, concept shifts refer to the changes over time in the underlying data distribution that misguide the detection model, mainly caused by unforeseen operating conditions [9]. In [13], techniques such as dimensionality reduction and statistical hypothesis are adopted to deal with concept shifts disturbing FDIA detection. Reference [9] proposes a Dynamic Neural Network (DyNN) framework to detect FDIA under concept shifts. But under the circumstances of concept shifts, there is no new type of FDIA emerging. While in the scenario of OOD FDIA, there exists a new class of FDIA. Therefore, these methods still fail to recognize a new type of OOD FDIA. To date, there exists a notable paucity of research focused on considering OOD FDIA issues in smart grids. However, in the broader scope of OOD detection for unknown samples, research has been conducted extensively. Two representative approaches of OOD detection are confidence-based and uncertainty-based methods.

1) Confidence-Based OOD Detection: Prediction confidence is often employed as a criterion to distinguish whether unknown samples are OOD or not. A baseline approach in [14] is introduced to leverage probabilities from softmax distributions of a trained classifier for detecting OOD and misidentified examples. To improve the softmax scores on OOD samples, OOD detector for neural networks (ODIN) is developed based on techniques of temperature scaling and input preprocessing [15]. To calibrate the prediction confidence, GAN is incorporated into the classifier training phase for enhancing OOD detection (GANOOD) [16]. In [17], an additional confidence estimation (CE) branch is introduced into common neural networks for OOD detection. Taking domain information into account, generalized ODIN (GODIN) [18] proposes decomposed confidence based on ODIN.

Nevertheless, confidence-based OOD methods show insufficient robustness to unknown samples and suffer from assigning high overconfidence to incorrect predictions [19]. Moreover, confidence-based methods fail to provide uncertainty information that is vital to ensure reliability under safety-critical FDIA scenarios.

2) Uncertainty-Based OOD Detection: To provide uncertainty estimation for OOD detection, Bayesian neural networks (BNNs) are regarded as a major solution [20]. Bayesian posterior characteristics of BNNs possess inherent robustness to OOD samples [21]. Here, uncertainty is disentangled into aleatoric and epistemic uncertainty. Epistemic uncertainty arising from a lack of training data reflects the inherent model uncertainty and it is especially sensitive to the deviating distribution of OOD samples [22]. Generally speaking, OOD samples are prone to result in high epistemic uncertainty than known samples. But stochastic weights of conventional BNNs double the total parameters compared to common neural networks and cause significant computational expenses. To reduce the computational burden of previous BNNs, it is proved that the common dropout regularization can be utilized to approximate the variational inference process of BNNs [20]. For dropout-based BNNs, uncertainty can be obtained from Monte Carlo estimation. However, common BNNs suffer from the uncertainty overconfidence issues [23]. It remains a barrier to accurate detection on OOD FDIA.

Given the aforementioned tough challenges in detecting OOD FDIA, a novel logit-normalized Bayesian ResNet (LNBRN) is proposed to distinguish OOD FDIA from known classes based on uncertainty estimation. The proposed LNBRN is constructed on a light-weight ResNet for fast training. Here, dropout is utilized to approximate Bayesian variational inference for computation reduction. To mitigate the crucial uncertainty overconfidence issues and improve OOD detection precision further, logit normalization is introduced to keep output norm constant [24]. Then accurate epistemic uncertainty is estimated via mutual information so that OOD FDIA can be recognized by its high uncertainty characteristics. The main contributions of this paper are summarized as below.

1) A novel LNBRN framework is proposed for addressing the unexplored OOD FDIA issues in smart grids. To the best of our knowledge, it is the first model that can detect OOD FDIA.
2) The key employment of logit normalization into Bayesian deep learning contributes to the alleviation of uncertainty overconfidence issues. It enhances the performance of LNBRN during further OOD FDIA detection.
3) The proposed light-weight ResNet structure is designed carefully to help LNBRN achieve the best training cost-efficiency.

Experimental simulations in IEEE 14-bus and 118-bus with real load profiles verify that the proposed LNBRN displays leading superiority during OOD detection and good scalability in larger power grids.

The rest of the paper is organized as follows. Section II displays a thorough literature review. Section III introduces AC state estimation and defines OOD FDIA. In Section IV, the proposed LNBRN for OOD FDIA detection is described in detail. Section V gives the case study including verification under the IEEE test systems. This paper is concluded and summarized in Section VI.

II. Relevant Literature Review

In this section, a detailed introduction of current FDIA detection methods is displayed. Then, a comprehensive comparison of the state-of-the-art and the proposed LNBRN is demonstrated.

A. Brief Review of FDIA Detection

In order to detect stealthy FDIA, a large amount of algorithms have been proposed. In general, there exist two major categories of detection approaches, namely model-based methods and data-driven methods [7].

1) Model-Based Methods: Methods in this category usually incorporate system operation information to detect FDIA and are primarily concerned with enhancing the ability of

TABLE I
COMPARISON WITH CURRENT STATE-OF-THE-ART

Category		Method	References	FDIA	SI	OOD	Precision	Robustness	Overconfidence
Model-based		Variants of Kalman Filter	[25], [26]	Yes	Yes	No	Low	No	None
		Graph Theory	[27]	Yes	Yes	No	Median	No	None
Data-driven	Statistics-based	Distribution Consistency	[30], [31]	Yes	No	No	Median	No	None
	Deep Learning	No scoring	[32], [33], [35], [36],[9]	Yes	No	No	High	No	None
		Scoring by Confidence	[14], [15], [16], [17], [18]	No	No	Yes	High	No	Yes
		Scoring by Uncertainty	BNN [22]	No	No	Yes	High	Yes	Yes
			Proposed LNBRN	Yes	No	Yes	High	Yes	No

system static or dynamic state estimation. Variants of Kalman filter like extended Kalman filter [25] and unscented Kalman filter [26] are utilized to model the nonlinear characteristics of AC SE accurately. In order to detect FDIA with incomplete information inside a certain region, graph theory [27] is applied into modeling power system to identify the outliers in the state estimation results. To mitigate the impact of FDIA, a scheme of multi-agent system is utilized to check the potential threats and a bi-level optimization strategy is employed to model the system risk analysis [28]. However, model-based methods rely heavily on the underlying model accuracy and are sensitive to changes of system parameters. Moreover, calibration and maintenance of model-based methods are quite cost-intensive and time-consuming [29].

2) Data-Driven Methods: To overcome the limitations of model-based methods, numerous data-driven algorithms have been proposed. Methods based on statistical consistency measure the difference between the distributions of measurement variations [30] or measurement residuals [31] where Kullback–Leibler distance and Jensen-Shannon distance are used as metrics respectively. But they are vulnerable to measurement errors and the difficult selection of threshold influences the detection accuracy a lot. Machine learning algorithms, particularly deep learning algorithms, have demonstrated exceptional performance and notable stability. In [32], deep neural network (DNN) is employed to distinguish the system state features extracted by discrete wavelet transform. The effectiveness of convolutional neural network (CNN) has been widely verified. For example, CNN acts as a multi-label classifier to locate FDIA in [33], and in [34], CNN utilizes temporal and spatial correlation features to indicate FDIA, etc. To mitigate the computation burden and stabilize the training, residual network (ResNet) is concatenated with attention LSTM to identify FDIA [35]. Under the circumstance of a few labeled measurement records, a semi-supervised method using generative adversarial network (GAN) is presented to detect unobservable FDIA [36].

B. Comprehensive Comparison With State-of-the-Art

For comparing the current state-of-the-art with proposed LNBRN comprehensively, the attributes below are utilized.

1) The attribute "FDIA" refers to whether the method is originally used to deal with FDIA or not.
2) The attribute "SI" is the abbreviation of system information. It indicates whether power system information is required in the method or not.
3) The attribute "OOD" represents if the method is capable of OOD detection or not.
4) The attribute "Precision" represents the in-distribution detection precision on known FDIA samples.
5) The attribute "Robustness" represents the model performance when faced with different unknown scenarios.
6) The attribute "Overconfidence" indicates if the method has the tendency to assign disproportionately high certainty to its predictions, thus leading to misjudgment.

The comprehensive comparison with the current state-of-the-art is shown in Table I. It can be observed that there is nearly no existing FDIA works considering OOD issues in Table I. And model-based FDIA detection methods fail to the essential detection requirements at all. Then, uncertainty-based OOD detection methods are more robust compared to confidence-based methods among different scenarios. Moreover, proposed LNBRN has an intuitive advantage of addressing overconfidence issues compared to BNN when detecting OOD FDIA.

III. OUT-OF-DISTRIBUTION FALSE DATA INJECTION ATTACK

This section introduces the preliminaries of OOD FDIA. FDIA requires the system and measurement information of SE. Firstly, the background information of AC SE and bad data detection are introduced. Then, FDIA and the relevant OOD issues are described.

A. AC State Estimation and Bad Data Detection

Based on the AC power flow equations, AC static state estimation (SE) utilizes a non-linear function to describe the relationship between the measurement vector $z \in \mathbb{R}^m$ and the state vector $x \in \mathbb{R}^n$:

$$\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{h}(\cdot)$ denotes the nonlinear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ and $e \in \mathbb{R}^m$ represents the measurement error vector. e is assumed to follow a Gaussian distribution that has a zero mean and its covariance matrix $\boldsymbol{R} = diag(\sigma_1^2, \sigma_2{}^2, \ldots, \sigma_m{}^2) \in \mathbb{R}^{m\times m}$ where σ_i is the variance of the ith measurement [31].

To estimate the optimal states $\hat{\boldsymbol{x}}$, the weighted least-squares (WLS) algorithm is widely adopted [37] is widely adopted and the corresponding objective function $\mathcal{J}(\boldsymbol{x})$ is minimized as follows:

$$\widehat{\mathcal{J}} = \min_{\boldsymbol{x}} [\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x})]^T \boldsymbol{R}^{-1} [\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x})]. \tag{2}$$

And (2) is solved by the Gauss–Newton algorithm that updates the states iteratively until the state deviation is lower than a preset threshold.

To detect bad data, the measurement residual $\boldsymbol{r} = \boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x})$ is evaluated. Simply, if the residual norm $\|\boldsymbol{r}\|$ is lower than a preset threshold τ, it is normal. Moreover, for the Chi-square test [38], the WLS objective function value $\widehat{\mathcal{J}}$ in (2) is evaluated. If no suspicious data is detected, it is required to satisfy:

$$\widehat{\mathcal{J}} = \left[\boldsymbol{z} - \boldsymbol{h}(\hat{\boldsymbol{x}})\right]^T \boldsymbol{R}^{-1} \left[\boldsymbol{z} - \boldsymbol{h}(\hat{\boldsymbol{x}})\right] \leq \mathcal{X}^2_{m-n,p} \tag{3}$$

where $\mathcal{X}^2_p$ is the Chi-square distribution with $m-n$ degrees of freedom at the confidence p. The test works well under common scenarios.

B. OOD Issues in Regional FDIA

Typically, the attacker performs a FDIA by injecting the false data $\boldsymbol{a}$ into the measurement data:

$$\boldsymbol{z}_a = \boldsymbol{z} + \boldsymbol{a}. \tag{4}$$

The key point for launching a successful stealthy FDIA is to blind the bad data detection in AC SE and the attack vector $\boldsymbol{a}$ should satisfy the requirement [31]:

$$\boldsymbol{a} = \boldsymbol{h}(\hat{\boldsymbol{x}} + \boldsymbol{c}) - \boldsymbol{h}(\hat{\boldsymbol{x}}). \tag{5}$$

where $\hat{\boldsymbol{x}} = \hat{V}\angle\hat{\theta}$ and $\boldsymbol{c}$ denotes the deviation of the estimated states after FDIA. Here, it is formulated under the assumption that the attacker has obtained full information of the power system including network parameters and access to real-time measurements. The corresponding residual norm is written as

$$\begin{aligned} \|\boldsymbol{r}_a\| &= \|\boldsymbol{z} + \boldsymbol{a} - \boldsymbol{h}(\hat{\boldsymbol{x}} + \boldsymbol{c})\| \\ &= \|\boldsymbol{z} + \boldsymbol{h}(\hat{\boldsymbol{x}} + \boldsymbol{c}) - \boldsymbol{h}(\hat{\boldsymbol{x}}) - \boldsymbol{h}(\hat{\boldsymbol{x}} + \boldsymbol{c})\| \\ &= \|\boldsymbol{z} - \boldsymbol{h}(\hat{\boldsymbol{x}})\| = \|\boldsymbol{r}\|. \end{aligned} \tag{6}$$

Here, a stealthy FDIA causes no change on the residuals so the bad data detection test can be passed easily [31].

In many smart grid applications, some specific FDIAs aim to manipulate the states of partial buses within a certain region of power grids [39]. In this work, the setting of regional FDIA is defined. The bus index set in this certain region is denoted as I. For the corresponding regional FDIA here, the bus sets of manipulated voltage magnitude and phase angle states are denoted as $I^{\hat{V}}, I^{\hat{\theta}} \subseteq I$. And the deviation vectors for voltage magnitude and phase angle are $\Delta\hat{V}$ and $\Delta\hat{\theta}$ respectively. The state deviation in the regional FDIA here is formulated as

$$\boldsymbol{c} = \Delta\hat{V}\angle\Delta\hat{\theta} = \left[\left\{\epsilon^{\hat{V}}{}_i \cdot \hat{V}_i\right\}|_{i\in I^{\hat{V}}}\right] \angle \left[\left\{\epsilon^{\hat{\theta}}{}_j \cdot \hat{\theta}_j\right\}|_{j\in I^{\hat{\theta}}}\right] \tag{7}$$

where $\epsilon^{\hat{V}}{}_i$ and $\epsilon^{\hat{\theta}}{}_j$ are the respective deviation proportions to $\hat{V}_i$ and $\hat{\theta}_j$. For indices outside $I^{\hat{V}}$ and $I^{\hat{\theta}}$, the matching proportions are 0. Then, based on (7), (5) and (4), a regional FDIA is performed by $\boldsymbol{z}_a = \boldsymbol{z} + \boldsymbol{h}(\hat{\boldsymbol{x}} + \boldsymbol{c}) - \boldsymbol{h}(\hat{\boldsymbol{x}})$.

It is common to encounter FDIA with OOD characteristics. A significant challenge in addressing OOD issues is the prevalence of unknown sample types as highlighted by [19]. Given an unknown measurement vector $\bar{\boldsymbol{z}}$, it belongs to an unseen type $\bar{\boldsymbol{y}}$. And the known in-distribution (ID) training dataset is given as $\mathcal{D} = \{(\boldsymbol{z}_k, \boldsymbol{y}_k)\}_{k=1}^N$ where $\boldsymbol{z}_k$ and $\boldsymbol{y}_k$ are the training measurements and indicating labels respectively. As illustrated in Fig. 1, assuming $\boldsymbol{z}_k \sim \mathbb{P}_{in}$ and $\bar{\boldsymbol{z}} \sim \mathbb{P}_{ood}$,

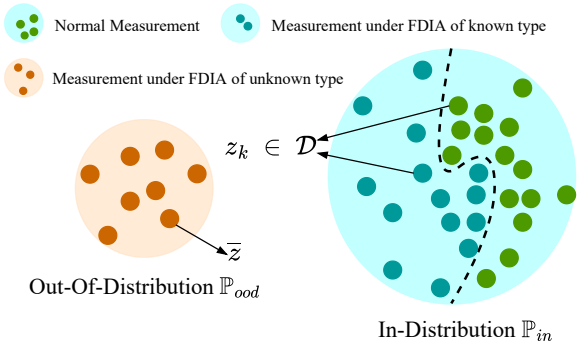


Fig. 1. Illustration about out-of-distribution (OOD) FDIA of unknown type.

$$\mathbb{P}_{ood} \neq \mathbb{P}_{in} \tag{8}$$

that holds for the open-world assumption [19]. It means that OOD FDIA of unknown type $\bar{\boldsymbol{y}}$ has a different data distribution from the training dataset distribution. Moreover, the unknown type $\bar{\boldsymbol{y}}$ is not covered in the training label set.

$$\bar{\boldsymbol{y}} \cap \{\boldsymbol{y}_k)\}_{k=1}^N = \emptyset. \tag{9}$$

Such OOD issues can be notably prevalent in regional FDIA owing to the presence of unforeseen target regions subject to attacks [11]. Notably, the characteristics of a regional FDIA are mainly reflected by the its state deviation $\boldsymbol{c}$. According to (7), regardless of the deviation proportions in our setting, a regional FDIA can be characterized by the bus sets of its manipulated states.

Definition 1: Suppose for an unknown regional FDIA event, $\bar{I}^{\hat{V}}$ and $\bar{I}^{\hat{\theta}}$ are its bus index sets of manipulated voltage magnitude and phase angle states. If it satisfies the following condition

$$\bar{I}^{\hat{V}} \cap \{I_k)\}_{k=1}^N = \emptyset \text{ and } \bar{I}^{\hat{\theta}} \cap \{I_k)\}_{k=1}^N = \emptyset \tag{10}$$

where $\{I_k)\}_{k=1}^N$ denotes the bus index sets of manipulated states among all N measurement samples in the training dataset, then this event is an OOD regional FDIA.

IV. Proposed Logit-Normalized Bayesian ResNet for OOD FDIA

We propose LNBRN to detect OOD FDIA. The overall framework is illustrated in Fig. 2. LNBRN includes three parts matching the three sub-sections below. The first part, Section IV-A, introduces how to approximate the Bayesian variational inference with dropout. The second part, Section IV-B, presents the successive offline training with logit-normalized loss. The last part, Section IV-B, describes how to leverage the trained model for online calculation of epistemic uncertainty for unknown measurements. Finally, we summarize the training and testing pipelines in Section IV-D.

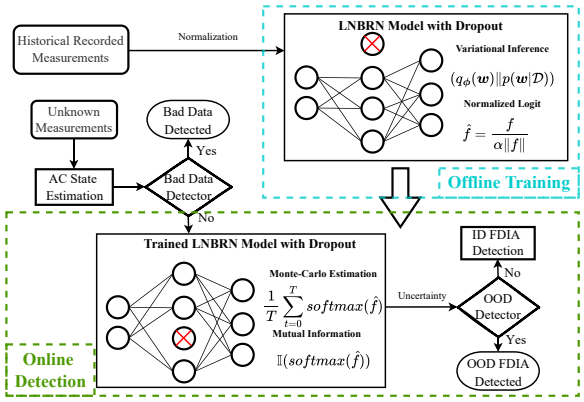


Fig. 2. Proposed framework of OOD FDIA Detection mechanism.

A. Bayesian Variational Inference With Dropout

The classical neural networks utilizes maximum likelihood to obtain the point estimates of parameters so they lack the capacity of uncertainty estimation. By contrast, the proposed LNBRN can provide the posterior distribution of weights.

Given the unknown measurement vector $\bar{\boldsymbol{z}}$, the distribution of its label $\bar{\boldsymbol{y}}$ is written as

$$p(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}}, \mathcal{D}) = \int p(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}}, \boldsymbol{w}) p(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w} \tag{11}$$

where $\boldsymbol{w}$ denotes the trainable parameters of the proposed LNBRN model $\boldsymbol{f}$ and $p(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}}, \boldsymbol{w})$ represents its likelihood function. For FDIA detection here, it is a categorical distribution. Remarkably, $p(\boldsymbol{w}|\mathcal{D})$ is the posterior distribution over the weights under observed $\mathcal{D}$ and it is highly intractable under complicated networks.

To approximate the intractable posterior distribution, variational inference [20] is employed with the utilization of a variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{w})$ where $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ denotes the variational parameters. In order to align the two distributions, the Kullback-Leibler (KL) divergence between them is minimized as the optimization objective:

$$\mathcal{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{w})\|p(\boldsymbol{w}|\mathcal{D})\big) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{w})} \log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{w})}{p(\boldsymbol{w}|\mathcal{D})}. \tag{12}$$

According to Bayes theorem, (12) is reformulated as

$$\mathcal{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{w})\|p(\boldsymbol{w}|\mathcal{D})\big) = -\mathcal{L}_{\boldsymbol{\phi}}(\mathcal{D}) + \log p(\mathcal{D}). \tag{13}$$

$$\mathcal{L}_{\boldsymbol{\phi}}(\mathcal{D}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{w})} \log p(\mathcal{D}|\boldsymbol{w}) - \mathcal{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{w})\|p(\boldsymbol{w})\big). \tag{14}$$

Here, $\mathcal{L}_{\boldsymbol{\phi}}(\mathcal{D})$ is called evidence lower bound (ELBO) that has a lower bound on the *log* marginal likelihood $\log p(\mathcal{D})$. $p(\boldsymbol{w})$ denotes the prior distribution imposed on weights $\boldsymbol{w}$ and $p(\mathcal{D}|\boldsymbol{w})$ describes the distribution of observed training data $\mathcal{D}$ given $\boldsymbol{w}$. As $\log p(\mathcal{D})$ is independent on $\boldsymbol{w}$, minimizing the KL divergence in (12) is equivalent to minimizing the negative ELBO term $-\mathcal{L}_{\boldsymbol{\phi}}(\mathcal{D})$ as follows:

$$\min_{\boldsymbol{\phi}} \mathcal{KL}\big(q_{\boldsymbol{\phi}}(\boldsymbol{w})\|p(\boldsymbol{w}|\mathcal{D})\big) = \min_{\boldsymbol{\phi}} -\mathcal{L}_{\boldsymbol{\phi}}(\mathcal{D}). \tag{15}$$

To derive $q_{\boldsymbol{\phi}}(\boldsymbol{w})$ for Bayesian variational inference, the integration of dropout in neural networks can be formulated as the relevant approximation. Dropout regularization is widely utilized to avoid over-fitting by discarding certain neurons with a Bernoulli distribution $\mathbb{B}$. Given a L-layer network, its weights are written as $\boldsymbol{w} = \{\boldsymbol{\Omega}_i\}_{i=1}^{L}$ and for layer i, the dimensions of $\boldsymbol{w}_i$ are $\mathrm{M_i} \times \mathrm{M_{i-1}}$. Notably, the corresponding $q_{\boldsymbol{\phi}}(\boldsymbol{w}_i)$ for approximation is defined as

$$\begin{aligned} \boldsymbol{\Omega}_i &= \boldsymbol{W}_i \cdot \mathrm{diag}\Big\{\big[s_{i,j}\big]_{j=1}^{\mathrm{M_i}}\Big\} \\ s_{i,j} &\sim \mathbb{B}(\beta_i, \beta_i(1-\beta_i)) \\ &\forall i \in \{1, \ldots, L\},\ j \in \{1, \ldots, \mathrm{M_{i-1}}\} \end{aligned} \tag{16}$$

where $\boldsymbol{\phi} = \{\boldsymbol{W}_i, \beta_i\}_{i=1}^{L}$ acts as the variational parameters and $s_{i,j}$ denotes the binary variable following the distribution $\mathbb{B}(\beta_i)$ where $s_{i,j} = 0$ indicates that unit j in layer $i-1$ linking to layer i has the probability β_i to be abandoned.

By Monte Carlo integration of N samples over $\boldsymbol{w}$, the first integral term in (14) denoted by $\mathcal{L}_{\mathrm{E}}$ can be approximated as:

$$\mathcal{L}_{\mathrm{E}} \approx \sum_{k=1}^{N} \log p\big(\boldsymbol{y}_k|\boldsymbol{z}_k, \boldsymbol{w}_k\big),\ \boldsymbol{w}_k \sim q_{\boldsymbol{\phi}}(\boldsymbol{w}) \tag{17}$$

where N is the total number of training samples. According to [20], the second term in (14) can also be approximated and the objective function in (15) is reformulated as

$$-\mathcal{L}_{\boldsymbol{\phi}}(\mathcal{D}) \propto -\frac{\mathcal{L}_{\mathrm{E}}}{N} + \lambda\left(\sum_{i=1}^{L} \beta_i \|\mathbf{W}_i\|_2^2 + \|\mathbf{b}_i\|_2^2\right) \tag{18}$$

where $\mathbf{b}_i$ is the bias vector for layer i and λ is weight decay for adjusting the regularization term. The target negative ELBO is proportional to the right side that is exactly equal to the objective of dropout network. It can be derived that sampling $\boldsymbol{w} \sim q_{\boldsymbol{\phi}}(\boldsymbol{w})$ is the identical process of dropout acting on the network with weights $\{\boldsymbol{W}_i\}_{i=1}^{L}$. Therefore, Dropout is capable of performing Bayesian variational inference.

B. Training With Logit-Normalized Loss

During in-distribution training, the common softmax cross-entropy $\mathcal{L}_{\mathrm{CE}}$ is utilized as the classification loss. For the measurements $\boldsymbol{z}_k$ and labels $\boldsymbol{y}_k$ with C classes, the corresponding softmax probability can be written as

$$p\big(\boldsymbol{y}_k|\boldsymbol{z}_k, \boldsymbol{w}_k\big) = \frac{e^{\boldsymbol{f}_{y_k}(\boldsymbol{z}_k;\boldsymbol{w}_k)}}{\sum_{i=1}^{\mathrm{C}} e^{\boldsymbol{f}_i(\boldsymbol{z}_k;\boldsymbol{w}_k)}} \tag{19}$$

where $\boldsymbol{f}_{y_k}(\boldsymbol{z}_k; \boldsymbol{w}_k)$ is the exact logit of the network output $\boldsymbol{f}(\boldsymbol{z}_k; \boldsymbol{w}_k)$ matching label $\boldsymbol{y}_k$. After one-hot encoding on known$\boldsymbol{y}_k$, it is derived easily that

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{\mathrm{C}} \boldsymbol{y}_k \log \frac{e^{\boldsymbol{f}_j(\boldsymbol{z}_k;\boldsymbol{w}_k)}}{\sum_{i=1}^{\mathrm{C}} e^{\boldsymbol{f}_i(\boldsymbol{z}_k;\boldsymbol{w}_k)}} = -\frac{\mathcal{L}_{\mathrm{E}}}{N}. \tag{20}$$

So that the loss term $-\frac{\mathcal{L}_{\mathrm{E}}}{N}$ in (18) is equivalent to the softmax cross-entropy $\mathcal{L}_{\mathrm{CE}}$.

As referred in [24], $\mathcal{L}_{\mathrm{CE}}$ enables the magnitude of the output logits to become larger for both ID and OOD samples so that high prediction confidence is produced. And it makes the model prediction miscalibrated. Such miscalibration issues are also observed for low uncertainty estimation under dropout-based Bayesian variational inference [23]. To alleviate such

issues, logit normalization operation is imposed on the output logits to keep the magnitude constant:

$$\hat{\boldsymbol{f}}(\boldsymbol{z}_k;\boldsymbol{w}_k)=\frac{\boldsymbol{f}(\boldsymbol{z}_k;\boldsymbol{w}_k)}{\alpha\|\boldsymbol{f}(\boldsymbol{z}_k;\boldsymbol{w}_k)\|} \tag{21}$$

where α adjusts the logit magnitude. $\alpha\|\boldsymbol{f}(\boldsymbol{z}_k)\|$ plays the role as temperature scaling to restrict high confidence.

Then the logit-normalized cross-entropy loss is given by:

$$\hat{\mathcal{L}}_{\mathrm{CE}}=-\frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{\mathrm{C}}\boldsymbol{y}_k\log\frac{e^{\hat{\boldsymbol{f}}_j(\boldsymbol{z}_k;\boldsymbol{w}_k)}}{\sum_{i=1}^{\mathrm{C}}e^{\hat{\boldsymbol{f}}_i(\boldsymbol{z}_k;\boldsymbol{w}_k)}} \tag{22}$$

The new loss in (22) encourages the model to calibrate the model uncertainty well [23]. Considering (18), the overall loss is written as

$$\mathcal{L}_{\mathrm{total}}=\hat{\mathcal{L}}_{\mathrm{CE}}+\lambda\left(\sum_{i=1}^{L}\beta_i\|\mathbf{W}_i\|_2^2+\|\mathbf{b}_i\|_2^2\right) \tag{23}$$

C. Out-of-Distribution Detection With Epistemic Uncertainty

When it comes to online detection, based on $q_{\boldsymbol{\phi}}(\boldsymbol{w})$ from standard dropout, the predictive distribution for unknown OOD FDIA measurements can be approximated as:

$$q_{\boldsymbol{\phi}}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})=\int p(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}},\boldsymbol{w})q_{\boldsymbol{\phi}}(\boldsymbol{w})d\boldsymbol{w} \tag{24}$$

According to [40], the expectation of the prediction $\bar{\boldsymbol{y}}$ belonging to class i is obtained by T times of Monte Carlo dropout in the forward pass:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})}(\bar{\boldsymbol{y}}=i)\approx\frac{1}{T}\sum_{t=0}^{T}p\big(\bar{\boldsymbol{y}}=i|\bar{\boldsymbol{z}},\boldsymbol{w}^t\big). \tag{25}$$

where $\{\boldsymbol{w}^t\}_{t=1}^{T}$ is sampled from $q_{\boldsymbol{\phi}}(\boldsymbol{w})$ defined in (16). Besides, in the tth forward pass with weight $\boldsymbol{w}^t$, the softmax probability with logit normalization for class i is given by

$$p\big(\bar{\boldsymbol{y}}=i|\bar{\boldsymbol{z}},\boldsymbol{w}^t\big)=\frac{e^{\hat{\boldsymbol{f}}_i(\bar{\boldsymbol{z}};\boldsymbol{w}^t)}}{\sum_{i=1}^{\mathrm{C}}e^{\hat{\boldsymbol{f}}_i(\bar{\boldsymbol{z}};\boldsymbol{w}^t)}}. \tag{26}$$

To measure uncertainty when detecting OOD FDIA, the predictive entropy $\mathbb{H}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})$ is considered:

$$\mathbb{H}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})=-\sum_{i=1}^{\mathrm{C}}q_{\boldsymbol{\phi}}(\bar{\boldsymbol{y}}=i|\bar{\boldsymbol{z}})\log q_{\boldsymbol{\phi}}(\bar{\boldsymbol{y}}=i|\bar{\boldsymbol{z}}). \tag{27}$$

Decomposed from the entropy $\mathbb{H}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})$, the mutual information $\mathbb{I}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})$ is a good measure of epistemic uncertainty [41]:

$$\begin{aligned}\mathbb{I}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})&=\mathbb{H}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})-\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{w})}\big[\mathbb{H}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}},\boldsymbol{w})\big]\\&=-\sum_{i=1}^{\mathrm{C}}\mathbb{E}_{q_{\boldsymbol{\phi}}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})}(\bar{\boldsymbol{y}}=i)\log\mathbb{E}_{q_{\boldsymbol{\phi}}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}})}(\bar{\boldsymbol{y}}=i)\\&\quad+\frac{1}{T}\sum_{t=0}^{T}\sum_{i=1}^{\mathrm{C}}p\big(\bar{\boldsymbol{y}}=i|\bar{\boldsymbol{z}},\boldsymbol{w}^t\big)\log p\big(\bar{\boldsymbol{y}}=i|\bar{\boldsymbol{z}},\boldsymbol{w}^t\big).\end{aligned} \tag{28}$$

The final formulation can be derived by substituting (25) and (26) into (28). For simplicity, it is omitted here.

Based on (28), the scoring function $\mathcal{S}(\bar{\boldsymbol{z}})$ is expressed as

$$\mathcal{S}(\bar{\boldsymbol{z}})=\mathbb{I}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{z}}). \tag{29}$$

Algorithm 1 ID Training of Proposed LNBRN

Input: Historical dataset $\mathcal{D}=\{(\boldsymbol{z}_k,\boldsymbol{y}_k)\}_{k=1}^{N}$, dropout rate $\{\beta_i\}_{i=1}^{L}$, temperature factor α, weight decay λ, learning rate η
Output: Trained LNBRN model

```
Initialize parameters of LNBRN model w
while the model has not converged do
    for epoch ← 1 to n_train do
        Sample z_k, y_k from D
        for layer i = 1, ..., L do
            Sample Ω_i with rate β_i from q_φ(w)
            w ← {Ω_i}_{i=1}^L
        end for
        Calculate logit f̂(z_k; w_k) ← f(z_k;w_k) / (α‖f(z_k;w_k)‖)
        Calculate total loss L_total(f̂(z_k; w_k), y_k, λ)
        Update w ← w + η∇_w[L_total(f̂(z_k; w_k), y_k, λ)]
    end for
end while
```

Algorithm 2 OOD Detection of Proposed LNBRN

Input: Unknown measurements $\bar{\boldsymbol{z}}$, trained LNBRN model with $q_{\boldsymbol{\phi}}(\boldsymbol{w})$, temperature factor α, OOD detector $\mathcal{G}$
Output: Epistemic uncertainty of $\bar{\boldsymbol{z}}$

```
while the measurements pass the Chi-square test do
    for Monte Carlo time t = 1, ..., T do
        Sample w^t from q_φ(w) of trained LNBRN model
        Calculate logit f̂(z̄; w^t) ← f(z̄;w^t) / (α‖f(z̄;w^t)‖)
        Prediction p(ȳ = i|z̄, w^t) ← e^{f̂_i(z̄;w^t)} / Σ_{i=1}^C e^{f̂_i(z̄;w^t)}
    end for
    Calculate epistemic uncertainty for scoring:
        S(z̄) ← I(p(ȳ = i|z̄, w^t))
    if S(z̄) > δ then
        G(z̄; δ, α) ← 1
    else
        G(z̄; δ, α) ← 0
    end if
end while
```

Then the OOD detector $\mathcal{G}(\bar{\boldsymbol{z}};\delta,\alpha)$ for evaluating FDIA measurements is written as

$$\mathcal{G}(\bar{\boldsymbol{z}};\delta,\alpha)=\begin{cases}0 & \text{if } \mathcal{S}(\bar{\boldsymbol{z}})\le\delta\\ 1 & \text{if } \mathcal{S}(\bar{\boldsymbol{z}})>\delta\end{cases} \tag{30}$$

where δ is the detection threshold and α is determined during training. If the score of the detected measurements exceeds δ, it is evaluated as OOD due to its larger uncertainty. Therefore, the potential OOD FDIA can be detected.

D. Training and Testing Pipelines

Algorithm 1 and Algorithm 2 demonstrate the training and testing procedures respectively.

In the phase of offline training, the historical recorded measurements $\{(\boldsymbol{z}_k,\boldsymbol{y}_k)\}_{k=1}^{N}$ are preprocessed and then passed into in-distribution training. Dropout is performed to approximate Bayesian variational inference so that posterior distribution on

weights $\boldsymbol{w}$ is obtained. In the forward pass, the output logit $\boldsymbol{f}(z_k; \boldsymbol{w}_k)$ is normalized to keep constant as α. The total loss $\mathcal{L}_{\text{total}}$ is calculated on the normalized logit $\hat{\boldsymbol{f}}(z_k; \boldsymbol{w}_k)$ for further back propagation. Until the loss is minimized, weight updating is over.

In the phase of online detection, firstly, the unknown measurements $\bar{\boldsymbol{z}}$ are fed into AC state estimation for Chi-square testing. If no bad data is detected, it comes to the OOD evaluation of trained LNBRN model. In this period, dropout works by applying Bernoulli distribution on the trained weights $\boldsymbol{w}$. With T times of Monte-Carlo estimation, the softmax probability $p(\bar{\boldsymbol{y}} = i|\bar{\boldsymbol{z}}, \boldsymbol{w}^t)$ is computed with the normalized logit vector. Then epistemic uncertainty $\mathbb{I}$ is calculated from the T times of softmax probabilities. The OOD detector discriminates the measurements OOD based on scores of epistemic uncertainty. $\bar{\boldsymbol{z}}$ with the score larger than the threshold is recognized as OOD FDIA.

V. Case Study

A. Experiment Setup

1) Data Generation: For verifying the performance, two test systems based on IEEE 14-bus and IEEE 118-bus are established with PYPOWER in *Python 3.8* respectively. The real-world load profiles at 5-minute intervals are collected from New York Independent System Operator (NYISO) via Pecan Street Dataport. For IEEE 14-bus test system, the load data ranges from October 2, 2021 to March 31, 2022. While for IEEE 118-bus test system, the load data is collected between January 1, 2024 and March 6, 2024. Following the instructions in [27], the load profiles of NYISO regions can be linked to load buses through specific combination and then normalized to the default active load of the test systems. Besides, the reactive load is obtained from calculation with the presetting power factor ranging from 0.9 to 0.98. To approach the practical characteristics, the power generation is set to change as the same proportion as the total load varies. Then power flow is operated to get the desired states and construct the relevant measurements below.

Normal Samples: Measurements are constructed from the power flow results directly as shown in (1).

Samples under regional FDIA: Compromised measurements are constructed using (4) and (5) so that they can pass the BDD stealthily. In particular, FDIA can be categorized based on the state deviations [42]. For IEEE 14-bus, as shown in Fig. 3, according to regional information, three types of FDIA events are defined for the convenience of out-of-distribution analysis in this paper. For constructing one certain type of FDIA here, it is defined that state variables $\boldsymbol{x}$ of the buses in the corresponding region are manipulated at random with the deviation $\boldsymbol{c}$ by FDIA from the attacker. Particularly, detailed elaboration of three FDIA events is displayed below.

FDIA Event 1: there exist certain deviations among states of bus 4, 7, 8, 9 and 10. The attacked buses are chosen randomly and the ratio of state deviation $\boldsymbol{c}$ varies between 20% and 30%.

FDIA Event 2: there exist certain deviations among states of bus 2, 3, 5 and 6. The attacked buses are chosen at random and the ratio of state deviation $\boldsymbol{c}$ changes between 20% and 30%.

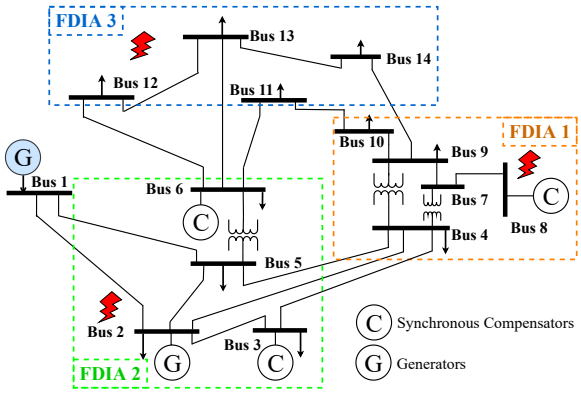


Fig. 3. Proposed scheme of regional FDIA events in IEEE 14-bus system.

FDIA Event 3: there exist certain deviations among states of bus 11, 12, 13, and 14. The attacked buses are chosen randomly and the ratio of state deviation $\boldsymbol{c}$ varies between 20% and 30%.

Similarly, for IEEE 118-bus, seven types of FDIA events are defined according to regional information for evaluating the scalability of proposed OOD detection framework. For a detailed explanation of the seven regional FDIA events, please refer to the Appendix.

All measurements are preprocessed with normalization. For in-distribution training, it involves normal measurements and measurements under known FDIA. Besides, the train-test-ratio is 7:3. Accordingly, measurements under unknown FDIA are utilized for OOD evaluation with the same ratio as each type of known FDIA.

2) Neural Network Configuration: The proposed LNBRN is built on the pre-activation ResNet structure and detailed parameters are shown in Table II. In this architecture, Batch Normalization (BN) and ReLU activation are added before the convolution layer. And then follows the dropout operation. It consists of feature extractor and classifier. Feature extractor is composed of 1 convolution block, 3 residual blocks, 3 downsampling (DS) blocks for feature reduction and 1 pooling block. In the convolution block, the filters are 128 and kernel size is 3. One residual block contains 3 convolution blocks where the kernel sizes are 1, 3 and 1. In Table II, the omitted identity layers for connection skipping are used in ResNet to addressing model degradation [43]. In the DS residual blocks, the strides of 3 convolution layers are 2, 1 and 1 respectively. Here, the DS convolution blocks work as residual connections where the kernel size is 1 and strides are 2. In the pooling block, pool size of the average pooling layer is 4 and follows by a global average pooling layer. Classifier includes one fully-connected layer with 3 units and Softmax activation one normalization layer follows.

During ID training, mini-batch size is 80 and weight decay $\lambda = 0.0005$. Total epochs are 300 and Adam optimizer is adopted. The learning rate η is updated by $\eta = \eta/2$ with initial value 0.001 when loss has no decreasing for 4 epochs.

TABLE II
PARAMETER SETTING OF PROPOSED LNBRN

Blocks	Layers	Parameters
Feature Extractor		
1 convolution block	1 (BN + ReLU + convolution + dropout) layer	filters: 128, kernel size: 3, strides: 1, padding: 'same'
1 residual block	3 (BN + ReLU + convolution + dropout) layers	filters: 32/32/128, kernel size:1/3/1, strides: 1, padding: 'same'
1 DS residual block	3 (BN + ReLU + convolution + dropout) layers 1 (BN + ReLU + DS convolution + dropout) layer	filters: 64/64/256, kernel size: 1/3/1, strides:2/1/1, padding: 'same' filters: 256, kernel size: 1, strides: 2, padding: 'same'
1 residual block	3 (BN + ReLU + convolution + dropout) layers	filters: 64/64/256, kernel size: 1/3/1, strides: 1, padding: 'same'
1 DS residual block	3 (BN + ReLU + convolution + dropout) layers 1 (BN + ReLU + DS convolution + dropout) layer	filters: 128/128/512, kernel size: 1/3/1, strides: 2/1/1, padding: 'same' filters: 512, kernel size: 1, strides: 2, padding: 'same'
1 residual block	3 (BN + ReLU + convolution + dropout) layers	filters: 128/128/512, kernel size: 1/3/1, strides: 1, padding: 'same'
1 DS residual block	3 (BN + ReLU + convolution + dropout) layers 1 (BN + ReLU + DS convolution + dropout) layers	filters: 256/256/1024, kernel size: 1/3/1, strides: 2/1/1, padding: 'same' filters: 1024, kernel size: 1, strides: 2, padding: 'same'
1 pooling block	1 average pooling layer 1 global average pooling layer	pool size: 4 -
Classifier		
1 dense block	1 fully-connected layer 1 normalization layer	units: 3, Softmax, bias -

The proposed LNBRN is constructed on Tensorflow Keras 2.4. Related experiments are simulated in DELL Precision Tower 7920 with Intel Xeon Gold 6140 2.30GHz CPU, Quadro P6000 GPU, and Ubuntu Linux 18.04 LTS system.

B. Evaluation Metrics

To evaluate the effectiveness of LNBRN, recognized metrics are employed [15]. During the ID training, accuracy and confusion matrix are considered. For quantification during the OOD detection, metrics of binary confusion matrix are used as follows. Here, OOD FDIA and ID measurements are set as negative and positive examples respectively.

1) FPR(TPR95) denotes the ratio of OOD FDIA erroneously detected as known ID measurements (false positive rate, FPR) when the correct detection rate of ID measurements (true positive rate, TPR) is 95%.
2) FNR(TNR95) is the proportion of ID measurements that are miscategorized as OOD (false negative rate, FNR) as the right categorized rate of OOD FDIA (true negative rate, TNR) is equal to 95%.
3) D-Error measures the detection error probability of both ID and OOD measurements when TPR is 95%.
4) AUROC evaluates the area under the receiver operating characteristic curve. It describes the change of TPR as FPR varies regardless of the detection threshold.
5) AUPR evaluates the area under the precision-recall curve. Precision Recall is identical to TPR. It is useful for the imbalanced dataset.

For FPR(TPR95), FNR(TNR95) and D-Error, the lower values indicate the better performance. While the higher AUROC and AUPR are, the better OOD detection is.

Overlap coefficients measure the extent of commonality between two data distributions, thereby serving as a metric to demonstrate the separation ability intuitively in the density histograms here. Gaussian kernel density functions are employed to estimate the values of overlap coefficients [44]. The smaller overlap value indicates the better separation ability. Moreover, t-distributed Stochastic Neighbor Embedding (t-SNE) is utilized here for feature visualization [45].

C. Introduction of Overall Verification

In IEEE 14-bus, 8000 normal measurements are generated and for each type of the aforementioned regional FDIA, 4000 samples are generated. Here, three OOD scenarios are constructed on the basis of three regional FDIA events.

Scenario 1: FDIA 1 is unknown with no record before while known records include the other events. In other words, ODD samples are FDIA 1 and ID samples are FDIA 2, FDIA 3 and normal samples.

Scenario 2: Similarly, ODD samples are FDIA 2 here and ID samples include FDIA 1, 3 and normal samples.

Scenario 3: ODD samples are FDIA 3 and ID samples consist of FDIA 1, 2 and normal samples.

In each OOD scenario of IEEE 14-bus, two types of FDIA and normal samples are used for ID training. Here, the ratio of normal samples to total FDIA samples is 1:1 for a balanced setting.

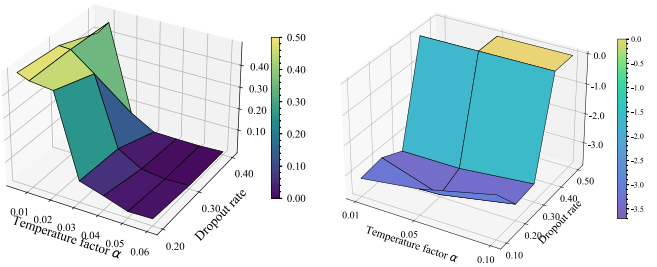
In the context of IEEE 118-bus, 1000 samples are generated for each type of regional FDIA respectively. Then, 12,000 samples are generated for normal measurements. With seven regional FDIA events, seven OOD scenarios are constructed. The thorough elaboration of seven OOD FDIA scenarios and the relevant regional FDIA events is provided in the Appendix. In particular, in each OOD scenario of IEEE 118-bus, only six types of FDIA events and normal measurements are involved for training so the ratio of normal samples to total FDIA samples is 2:1 for an imbalanced setting.

Then the overall evaluation is divided into three parts. The first part reveals the ID training performance of proposed LNBRN and the second part verifies the superiority of LNBRN on OOD detection in IEEE 14-bus. The last part evaluates the scalability of LNBRN or deployment in larger smart grids like IEEE 118-bus.

D. Verification of Proposed LNBRN on ID Training

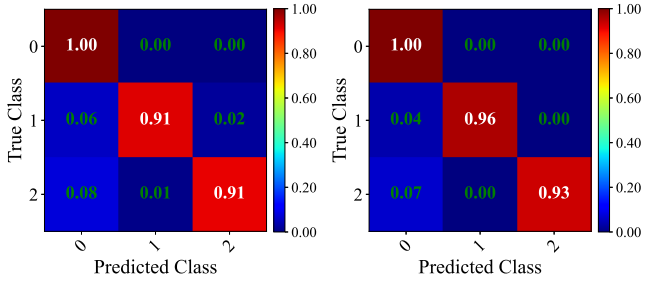
For comparison of ID training, two typical Bayesian structures are selected below.

1) *Dropout BCNN* [22] is dropout-based Bayesian convolutional network. It follows the VGG-16 CNN structure but



(a) (b)

Fig. 4. FNR(TNR95) of LNBRN under different α values and dropout rates. (a) Scenario 3 of IEEE 14-bus. (b) Scenario 1 of IEEE 118-bus.



(a) (b)

Fig. 5. Accuracy analysis with confusion matrix under Scenario 3. (a) Flipout BRN. (b) Proposed LNBRN.

TABLE III
COMPARISON OF DIFFERENT BAYESIAN NEURAL NETWORK STRUCTURES

Structure	Parameters	Training Time/s	Accuracy
Proposed LNBRN	1,724,359	1395.06	0.9721±0.0001
Dropout BCNN	69,210,691	2147.88	0.9614±0.0003
Flipout BRN	3,434,311	2694.03	0.9553±0.0011

has 4 blocks with 2 convolution layers each. Pool size is 3 and dense units are 1024, 512 and 3.

2) *Flipout BRN* [46] is Bayesian ResNet based on variational Flipout estimator. It has the same ResNet structure as LNBRN but double the total parameters.

To tune crucial hyperparameters such as temperature factor α and dropout rate, grid search is adopted for LNBRN to enhance performance robustness. In IEEE 14-bus, taking Scenario 3 as an example, FNR(TNR95) values of LNBRN under different α and dropout rates are shown in Fig. 4(a). It is seen that when dropout rate β is 0.4 and α is about 0.05, FNR(TNR95) has the lowest value. In particularly, the settings of $\beta = 0.4$ and $\alpha = 0.05$ are kept for all three scenarios in IEEE 14-bus. Similarly, in IEEE 118-bus, Fig. 4(b) shows that the lowest FNR(TNR95) under Scenario 1 is achieved at the point of $\beta = 0.2$ and $\alpha = 0.05$. It means that LNBRN reaches the minimum error rate for ID samples under these hyperparameter settings.

Moreover, Table III demonstrates its ID training performance of different Bayesian neural network structures in Scenario 3. The proposed method achieves the best training results. Compared to Flipout BRN, dropout-based Bayesian variational inference in LNBRN reduces the training parameters by half and decreases the training time by 48.22%. Compared to Dropout BCNN, the proposed ResNet structure obtains the better classification accuracy using only one fortieth of the parameters. Detailed accuracy analysis with confusion matrix is shown in Fig. 5. Therein, class 0, 1, 2 denote normal samples, FDIA 1, and FDIA 2 respectively.

Both Flipout BRN and LNBRN perform greatly on recognizing normal samples (label 0). Compared to Flipout BRN, LNBRN has a higher classification accuracy on FDIA 1 and FDIA 2, i.e., it improves the accuracy by 5.49% on FDIA 1. In general, the proposed LNBRN shows the most cost-efficient performance during ID training.

TABLE IV
DETECTION PERFORMANCE OF DIFFERENT METHODS FOR THREE OUT-OF-DISTRIBUTION SCENARIOS

Scenario	Method	D-Error↓	FPR(TPR95)↓	FNR(TNR95)↓	AUROC↑	AUPRC↑
Scenario 1	Baseline	0.4717	0.8933	1.0000	0.4049	0.7671
	GANOOD	0.4442	0.8383	0.9896	0.4901	0.8023
	ODIN	0.4517	0.8533	1.0000	0.5490	0.8161
	GODIN	0.4467	0.8433	0.8692	0.6731	0.8779
	CE	0.3750	0.7000	0.9090	0.6924	0.8722
	BRN	0.0892	0.1283	0.3569	0.9542	0.9843
	Proposed	**0.0250**	**0.0000**	**0.0002**	**1.0000**	**1.0000**
Scenario 2	Baseline	0.4200	0.7900	0.7013	0.5671	0.8537
	GANOOD	0.4275	0.8050	1.0000	0.5177	0.7915
	ODIN	0.4108	0.7717	1.0000	0.5845	0.8279
	GODIN	0.3858	0.7217	0.6963	0.7881	0.9282
	CE	0.3608	0.7167	0.9271	0.6734	0.8602
	BRN	0.0621	0.0742	0.0681	0.9894	0.9973
	Proposed	**0.0254**	**0.0008**	**0.0321**	**0.9969**	**0.9992**
Scenario 3	Baseline	0.4096	0.7692	1.0000	0.5031	0.7987
	GANOOD	0.4579	0.8658	0.8256	0.4852	0.8185
	ODIN	0.3808	0.7116	1.0000	0.5518	0.8158
	GODIN	0.3908	0.7317	0.7396	0.7884	0.9319
	CE	0.3992	0.7483	0.7850	0.7378	0.9071
	BRN	0.1313	0.2125	0.6479	0.9179	0.9704
	Proposed	**0.0250**	**0.0000**	**0.0015**	**0.9996**	**0.9999**

E. Verification of Proposed LNBRN on OOD Detection

For verification of LNBRN on OOD detection, six state-of-the-art algorithms for comparison are introduced below.

1) *Baseline* [14] uses maximum softmax probabilities to distinguish OOD samples from ID ones.
2) *GANOOD* [16] improves Baseline with temperature scaling and input preprocessing.
3) *ODIN* [15] improves Baseline with temperature scaling and input preprocessing.
4) *GODIN* [18] utilizes the decomposed confidence scoring based on ODIN.
5) *CE* [17] introduces confidence learning by providing the network with hints during training.
6) *BNN* [22] is based on dropout Bayesian variational inference to model uncertainty for OOD detection.

All benchmark algorithms are constructed based on the same ResNet structure as the proposed LNBRN. For detection of Baseline and ODIN, OOD data has lower maximum softmax probabilities than ID ones. And for GODIN and CE, it is expected that OOD data has lower confidence. While for BNN and proposed LNBRN, OOD FDIA tends to cause larger epistemic uncertainty than ID samples.

Table IV shows the OOD detection results of the seven methods. It can be observed that the proposed method performs best for all metrics under the three scenarios. The proposed LNBRN has the highest AUROC and AUPRC, and also the lowest error rate. Especially, FPR(TPR95) is reduced

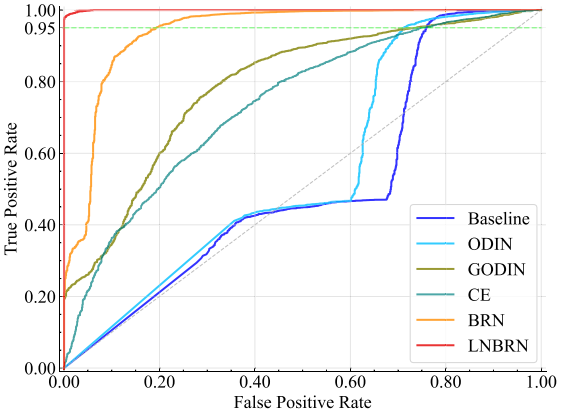


Fig. 6. ROC Curves of different OOD detection methods under Scenario 3.

close to 0 and it means that the proposed method hardly makes errors on detecting OOD FDIA. In contrast to the other methods under all 3 scenarios, LNBRN decreases the error rate by at least 59.10% in D-error, 52.86% in FNR(TNR95) and 98.92% in FPR(TPR95). For instance, LNBRN reduces the D-error by 94.18% on average compared to Baseline and 80.96% under Scenario 3 compared to BRN that also uses uncertainty. Besides, AUROC and AUPRC of LNBRN approach the ideal value 1 that means the proposed method has an excellent performance of distinguishing OOD FDIA from ID samples. And when Scenario 3 happens, AUROC of LNBRN is increased by 98.69% compared to Baseline and 8.90% compared to BRN. The related ROC curves under Scenario 3 are displayed in Fig. 6. The red curve of LNBRN is almost coincident to the horizontal line with value 1 and it indicates the near-perfect discrimination. When TPR is 95% (green dash in Fig. 6), FPR is reduced by around 0.2 from BRN to LNBRN that verifies the reliability under such safety-critical scenarios.

To demonstrate the separation capacity between ID samples and OOD FDIA, histograms that describes the density distributions of detection scores are shown in Fig. 7. Three types of scores are mentioned here and the smaller overlap of ID samples and OOD FDIA indicates the better separation capacity. Under Scenario 3, it is observed that the proposed LNBRN displays the best separation performance with the lowest overlap coefficient. Although GODIN and CE based on confidence scoring reduce the detection error compared to Baseline, their overlap values are increased by 81.54% and 91.05%. It means that GODIN and CE exhibit weak efficacy in terms of separation capacity. While for the uncertainty-based methods in Figs. 7e and 7f, the overlap of LNBRN is significantly reduced by 74.03% compared to that of BRN. Above all, the proposed LNBRN possesses the best separation capacity than any other.

For presenting the efficiency of logit-normalization-based uncertainty calibration on feature representation, 2-D t-SNE visualization of feature extracted by BRN and LNBRN under Scenario 3 is displayed in Fig. 8 as an example. The lower two figures show features of LNBRN on ID samples and OOD FDIA while the upper two are for BRN. Fig. 8c shows ground truth features where red points denote OOD FDIA

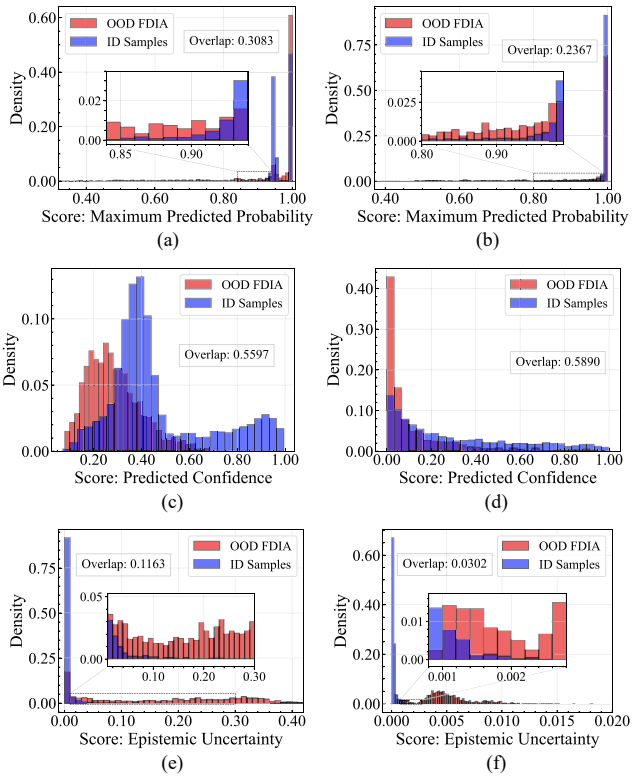


Fig. 7. Histograms of separation performance between ID samples and OOD FDIA by different OOD methods under Scenario 3. (a) Baseline. (b) ODIN. (c) GODIN. (d) CE. (e) BRN. (f) LNBRN.

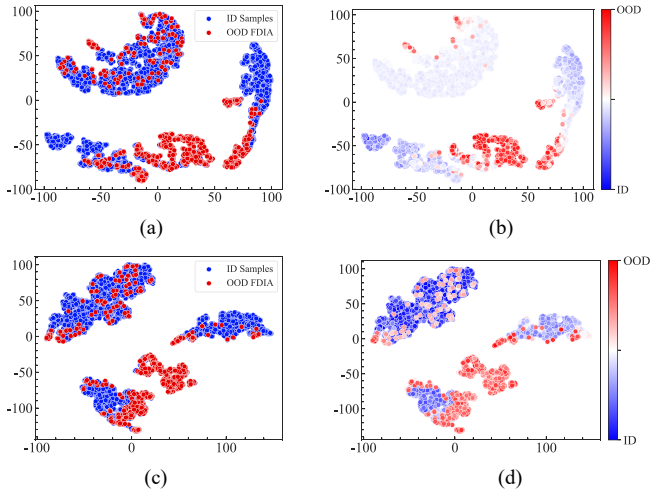


Fig. 8. Feature Visualization by t-SNE of BRN and LNBRN under Scenario 3. (a) Ground truth of BRN. (b) Predicted by epistemic uncertainty of BRN. (c) Ground truth of LNBRN. (d) Predicted by epistemic uncertainty of LNBRN.

and blue points represent ID samples. Fig. 8d shows features predicted by epistemic uncertainty where points with larger uncertainty are assigned the redder color and ones with lower uncertainty have the bluer color. The diverging center of the colorbar is the detection threshold when TPR is 95%. It is the same for BRN in Fig. 8(a) and Fig. 8(b). It can be seen that the features predicted by epistemic uncertainty of LNBRN align better with the matching ground truth in color than BRN. The feature boundary of ID and OOD FDIA predicted

TABLE V
DETECTION PERFORMANCE OF DIFFERENT METHODS FOR THREE OUT-OF-DISTRIBUTION SCENARIOS IN IEEE 118-BUS

Scenario	Method	D-Error↓	FPR(TPR95)↓	FNR(TNR95)↓	AUROC↑	AUPRC↑
Scenario 1	Baseline	0.2267	0.4033	0.5280	0.8913	0.9922
	ODIN	0.2330	0.4160	0.3867	0.8858	0.9673
	GODIN	0.2085	0.3670	0.3735	0.9131	0.9787
	CE	0.5117	0.9733	0.3348	0.7576	0.9848
	BRNN	0.0383	0.0267	0.0419	0.9859	0.9992
	LNBRN	**0.0250**	**0.0000**	**0.0002**	**1.0000**	**1.0000**
Scenario 2	Baseline	0.3967	0.7433	0.8219	0.5567	0.9588
	ODIN	0.3020	0.5540	0.6167	0.8631	0.9656
	GODIN	0.3115	0.5730	0.9028	0.6571	0.8878
	CE	0.3195	0.5890	0.9012	0.6382	0.9474
	BRN	0.0533	0.0567	0.1126	0.9718	0.9975
	LNBRN	**0.0250**	**0.0000**	**0.0002**	**1.0000**	**1.0000**
Scenario 3	Baseline	0.0967	0.1433	0.2361	0.9582	0.9970
	ODIN	0.0910	0.1320	0.1667	0.9523	0.9849
	GODIN	0.3410	0.6320	0.9481	0.5531	0.8444
	CE	0.5100	0.9700	0.8427	0.6506	0.9691
	BRN	0.0550	0.0600	0.1146	0.9756	0.9981
	LNBRN	**0.0250**	**0.0000**	**0.0026**	**0.9998**	**1.0000**
Scenario 4	Baseline	0.0850	0.1200	0.1537	0.9704	0.9980
	ODIN	0.0725	0.0950	0.1435	0.9729	0.9937
	GODIN	0.3190	0.5880	0.9372	0.6816	0.8857
	CE	0.4650	0.8800	0.5967	0.8615	0.9862
	BRN	0.0483	0.0467	0.0504	0.9817	0.9986
	LNBRN	**0.0250**	**0.0000**	**0.0002**	**0.9999**	**1.0000**
Scenario 5	Baseline	0.0950	0.1400	0.3824	0.9523	0.9961
	ODIN	0.0945	0.1390	0.3357	0.9532	0.9864
	GODIN	0.4075	0.7650	0.9607	0.6866	0.8915
	CE	0.3633	0.6767	0.8780	0.7870	0.9748
	BRN	0.0567	0.0633	0.1417	0.9690	0.9968
	LNBRN	**0.0250**	**0.0000**	**0.0000**	**1.0000**	**1.0000**
Scenario 6	Baseline	0.1467	0.2433	0.4372	0.9355	0.9956
	ODIN	0.1000	0.1500	0.2726	0.9538	0.9861
	GODIN	0.3160	0.5820	0.7767	0.7466	0.9323
	CE	0.5233	0.9967	0.3743	0.7048	0.9804
	BRN	0.0583	0.0667	0.1078	0.9701	0.9974
	LNBRN	**0.0250**	**0.0000**	**0.0008**	**0.9999**	**1.0000**
Scenario 7	Baseline	0.0900	0.1300	0.2063	0.9661	0.9978
	ODIN	0.0865	0.1230	0.1850	0.9672	0.9918
	GODIN	0.2585	0.4670	0.6724	0.8375	0.9601
	CE	0.5217	0.9933	0.7763	0.2934	0.9369
	BRN	0.0417	0.0333	0.0356	0.9772	0.9981
	LNBRN	**0.0250**	**0.0000**	**0.0002**	**0.9999**	**1.0000**

by LNBRN is more clear and precise in Fig. 8(d). While for BRN in Fig. 8(b), some feature points of ID samples and OOD FDIA almost have the same color that is related to uncertainty miscalibration issues. By logit normalization, epistemic uncertainty values of LNBRN correspond better to the feature points of ID samples and OOD FDIA than BRN so that the uncertainty miscalibration issues are alleviated. It reveals that LNBRN achieves better uncertainty calibration on feature representation with logit normalization.

F. Scalability Evaluation in Large-Scale Smart Grids

The comprehensive comparison of the aforementioned six methods and proposed LNBRN is demonstrated in Table IX.

As shown in Table IX, the proposed LNBRN consistently outperforms competing models across all metrics in each of the seven scenarios considered. Specifically, the AUROC and AUPRC for the proposed LNBRN are equal to the maximum possible value of 1 in the majority of scenarios, while the FPR(TPR95) metric reaches the minimum value of 0 and the detection error D-Error achieves the lowest rate 0.0250 compared to other methods. Relative to the second-ranked BRN, the LNBRN reduces the error rate by 49.10% in D-error, 99.34% in FNR(TNR95), and 100% in FPR(TPR95) on average.

In the context of AUROC for Scenario 1, the proposed LNBRN demonstrates a 32.00% improvement over the least performing CE method. The ROC curves for all methods under Scenario 1, presented in Fig. 9, further corroborate these findings. Notably, the LNBRN ROC curve, marked in red, aligns more closely with the boundary lines, emphasizing its

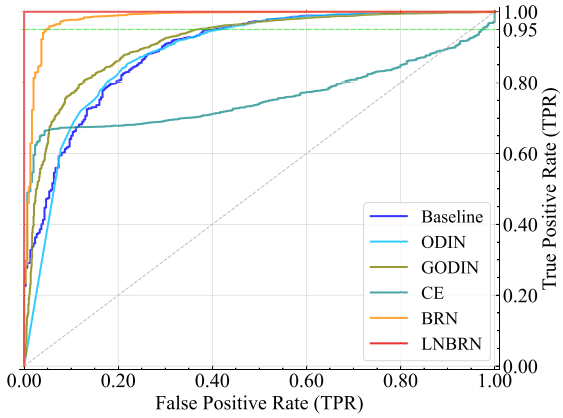


Fig. 9. ROC Curves of different OOD detection methods under Scenario 1 of IEEE 118-bus.

superior discrimination capability. Moreover, as indicated by the green dashed line in Fig. 9, representing a true positive rate (TPR) of 95%, it is apparent that the LNBRN achieves a false positive rate (FPR) of zero. This observation underscores the robustness of the LNBRN model in avoiding the misclassification of OOD FDIA as known classes. Collectively, these results suggest that the LNBRN exhibits commendable scalability for deployment in large-scale smart grids.

VI. CONCLUSION

In this paper, a novel LNBRN algorithm is proposed to handle the unexplored OOD FDIA problems in smart grids based on uncertainty estimation. During ID training with known measurements, dropout is utilized to approximate the costly Bayesian variational inference, thus reducing the computation burden. Moreover, the pivotal employment of logit normalization keeps the model output norm constant so that the tricky model overconfidence issues can be alleviated. When it comes to online OOD detection, uncertainty information of unknown measurements is estimated by mutual information. And then the detector indicates OOD FDIA based on the uncertainty scores. Experimental evidence in IEEE 14-bus demonstrates that LNBRN outperforms six state-of-the-art methods remarkably. While training at low computation cost, LNBRN significantly improves the OOD detection performance by reducing the detection error by an average of 94.18% compared to Baseline. Besides, LNBRN demonstrates the superior ability to distinguish between known measurements and OOD FDIA, as evidenced by the lowest overlap coefficient. The efficiency of uncertainty calibration on feature representation is ensured by logit normalization. As the OOD pattern of FDIA is reflected as spatial relationship in this work, in the future, it is meaningful to investigate how temporal OOD patterns like seasonality and long-term trends influence the relevant FDIA detection.

APPENDIX
PROPOSED OOD SCENARIOS IN IEEE 118-BUS

FDIA 1: The attacked buses are chosen randomly from {1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 14, 15, 16, 117}. The state deviation ranges from 0 to 10%.

FDIA 2: The attacked buses are chosen randomly from {8, 9, 10, 17, 18, 19, 20, 21, 22,23, 25, 26,27, 28, 29, 30, 31, 32, 113, 114, 115 }. The state deviation ranges from 0 to 10%.

FDIA 3: The attacked buses are chosen randomly from {33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49}. The state deviation ranges from 0 to 10%.

FDIA 4: The attacked buses are chosen randomly from {50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67}. The state deviation ranges from 0 to 10%.

FDIA 5: The attacked buses are chosen randomly from {24, 38, 68, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 97, 98, 99, 116, 118}. The state deviation ranges from 0 to 10%.

FDIA 6: The attacked buses are chosen randomly from {82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96}. The state deviation ranges from 0 to 10%.

FDIA 7 The attacked buses are chosen randomly from {100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112}. The state deviation ranges from 0 to 10%.

For evaluation in IEEE 118-bus test system, seven OOD scenarios are constructed with the regional FDIA events above.

Scenario 1: FDIA 1 is unknown with no record before while known records include the other events. In other words, ODD samples are FDIA 1 and ID samples are FDIA 2, 3, 4, 5, 6, 7 and normal samples.

Scenario 2: Similarly, ODD samples are FDIA 2 here and ID samples include FDIA 1, 3, 4, 5, 6, 7 and normal samples.

Scenario 3: ODD samples are FDIA 3 and ID samples consist of FDIA 1, 2, 4, 5, 6, 7 and normal samples.

Scenario 4: ODD samples are FDIA 4 and ID samples are composed of FDIA 1, 2, 3, 5, 6, 7 and normal samples.

Scenario 5: ODD samples are FDIA 5 and ID samples consist of FDIA 1, 2, 3, 4, 6, 7 and normal samples.

Scenario 6: ODD samples are FDIA 5 and ID samples include FDIA 1, 2, 3, 4, 5, 7 and normal samples.

Scenario 7: ODD samples are FDIA 5 and ID samples comprise of FDIA 1, 2, 3, 4, 5, 6 and normal samples.

For each type of FDIA, 1000 samples are generated respectively; for normal measurements, 12,000 samples are produced. The ratio of normal samples to total FDIA samples during training is maintained at 2:1, aligning with the possible proportion requirement encountered in real-world settings.

REFERENCES

[1] S. Yang, K.-W. Lao, H. Hui, and Y. Chen, "Secure distributed control for demand response in power systems against deception cyber-attacks with arbitrary patterns," *IEEE Trans. Power Syst.*, early access, Mar. 25, 2024, doi: 10.1109/TPWRS.2024.3381231.

[2] H. T. Reda, A. Anwar, and A. Mahmood, "Comprehensive survey and taxonomies of false data injection attacks in smart grids: Attack models, targets, and impacts," *Renew. Sustain. Energy Rev.*, vol. 163, Jul. 2022, Art. no. 112423.

[3] G. Feng and K.-W. Lao, "Wasserstein adversarial learning for identification of power quality disturbances with incomplete data," *IEEE Trans. Ind. Informat.*, vol. 19, no. 10, pp. 10401–10411, Oct. 2023.

[4] K. Wang, M. Du, S. Maharjan, and Y. Sun, "Strategic honeypot game model for distributed denial of service attacks in the smart grid," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2474–2482, Sep. 2017.

[5] L. Che, X. Liu, Z. Li, and Y. Wen, "False data injection attacks induced sequential outages in power systems," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1513–1523, Mar. 2019.

[6] J. Zhao, G. Zhang, M. La Scala, Z. Y. Dong, C. Chen, and J. Wang, "Short-term state forecasting-aided method for detection of smart grid general false data injection attacks," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1580–1590, Jul. 2017.

[7] A. S. Musleh, G. Chen, and Z. Y. Dong, "A survey on the detection algorithms for false data injection attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2218–2234, May 2020.

[8] R. Huang and Y. Li, "Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system," *IEEE Trans. Smart Grid*, vol. 14, no. 3, pp. 2367–2376, May 2023.

[9] D. Hu, S. Wu, J. Wang, and D. Shi, "Training a dynamic neural network to detect false data injection attacks under multiple unforeseen operating conditions," *IEEE Trans. Smart Grid*, vol. 15, no. 3, pp. 3248–3261, May 2024.

[10] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. 32nd Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[11] R. Jiao, G. Xun, X. Liu, and G. Yan, "A new AC false data injection attack method without network information," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5280–5289, Nov. 2021.

[12] H. Goyel and K. S. Swarup, "Data integrity attack detection using ensemble-based learning for cyber–physical power systems," *IEEE Trans. Smart Grid*, vol. 14, no. 2, pp. 1198–1209, Mar. 2022.

[13] M. Mohammadpourfard, Y. Weng, M. Pechenizkiy, M. Tajdinian, and B. Mohammadi-Ivatloo, "Ensuring cybersecurity of smart grid against data integrity attacks under concept drift," *Int. J. Electr. Power Energy Syst.*, vol. 119, Jul. 2020, Art. no. 105947.

[14] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–12.

[15] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15.

[16] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16. [Online]. Available: https://openreview.net/forum?id=ryiAv2xAZ

[17] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," 2018, *arXiv:1802.04865*.

[18] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10951–10960.

[19] J. Yang et al., "OpenOOD: Benchmarking generalized out-of-distribution detection," in *Proc. 36th Conf. Neural Inf. Process. Syst.*, 2022, pp. 1–14.

[20] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[21] G. Carbone, M. Wicker, L. Laurenti, A. Patane, L. Bortolussi, and G. Sanguinetti, "Robustness of Bayesian neural networks to gradient-based attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 15602–15613.

[22] A. T. Nguyen, F. Lu, G. L. Munoz, E. Raff, C. Nicholas, and J. Holt, "Out of distribution data detection using dropout Bayesian neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 7877–7885.

[23] M.-H. Laves, S. Ihler, K.-P. Kortmann, and T. Ortmaier, "Well-calibrated model uncertainty with temperature scaling for dropout variational inference," in *Proc. 4th Workshop Bayesian Deep Learn.*, 2019, pp. 1–8.

[24] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with Logit normalization," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 23631–23644.

[25] A. Meng, H. Wang, S. Aziz, J. Peng, and H. Jiang, "Kalman filtering based interval state estimation for attack detection," *Energy Procedia*, vol. 158, pp. 6589–6594, Feb. 2019.

[26] N. Živković and A. T. Sarić, "Detection of false data injection attacks using unscented Kalman filter," *J. Mod. Power Syst. Clean Energy*, vol. 6, pp. 847–859, Sep. 2018.

[27] M. Jorjani, H. Seifi, and A. Y. Varjani, "A graph theory-based approach to detect false data injection attacks in power system AC state estimation," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2465–2475, Apr. 2021.

[28] T. R. B. Kushal, K. Lai, and M. S. Illindala, "Risk-based mitigation of load curtailment Cyber attack using intelligent agents in a shipboard power system," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 4741–4750, Sep. 2019.

[29] M. Jorjani, H. Seifi, A. Y. Varjani, and H. Delkhosh, "An optimization-based approach to recover the detected attacked grid variables after false data injection attack," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5322–5334, Nov. 2021.
[30] S. K. Singh, K. Khanna, R. Bose, B. K. Panigrahi, and A. Joshi, "Joint-transformation-based detection of false data injection attacks in smart grid," *IEEE Trans. Ind. Informat.*, vol. 14, no. 1, pp. 89–97, Jan. 2018.
[31] G. Cheng, Y. Lin, J. Zhao, and J. Yan, "A highly discriminative detector against false data injection attacks in AC state estimation," *IEEE Trans. Smart Grid*, vol. 13, no. 3, pp. 2318–2330, May 2022.
[32] J. J. Q. Yu, Y. Hou, and V. O. K. Li, "Online false data injection attack detection with wavelet transform and deep neural networks," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3271–3280, Jul. 2018.
[33] S. Wang, S. Bi, and Y.-J. A. Zhang, "Locational detection of the false data injection attack in a smart grid: A multilabel classification approach," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8218–8227, Sep. 2020.
[34] G. Zhang, J. Li, O. Bamisile, D. Cai, W. Hu, and Q. Huang, "Spatio-temporal correlation-based false data injection attack detection using deep convolutional neural network," *IEEE Trans. Smart Grid*, vol. 13, no. 1, pp. 750–761, Jan. 2022.
[35] X. Gao, X. Yang, L. Meng, and S. Wang, "Fast economic dispatch with false data injection attack in electricity-gas cyber–physical system: A data-driven approach," *ISA Trans.*, vol. 137, pp. 13–22, Jun. 2022.
[36] Y. Zhang, J. Wang, and B. Chen, "Detecting false data injection attacks in smart grids: A semi-supervised deep learning approach," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 623–634, Jan. 2021.
[37] K. Dehghanpour, Z. Wang, J. Wang, Y. Yuan, and F. Bu, "A survey on state estimation techniques and challenges in smart distribution systems," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2312–2322, Mar. 2019.
[38] S. Wei, J. Xu, Z. Wu, Q. Hu, and X. Yu, "A false data injection attack detection strategy for unbalanced distribution networks state estimation," *IEEE Trans. Smart Grid*, vol. 14, no. 5, pp. 3992–4006, Sep. 2023.
[39] H. Margossian, M. A. Sayed, W. Fawaz, and Z. Nakad, "Partial grid false data injection attacks against state estimation," *Int. J. Electr. Power Energy Syst.*, vol. 110, pp. 623–629, Sep. 2019.
[40] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5580–5590.
[41] J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deep deterministic uncertainty: A new simple baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 24384–24394.
[42] Y. Li, X. Wei, Y. Li, Z. Dong, and M. Shahidehpour, "Detection of false data injection attacks in smart grid: A secure federated deep learning approach," *IEEE Trans. Smart Grid*, vol. 13, no. 6, pp. 4862–4872, Nov. 2022.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[44] O. M. Eidous and E. A. Ananbeh, "Kernel method for estimating overlapping coefficient using numerical integration methods," *Appl. Math. Comput.*, vol. 462, Feb. 2024, Art. no. 128339.
[45] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[46] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse, "Flipout: Efficient pseudo-independent weight perturbations on mini-batches," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.

Guangxu Feng (Graduate Student Member, IEEE) received the B.S. degree in automation science and technology from Xi'an Jiao Tong University, Xi'an, China, in 2018. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China.

His research interests include cyber-physical security of power systems and out-of-distribution deep learning application in smart grids.

Keng-Weng Lao (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and electronics engineering from the Faculty of Science and Technology, University of Macau, Macau, China, in 2009, 2011, and 2016, respectively.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering and the State Key Laboratory of Internet of Things for Smart City, University of Macau. He was a Research Scholar with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA, from June 2017 to June 2019. His research interests include cyber-physical security, renewable energy integration, energy Internet of Things, smart energy system protection, and smart grid. He was the recipient of the Macao Science and Technology Development Fund Postgraduate Award for Ph.D. student in 2016, and the first runner-up of the Challenge Cup National Inter-Varsity Science and Technology Competition in 2013.

Ge Chen (Member, IEEE) received the B.S. degree in thermodynamic engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2015, the M.S. degree in thermodynamic engineering from Xi'an Jiaotong University in 2018, and the Ph.D. degree in electrical and computer engineering from the University of Macau, Macau, China, in 2023. He is currently a Postdoctoral Research Associate with Purdue University. His research interests include Internet of Things for smart energy, optimal operation, and data-driven optimization under uncertainty.