A Data-Driven Framework for Power System Event Type Identification via Safe Semi-Supervised Techniques

Yuxuan Yuan, *Member, IEEE*, Yanchao Wang, *Graduate Student Member, IEEE*, and Zhaoyu Wang, *Senior Member, IEEE*

***Abstract*—This paper investigates the use of phasor measurement unit (PMU) data with deep learning techniques to construct real-time event identification models for transmission networks. Increasing penetration of distributed energy resources represents a great opportunity to achieve decarbonization, as well as challenges in systematic situational awareness. When high-resolution PMU data and sufficient manually recorded event labels are available, the power event identification problem is defined as a statistical classification problem that can be solved by numerous cutting-edge classifiers. However, in real grids, collecting tremendous high-quality event labels is quite expensive. Utilities frequently have a large number of event records without in-depth details (i.e., unlabeled events). To bridge this gap, we propose a novel semi-supervised learning-based method to improve the performance of event classifiers trained with a limited number of labeled events by exploiting the information from massive unlabeled events. In other words, compared to existing data-driven methods, our method requires only a small portion of labeled data to achieve a similar level of accuracy. Meanwhile, this work discusses and addresses the performance degradation caused by class distribution mismatch between the training set and the real applications. Specifically, this method utilizes pseudo-labeling technique to investigate the value of unlabeled events and incrementally expands the training dataset. Moreover, a safe learning mechanism is developed to mitigate the impacts of class distribution mismatch and prevent performance degradation. Based on the proposed safe learning mechanism, our model does not directly use all unlabeled events during model training, but selectively uses them through a comprehensive evaluation procedure. Numerical studies on a sizable PMU dataset have been used to validate the performance of the proposed method.**

***Index Terms*—Event identification, phasor measurement unit, safe learning, semi-supervised model, unlabeled event.**

Nomenclature

CNN	Convolutional neural network.
FP	False positive.
FN	False negative.
MCC	Matthews correlation coefficient.
PMU	Phasor measurement unit.
TP	True positive.
TN	True negative.
d	Length of analysis window.
D_l	New labeled data in each iteration.
$f(\cdot)$	Rectified linear function.
$h(\cdot)$	Encoder network.
K_l^m	Kernel filter of the m-th feature map of the l-th layer.
k	Number of unlabeled events.
$L(\cdot)$	Softmax cross-entropy loss.
m	Number of labeled events.
n	Size of events.
N_l	Labeled event set.
N_u	Unlabeled event set.
N^t	Number of events marked in the t iteration.
P_l	Size of feature maps in the l-th layer.
S_f	Shared feature extractor.
S_1, S_2, S_3	Three event classifiers.
u	Classification noise rate.
U^t	Upper bound of the classification error rate.
W	Number of repeated estimations in each iteration.
$w(\cdot)$	Weight function.
x_i	PMU measurement for event i
y_i	Label for event i
z_i	Samples from the standard normal distribution.
ω	Frequency with which a classifier differs from other classifiers.
γ	Parameter of weight function.
$\Omega(\cdot)$	Regularization term.
ε_G	Gaussian noise.
θ	Parameter of classifier.
η_θ	Learning rate for θ
η_γ	Learning rate for γ
τ	Search space in convolution layer.
ϵ	Hypothesis worst-case classification error rate.

Manuscript received 6 August 2022; revised 13 November 2022 and 14 February 2023; accepted 7 April 2023. This work was supported in part by the U.S. Department of Energy Office of Electricity under Grant DE-OE000910 and in part by the National Science Foundation under Grants EPCN 1929975 and EPCN 2042314. Paper no. TPWRS-01152-2022. *(Corresponding author: Zhaoyu Wang.)*

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: yuanyx@iastate.edu; yanchaow@iastate.edu; wzy@iastate.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPWRS.2023.3266153.

Digital Object Identifier 10.1109/TPWRS.2023.3266153

I. Introduction

With the modernization of power systems, system operators are expected to meet the growing demands of their customers while maintaining the reliability of the power supply. Recently, the increasing penetration of phasor measurement units (PMUs) [1] provides a unique opportunity to improve situational awareness of the system [1]. Typically, PMUs are installed into selected substations and interfaced to the grid via instrument transformers to measure frequency, rate of frequency change, voltage, and current phasors based on the united Coordinated Universal Time reference. PMUs are more accurate and faster (i.e., 30-60 samples per cycle) than supervisory control and data acquisition systems with low sampling rates (i.e., 2-4 samples per cycle) [2]. Inspired by these benefits of PMUs, researchers have dedicated great efforts on data-driven methods for real-time system monitoring and protection using PMU data [3]. Compared to conventional model-based event identification methods, data-based approach has the unique advantage of operating independently of the system.

Depending on whether the model requires a large number of recorded event labels, two categories of existing data-driven event classification methods are summarized. Studies in the first category follow a supervised learning fashion to associate PMU measurements with recorded event labels [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. In [4], a discrete wavelet transform-based deep neural network model was proposed to reduce false disturbance detection and validate true events. In [5], a three-stage framework was proposed for training robust event classifiers to address the data quality issues of PMU measurements. In [6], two well-established supervised learning methods (i.e., k-nearest neighbor and support vector machine (SVM)) were trained and tested on the basis of thousands of simulated events created by GE's PSLF software. In [7], a three-layer deep neural network-based method was designed to identify power system events using data from 187 PMUs and 1,000 real-world events. In [8], an empirical wavelet transform-based random forest method was proposed to assess power system events. The model was trained and tested based on PSS/E simulation. In [9], a one-versus-many extreme learning machine model was developed to perform event diagnosis by combining 3,495 simulated events and 81 real-world events collected from four PMUs located in Western Electricity Coordination Council. [10] introduced a dictionary of row subspaces of different event types and identified an event by comparing the subspace of the obtained PMU data with the dictionary. In [11], an event characterization algorithm was proposed to calculate spectral kurtosis and used it as the input to SVM for event identification. In [12], a threshold-based OR rule was presented to identify events using rank signatures of PMU measurements. In [13], a deep learning-based event classification model was designed to introduce robustness against bad data issues in online applications. In [14], a symbolic aggregation approximation technique was used to compress and convert PMU data features. Ensemble learning and SVM algorithms were utilized to perform event classification. These efforts have generally shown good results. However, the main concern with category I models is that good performance depends on the availability of sizable labeled events (e.g., thousands of simulated events). As demonstrated concretely in [15], limited training samples usually reduce the accuracy and generalization of supervised event classification models. In reality, even for stable grids with long-term operations and few events, the number of event labels is limited.

[1] According to statistical data provided by the North American SynchroPhasor Initiative, over 1,900 PMUs have been installed in the U.S., which is a nine-fold growth from 2009.

Utilities often have records of events without in-depth details. Of the 2,226 recorded events observed by Public Service Company of New Mexico over four years, only 97 events were registered in the event logs [16]. Considering that category I methods typically struggle to perform adequately with few labeled events, researchers are exploring a variety of unsupervised and transfer learning strategies to perform event detection and identification [17], [18], [19], [20], [21], [22], [23]. In [17], a heterogeneous joint domain adaptation method with a transfer learning strategy was proposed to transfer knowledge from a data-rich source grid to the data-limited target grid to boost the machine learning performance in the target grid. In [18], a statistics-based framework was proposed to detect events using PMU data. In [19], a two-stage framework was proposed to achieve real-time event detection, physically meaningful event type distinction, and localization using principal component analysis and hierarchical clustering technique. In [20], a transfer learning-based mechanism was proposed to address the issue of event detection from a remarkably small number of labeled events. In [21], three existing clustering algorithms (i.e., partitioning, hierarchical, and density-based methods) were evaluated to group disturbance files. In [22], a novel characteristic ellipsoid method was proposed to identify types and locations of transient events. In [23], a kernelized tensor decomposition and classification framework was proposed to incorporate rich unlabeled data. While existing unsupervised and transfer learning-based event identification works provide valuable results, several questions remain open. For example, unsupervised learning-based methods cannot provide the physical meaning of event types. The results of these methods are usually broadly defined categories and thus can only provide limited help for real-time system monitoring. A natural way to deal with this question is to associate and define each category using data from labeled events and domain knowledge. However, this solution relies on an important assumption that labeled events and unlabeled event types are identical. In other words, the utilities need to observe and register all possible event types. In practice, it is difficult to maintain such an assumption. Unlabeled events often hide a variety of new event types, which is also mentioned by previous work [6]. In this paper, this situation is referred to as the *class distribution mismatch problem* (as shown in Fig. 1), which greatly increases the difficulty of data-driven event classification tasks. Last but not least, the results of unsupervised techniques tend to have low accuracy due to the lack of labeling information.

To address these problems, this paper proposes a novel data-driven model to identify power event types in a semi-supervised learning manner. Compared to supervised learning-based models, the proposed model is better suited for real-world tasks because collecting tremendous high-quality event labels is quite expensive. To achieve this, our method leverages an output smearing strategy to build three different classifiers and initially

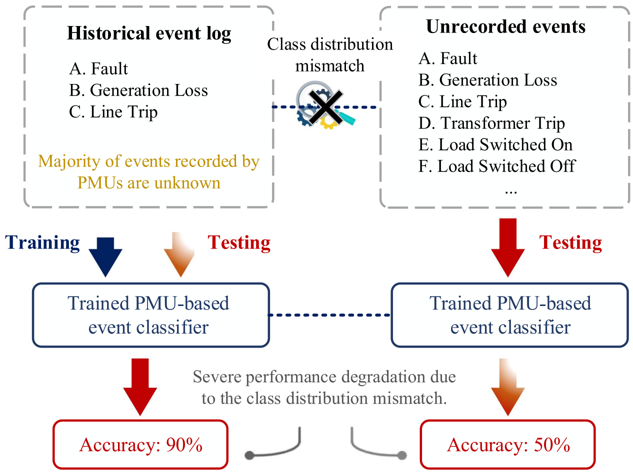


Fig. 1. Description of the event identification problem under the class mismatch problem.

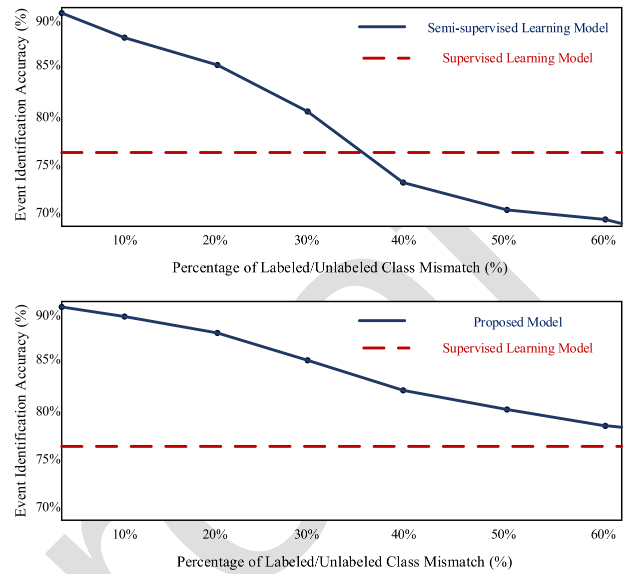


Fig. 2. As the class distribution mismatch ratio between the labeled and unlabeled data rises, the performance of traditional semi-supervised learning approaches drastically declines. When the mismatch exceeds a certain range, the performance of the traditional semi-supervised learning method is even worse than that of the supervised learning method (top). Such a performance degradation hinders the motivation to use semi-supervised learning techniques in the vast majority of real applications. In contrast, the performance of the proposed method similarly declines as the class distribution mismatch between labeled and unlabeled data increases, but it never performs worse than the performance of the supervised learning method (bottom).

trains them using labeled events in parallel. Considering the high model complexity due to the high dimensionality of PMU measurements, convolutional neural networks (CNNs) are used as the underlying classifier in this work. The unique benefit of utilizing three event identifiers is that it provides a workaround for marking unlabeled events. Specifically, if any two of classifiers have a consistent estimate for an unlabeled event, then this estimate is confident and can be added to the training set. The three event identifiers are retrained using the updated training set in order to consistently benefit from the abundance of unlabeled events. Considering that unseen event types do not exist in the initial training set, it is impossible for three classifiers to give meaningful estimates for these types. Therefore, the training process of each model is projected as a bi-level optimization problem to avoid pseudo-labeling of events under unseen types as much as possible, which is defined as a safe learning mechanism. A weighted empirical risk minimization model is to be obtained in the inner-layer optimization. Additionally, the goal of the outer-layer optimization is to minimize classification loss on a given training set. An online approximation method is applied to solve this bi-level optimization. By combining these novel modules, a better generalization ability can be achieved. The main contributions of this paper can be summarized as follows:

- The proposed framework can improve the performance of event classifiers trained with a limited number of labeled events. The proposed method is able to achieve similar accuracy as supervised learning methods using all labeled data, but using only 25% of the labeled data.
- The proposed framework not only exploits the value of unlabeled events, but also provides a basis for significantly reducing the impact of the class distribution mismatch problem to enhance event classifier performance.
- The proposed safe learning strategy prevents features of unseen events from becoming entangled with features of observed events, thus avoiding performance degradation of the model on known event types. Such a mechanism can help the proposed model to perform no worse than its supervised counterpart in extreme cases.
- The proposed model was developed and tested based on two years of data from hundreds of PMUs and approximately 4,800 event records from Western Interconnection. In our experiments, we constantly assume that a portion of the event records are unknown to simulate different real situations. All results are derived by comparing predictions and ground truths.

The rest of this paper is structured as follows. The preliminaries of the proposed framework are shown in Section II, including the data description and problem formulation. Section III introduces the semi-supervised learning-based event identification. Section IV presents the safe learning process. Case studies are demonstrated in Section V. Research conclusions are provided in Section VI.

II. Preliminaries

A. Data Description and Pre-Processing

The available 2-year PMU measurements were initially collected by regional system operators and utilities in the Texas, Western, and Eastern Interconnections of the U.S. and then formatted by Pacific Northwest National Laboratory. Each PMU monitors the system frequency, voltage, and current phasors, as well as the rate at which the frequency changes. The majority of PMU data segments are archived at 30 frames per second and the rest at 60 frames per second. In addition to 20 TB of

PMU streaming data, this dataset has the particular advantage of containing enough real event labels (i.e., 6,767 events from utilities), which creates a solid foundation for designing an effective event classification model. Note that complete detection criteria for all types of events and historical protection records are not provided in this work due to the safeguarding of sensitive information, making them unavailable for classification model development.

The data pre-processing is done prior to model development to assure the quality of the training data, preventing inaccurate event detection brought on by data quality issues. This procedure is empirical and follows the guidance of our industrial partners. Briefly, the first phase in data pre-processing is to use PMU status flag information to identify data with data quality issues. According to IEEE C37.118.2-2011 standard, when the decimal status flag value is 0, PMU measurements can be used to accurately describe the system status. Instead, PMU is in a malfunction state. In addition, based on engineering intuition, we designed several threshold-based methods to identify data quality issues that are not detected by the PMU itself, such as out-of-range issues. In the second phase, when consecutive missing or bad data happens, the data is removed from our study. The justification for this is that it is challenging to offer precise data imputation for these consecutive bad data, which is also out-of-scope of this work. Linear interpolation is then used to fill in and repair the remaining missing or bad data. After data processing, the latent data features are extracted using Markov transition field techniques. By calculating Markov transition probabilities and converting that data into graphs, Markov transition fields can preserve all time-domain information. More details can be found in our previous works [13], [24]. Note that the system topology, PMU locations, and historical event locations are not available. Hence, this work cannot be extended to identify the location of events. We leave it for future work. Once they are available, more comprehensive results will be provided.

B. Problem Formulation

In terms of notation, let x_i denote its i-th entry in a column vector $\mathbf{x}$. Given a matrix $\mathbf{X}$, let $X_{(i,j)}$ denote its entry at i-th row and j-th column and $[\mathbf{X}]_i$ denotes its i-th row. The *estimation* is indicated by the superscript $\widehat{(\bullet)}$ and the *optimum* is shown by the superscript $(\bullet)^\star$.

Consider a set of PMU data $N_l = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, the data-driven event identification problem can be formulated as an n type classification problem [7], where $x_i \in \mathbb{R}^{d\times 1}$ is the measurement data of PMU with d-length analysis window[2], $y_i \in \{1, \ldots, n\}$ is the label recorded in the disturbance files after using a label encoding technique, m is the total number of recorded events, and n is the number of event types. In order to achieve satisfactory event identification accuracy, a large amount of labeled events are necessary.[3] However, in power systems, such a condition is difficult to meet because obtaining labeled events is costly in terms of human and financial resources. As mentioned in previous work [6], most of the events recorded by PMUs are unknown. Therefore, the event identification problem needs to be refined to an $n+1$ type classification problem, where $N_l = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ and k unlabeled events $N_u = \{x_{m+1}, \ldots, x_{m+k}\}$. Here, $y_i \in \{1, \ldots, n+1\}$, where $(n+1)^{\text{th}}$ represents an unspecified type recorded in event logs. For the $(n+1)^{\text{th}}$ type, a natural assumption is that the $(n+1)^{\text{th}}$ type is a mixture of known event types. This is one of the common assumptions used in previous works [17], [19], [20], [22]. Under this assumption, the lack of labeled event data can be overcome by finding associations between known event types and the $(n+1)^{\text{th}}$ type using state-of-the-art unsupervised techniques. However, this assumption is not practical in many cases. In reality, the number of unlabeled data is much larger than the number of labeled data (i.e., $k >> 0$). This results in the $(n+1)^{\text{th}}$ type often consisting of two parts: the events belonging to the known types but not identified by utilities and all other types of events that are not seen in the event logs. Hence, unrecorded events and recorded events do not share the same distribution, which is known as class distribution mismatch, as shown in Fig. 1. Note that our model is built based on this actual situation rather than on the previous assumption. When the different unknown events that are classified in the $(n+1)^{\text{th}}$ type have markedly different underlying physics, they may have highly distinct characteristics and cannot be categorized in any of the known types. Face with this situation, since conventional semi-supervised models have never seen the types of these events, it is impossible for the model to provide correct estimation for unlabeled set and derive any useful information from them. Moreover, the characteristics of the unknown events are entangled with the characteristics of the observed events, which significantly impairs the trained model's ability to judge events of known types (also known as performance degradation). This is the reason why most semi-supervised learning algorithms no longer work well, and may even be worse than a simple supervised learning model (i.e., support vector machine, logistic regression, and random forest) [25]. It should be noted that supervised models do not suffer from this problem, as they only focus on those labeled events. Such shortcomings limit the application of deep semi-supervised models in power event classification problems.

To develop a practical event identification model, we propose a safe tri-net-based method that only requires limited labeled events without any class distribution assumptions. Briefly, our work uses the idea of pseudo labeling to discover the value of unlabeled events[4] to improve the performance of the event

[2]In this work, a 2-second analysis window is utilized to intercept PMU measurements based on event logs. This 2-second analysis window consists of 0.5 pre-event data and 1.5 post-event data. The value of d is determined based on previous studies [13], [24]. Note that the selection of d is a trade-off between event information and the curse of dimensionality. Also, as the input dimension increases, the computation complexity of the data-driven event identification model grows significantly, which can impact the real-time application of models.

[3]The amount of data required for machine learning depends on many factors, including the complexity of the problem and the complexity of the learning algorithm. Based on the high sampling rate of PMUs, the amount of data required to realistically train and test a classifier is enormous.

[4]Pseudo labeling is a commonly-used method to perform semi-supervised learning tasks. The basic idea of this method is to seek the generation of pseudo labels for unlabeled samples to guide the learning process in an alternating manner. Specifically, the initial model is trained using the limited labeled data.

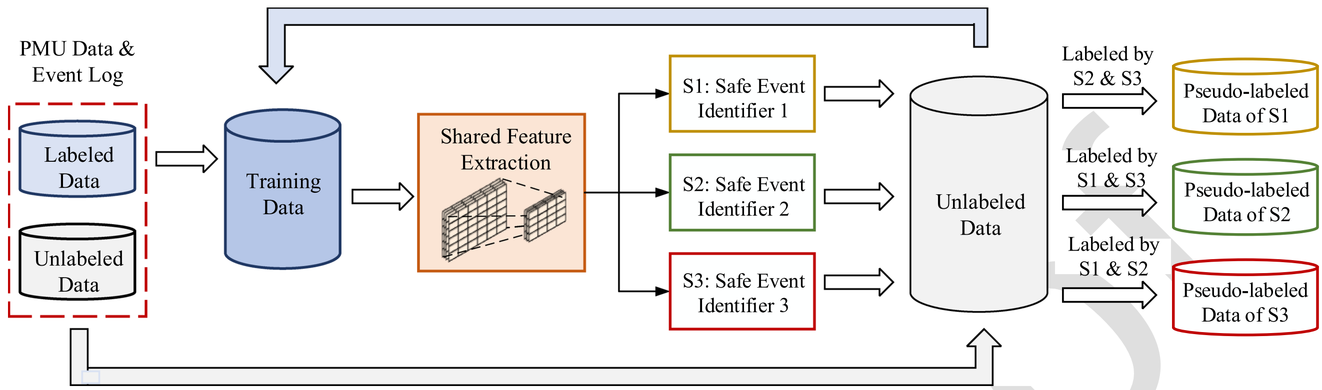


Fig. 3. Overall framework of the proposed safe learning-based event identification model.

identifiers when training with limited event logs [26]. However, unlike previous models, our method can handle class distribution mismatch by incorporating a bi-level optimization in the backpropagation process. By designing a weight function, the proposed method uses unlabeled data selectively. In each iteration, the model searches for the optimal model parameters based on weighted empirical risk minimization. The weight function parameters are then improved to continuously track the supervised performance once the obtained model parameters are evaluated on labeled events. The trained event identifier will therefore not perform worse than a supervised learning-based event identification model when utilizing our method, even if event logs do not cover all event types. We will demonstrate in the following section that the event identifier learned using the proposed approach is always better than the model developed using simply labeled data.

C. Proposed Event Identification Framework

The objective of this work is to design a framework to improve the performance of event classifiers in a safe manner. Given the prevalence of unlabeled data in all grids, the data resources required to train the proposed event classification model consist of unlabeled data and a limited amount of labeled data. Different stages of the proposed framework are demonstrated in Fig. 3.

- *Stage I - Tri-net Classifier Initialization:* A tri-net-based framework is developed to perform event identification in a semi-supervised learning manner. As shown in Fig. 3, the proposed framework consists of a shared feature extractor (S_f) and three safe event identification modules (S_1, S_2, and S_3) with different structures. The three event classifiers build the mapping relationship between shared features and event types. An output smearing strategy is used to construct three diverse training sets, thus augmenting diversity between three classifiers (detailed in Section III).
- *Stage II - Safe Learning-based Parameter Optimization:* A safe learning mechanism is proposed to update model parameters for each classifier. Such a mechanism can prevent performance degradation due to the class distribution mismatch problem. The basic idea is to weaken unlabeled data with unseen classes by adding a weight function and tracking supervised loss by designing a bi-level optimization (detailed in Section IV).
- *Stage III - Pseudo-Label Dropout:* To further deal with the low-confidence pseudo labels, a dropout strategy is applied during the training process. Basically, this strategy exploits the disagreements among the three classifiers. With three classifiers, if any two of them have a consistent estimate for an unlabeled event, then this estimate is confident and can be added to the training set, as shown in Fig. 3. Such an augmented training set is utilized to refine the three classifiers until the end of the training process (detailed in Section IV).

Then, the trained model is utilized to generate pseudo labels for the unlabeled samples. Based on the updated training dataset, the model is retrained.

III. Semi-Supervised Learning-Based Event Identification

This section outlines the proposed safe tri-net-based approach. We quickly review the concepts and characteristics of conventional semi-supervised learning techniques before describing our method in depth to help the reader comprehend the proposed model.

Semi-supervised learning is a learning paradigm linked with developing models using all available data, including labeled and unlabeled data, and is conceptually positioned between supervised and unsupervised learning. Compared to supervised learning approaches, Semi-supervised learning techniques are better suited for real-world tasks where unlabeled data are easily accessible whereas labeled cases need more resources and time to collect. The goal of the semi-supervised learning model is to use all available data to generate a predictive function that is more accurate than the one obtained using only labeled data. When dealing with classification problems, leveraging unlabeled data with a semi-supervised method can provide us with additional information about the shape of the decision boundary among different classes. According to previous studies, semi-supervised learning methods can be broadly divided into two categories: transductive learning and inductive learning [27]. Basically, transductive learning aims to apply the trained models to the unlabeled data observed at training time; in this case, it does not generalize to unobserved data. In contrast, the goal of inductive learning is to learn a model capable of generalizing to unobserved data at test time. This categorization applies to the proposed approach.

One of the major challenges of the semi-supervised event identification method is how to produce additional training data by labeling instances of the unlabeled set. Inspired by the tri-training methodology [26], the proposed model utilizes three different classifiers to handle the challenge of identifying unlabeled events. It should be noted that the initial classifiers should be diverse. When all classifiers are the same, they will all produce the same estimate for each unlabeled event, which will impede the model training. In this work, to construct three diverse modules, an output smearing strategy is applied [28]. By adding random noise to true labels, this strategy can construct diverse training sets, which can be formulated as follows:

$$\hat{y_i} = y_i + f(z_i \times \sigma_i) \tag{1}$$

where, z_i is sampled independently from the standard normal distribution, σ_i is the standard deviation, and $f(\cdot)$ is represented by the rectified linear function. With the output smearing strategy, three diverse training sets can be obtained from the initial labeled set. Then, the objective function of our method is to minimize the sum of the three identifiers' losses, which is defined as follows:

$$\begin{aligned} \min_{\Theta} \sum_{i=1}^{m} \{ & L(S_1(S(x_i), \hat{y_i^1})) \\ & + L(S_2(S(x_i), \hat{y_i^2})) + L(S_3(S(x_i), \hat{y_i^3}))\} \end{aligned} \tag{2}$$

where, $L(\cdot)$ denotes the standard softmax cross-entropy loss in this work (i.e., a softmax activation plus a Cross-Entropy loss). The shared module S_f is designed by using one convolutional and max-pooling layers. The parameters of the S_f are updated by learning all gradients from S_1, S_2, and S_3. The structure of classifiers S_1, S_2, and S_3 are derived from state-of-the-art convolutional neural network architecture [29]. To get more diversity among the three classifiers, different structures (i.e., different network depths and convolution parameters) were used for the three classifiers. In order to assist readers unfamiliar with deep learning, we outline each typical layer below:

- *Convolutional Layer:* Convolutional layers typically run an operation ($*$) on the input and pass the result to the following layer. In this work, after feature reconstruction, all event signals are considered as two-dimensional graphs, making the convolutional layer mathematically formulated as follows:

$$\begin{aligned} & (x_{l-1} * K_l^m)(i,j) \\ & = \sum_{\tau_i=0}^{P_l} \sum_{\tau_j=0}^{P_l} x_{l-1}^s(i-\tau_i, j-\tau_j) K_l^m(i,j) \end{aligned} \tag{3}$$

  where, K_l^m is the kernel filter of the m-th feature map of the l-th layer, P_l refers to the size of feature maps in the l-th layer, and τ_i and τ_j are the search paces in the horizontal and vertical directions, respectively. As a result, the convolutional layer performs an element-wise multiplication in a sliding-window manner. It will summarize the results into a single output and transform a feature matrix into a different feature matrix, whose dimensionality of the new matrix is determined by the dimensionality of the original matrix and the dimensionality of the kernel filter.
- *Activation Layer:* To compensate for the limitations of linear modeling in the convolutional layer, the results of the convolutional layer are given to a nonlinear function (e.g., sigmoid, tanh, softmax, ReLU, leaky ReLU, etc.). The activation layer is the name given to this nonlinear function. In this study, all layers but the fully linked layer are activated using Leaky-ReLU, while the fully connected layer is activated using soft-max.
- *Max-pooling Layer:* The feature maps are aggregated using a maximum pooling layer following activation function and batch normalization. Max pooling is essentially a pooling procedure that chooses the largest element from the feature map region that the filter covers. In other words, a feature map comprising the standout features from the prior feature map will be the output following the maximum pooling layer. In this paper, a 2×2 max-pooling is used.

In contrast to conventional semi-supervised models that require explicitly measuring confidence in pseudo-labeling (i.e., self-training), our method provides a natural and efficient mechanism for evaluating pseudo labels of unlabeled events. As demonstrated in Fig. 3, for any identifier, an unlabeled event can be labeled when two other identifiers agree on the label of this event. For example, x_i can be added to the training set for S_3 if S_1 and S_2 concur on the label of the event. Following this strategy, each classifier is retrained using the augmented training set in each iteration. Note that the structure of the classifiers should be different. Otherwise, the unlabeled events identified by the other two classifiers will be the same as those labeled by the other two classifiers for either of the classifiers. Obviously, even if our method uses two classifiers to increase the confidence of pseudo labels, incorrect pseudo-labeling is inevitable. These incorrect pseudo labels would degrade the performance of the classifiers during the training process. Therefore, we will show that the increase in the classification error can be offset if the amount of newly labeled data can adhere to certain requirements:

S_1 and S_2 classified instances with pseudo-labels are added to the training set of S_3 as examples to prove our conclusion above. First, let N^t and N^{t-1} refer to the number of data that are labeled for S_3 in the t-th and $t-1$-th iteration, respectively. Let u_{N_L} and $U^t_{S_1,S_2}$ denote the classification noise rate of the original training set N_l and the upper bound of the classification error rate caused by S_1 and S_2 at the $t-1$-th iteration. According to the finding of [30], the inverse of the square of the error at the t-th iteration (i.e., $\frac{1}{(\epsilon^t)^2}$) can be formulated as:

$$\frac{1}{(\epsilon^t)^2} = |N_l \cup N^t| \left(1 - 2\frac{(u_{N_l}|N_l| + U^t_{S_1,S_2}|N^t|)}{|N_l \cup N^t|}\right)^2 \tag{4}$$

Basically, if $\epsilon^t < \epsilon^{t-1}$, it implies that S_3 can be improved through using newly labeled data (i.e., D_l) from S_1 and S_2:

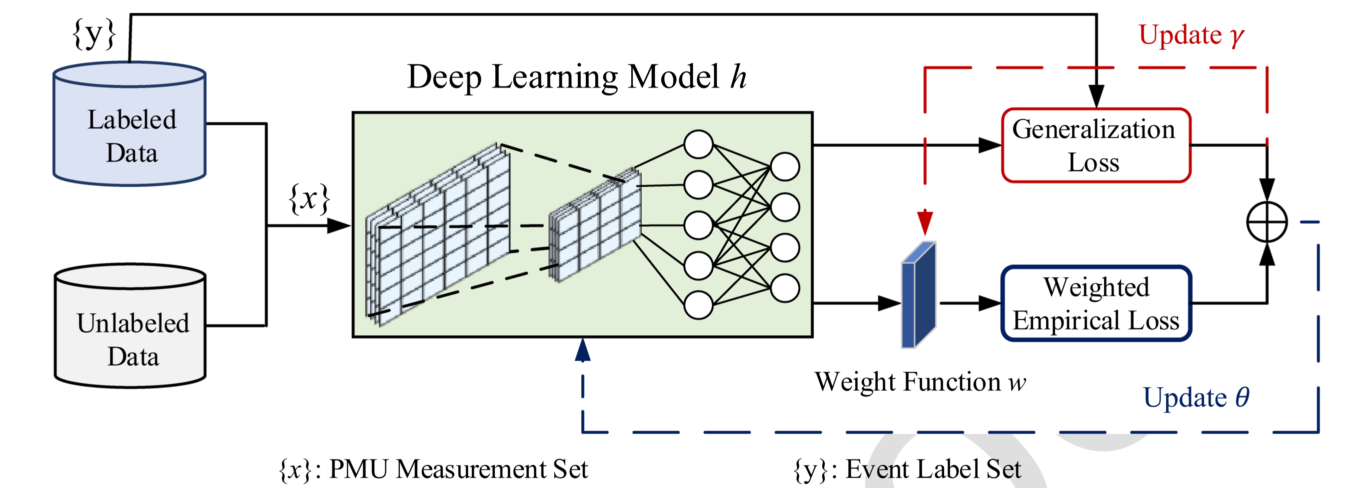


Fig. 4. Illustration of safe event identification model.

$$|N_L \cup N^t| \left(1 - 2\frac{(u_{N_l}|N_l| + U^t_{S_1,S_2}|N^t|)}{|N_l \cup N^t|}\right)^2 >$$
$$|N_l \cup N^{t-1}| \left(1 - 2\frac{(u_{D_l}|D_l| + U^{t-1}_{S_1,S_2}|N^{t-1}|)}{|D_l \cup N^{t-1}|}\right)^2 \quad (5)$$

When $U^t_{S_1,S_2}$ and $U^{t-1}_{S_1,S_2} \in [0, 0.5)$, (5) always holds if $|N^t| > |N^{t-1}|$ and $U^{t-1}_{S_1,S_2}|N^{t-1}| > U^t_{S_1,S_2}|N^t|$. In sum, S_3 can be improved when the following constraint is satisfied:

$$0 < \frac{U^t_{S_1,S_2}}{U^{t-1}_{S_1,S_2}} < \frac{|N^{t-1}|}{|N^t|} < 1 \quad (6)$$

This constraint cannot hold when $|N^t|$ is far bigger than $|N^{t-1}|$, which is possible. When this occurs, a subsampling method is applied for N^t randomly remove a portion of the data to maintain (6). It is combined with the proposed structure to specify the conditions under which unlabeled data may be labeled for a classifier.

IV. Safe Event Identification Model

A. Model Formulation

Considering that class distribution mismatch occurs in actual grids, it is not reasonable to estimate pseudo labels to unlabeled data under unseen classes because the training model never learns the features of this class. Such a problem leads to severe performance degradation when applying conventional semi-supervised learning techniques in power event identification. To solve this question, a safe learning mechanism is proposed based on the structure mentioned in the previous section. Specifically, the proposed mechanism designs a weight function to use unlabeled data selectively and continuously tracks the performance of the supervised learning model to prevent performance degradation. To achieve this, each event classifier (i.e., S_1, S_2, and S_3) is destined as a bi-level optimization problem, where one optimization problem is nested inside another issue. Fig. 4 describes this process. The basic idea is to use as many beneficial unlabeled events as possible and keep track of supervised loss to prevent performance degradation. To achieve this, first, our method minimizes a weighted empirical risk[5] by integrating a weight function with a regularization strategy for the unlabeled events. The objective function can be formulated as follows:

$$\hat{\theta} = \min_{\theta} \sum_{i=1}^{m} L(S(x_i;\theta), y_i) + \sum_{i=m+1}^{m+k} w(x_i;\gamma)\Omega(x_i;\theta) \quad (7)$$

where, $\hat{\theta}$ is denoted as the model trained with the weight function parameterized by γ, and $\Omega(\cdot)$ refers to the regularization term. In this work, we have applied a consistency regularization strategy to formulate $\Omega(\cdot)$ [31]:

$$\Omega(x;\theta) = ||h(x + \varepsilon_G;\theta) - h(x;\theta)||_2^2 \quad (8)$$

where, $h(\cdot)$ is a standard encoder network that maps input data to a lower dimensional space and ε_G refers to Gaussian noise. The aim of the regularization term is to train a model that is invariant to various data augmentations, which provides the basis for using unlabeled data to augment prediction function [32].[6] Using the weight function, unlabeled events can be utilized selectively, thus reducing the impact of the distribution mismatch problem. Then, the proposed model evaluates $\hat{\theta}$ on m labeled events and optimizes the weight function parameter γ to avoid severe performance degradation. This optimization can be formulated as follows:

$$\hat{\gamma} = \min_{\theta} \sum_{i=1}^{m} L(S(x_i;\theta), y_i) \quad (9)$$

[5]Empirical risk minimization is a principle in statistical learning theory, which is commonly used to give theoretical bounds on their performance. The basic idea is to measure model performance on a known set of training data rather than an unknown true data distribution.

[6]Mathematically, using pseudo-labeled data to augment the training set first requires adherence to the notion: if an actual perturbation is applied to an unlabeled data, the prediction should not change significantly. The underlying rationale behind this is that data points with different labels should be low density separation based on cluster assumption.

In summary, the first optimization of the proposed safe event identification (7) is to seek the optimal model parameters $\hat{\theta}$ using the entire dataset. For convenience, let $A(\theta, \gamma)$ denotes as (7). Next, the learned model parameter $\hat{\theta}$ is evaluated in the labeled dataset and the weight function parameters γ are optimized, as shown in (9), to achieve a better reliable performance, which is represented by $B(\theta)$. Consequently, the following bi-level optimization problem can be expressed as the objective of the proposed safe event identification model:

$$\min_{\gamma} \sum_{i=1}^{m} L(S(x_i; \hat{\theta}), y_i) \tag{10}$$

$$s.t.\ \ \hat{\theta} = \min_{\theta} \sum_{i=1}^{m} L(S(x_i; \theta), y_i) + \sum_{i=m+1}^{m+k} w(x_i; \gamma)\Omega(x_i; \theta) \tag{11}$$

The unique benefit of the proposed safe learning mechanism is to introduce safeness in terms of empirical error. In other words, by optimizing γ, the proposed method does not perform worse than its supervised counterpart.

B. Model Training

Since there is no closed-form expression for this bi-level optimization problem, it necessitates two nested loops of optimization to obtain the optimal $\gamma^{\star}$.[7] As a result, the computational complexity of the training process increases significantly as the size of the training data increases. To address this issue, the parameter optimization in the proposed model follows an alternating manner. Such a strategy can significantly reduce the computation burden. Mathematically, given a weight function w with parameters γ_t, the update of θ_{t+1} can be obtained by the following equation:

$$\theta_{t+1} = \theta_t - \eta_\theta \nabla_\theta A(\theta_t, \gamma_t) \tag{12}$$

where, η_θ is the learning rate for classifier network. Then, following (11), γ_{t+1} can be formulated as:

$$\gamma_{t+1} = \gamma_t - \eta_\gamma \nabla_\gamma B(\theta_{t+1}) \tag{13}$$

Follow the chain rule, the gradient of $B(\theta_{t+1})$ can be reformulated as $\nabla_\theta B(\theta_t) - \eta_\theta \nabla_\gamma \nabla_\theta A(\theta_t, \gamma_t)$. To efficiently calculate this, an automatic differentiation strategy is applied [33]. Basically, for each iteration, the local descent directions of the training data are first examined on the training loss surface. Then, they are recalculated based on their similarity to the descent directions of the supervised loss surface. This strategy requires two full forward and backward passes of the network on training loss and supervised loss for parameter update, respectively. The first forward and backward pass is used to calculate the loss using $A(\theta_t, \gamma_t)$ and obtain $\nabla_\theta A(\theta_t, \gamma_t)$. Then, model parameter θ_{t+1} can be updated using (12). The weight function is then subjected to the second forward and backward pass in order to calculate the loss using $B(\theta_{t+1})$ and $\nabla_\gamma B(\theta_{t+1})$. After that, γ_{t+1} can be updated using (13). Finally, the last forward and backward pass is performed to minimize the reweighted objective to finish one iteration. Note that this process can be easily implemented using popular deep learning frameworks such as TensorFlow [34]. See Algorithm 1 and [33] for more details.

[7] For each γ, we need to compute the optimal $\hat{\theta}$. The computational complexity is $O(n^2)$. Thus, each single loop can be very expensive.

Algorithm 1: Safe Event Classifier Training using Automatic Differentiation.

Require: Labeled data $N_l = \{(x_1, y_1), \ldots, (x_m, y_m)\}$; unlabeled data $N_u = \{x_{m+1}, \ldots, x_{m+k}\}$; initial model parameter θ_0; initial weight function parameter γ_0; learning rate for model parameter η_θ; learning rate for weight function parameter η_γ; iteration number T.

```
for t = 0, ..., T − 1 do
    Select sample batch from N_l → {x_l, y_l}.
    Select sample batch from N_u → {x_u}.
    Compute generalization loss and weighted
    empirical loss using (7) → A(θ_t, γ_t).
    Calculate the gradient of model parameter
    → ∇_θ A(θ_t, γ_t).
    Update model parameter using η_θ and (12)
    → θ_{t+1}.
    Recompute generalization loss using (9).
    Calculate the gradient of weight function
    parameter → ∇_γ B(θ_{t+1}).
    Update weight function parameter using η_γ and
    automatic differentiation strategy → γ_{t+1}.
end for
```

C. Pseudo Label Dropout

During the training process, based on the estimated results of three safe event identifiers, a part of unlabeled events will be labeled and added into the training dataset. In this work, a dropout strategy is applied in pseudo labeling to exclude those pseudo-labels with low confidence and ensure the stability of the training set during the training process. Specifically, each classifier is used to estimate the label of x_i for W times throughout each iteration and record the frequency ω at which the outcome differs from the rest of the classifiers. When $\omega < \frac{W}{3}$, this pseudo label is recognized as a stable label and can be utilized for model retraining. As the value of W gets larger, it takes longer to estimate the pseudo labels in each iteration, thus greatly increasing computational burden. In other words, the selection of W is a trade-off between the stability of the pseudo labels and computational burden. In this work, different values of W are tested based on the performance of the validation set. The appropriate value of W is obtained when the accuracy of the validation set no longer increases significantly. Here, the value of W is assigned as 12.

V. Numerical Result

This section investigates the performance of our framework utilizing PMU data and related event logs from Western Interconnection. The full dataset consists of 4,800 data points taken under normal behaviors as well as 4,800 recorded events,

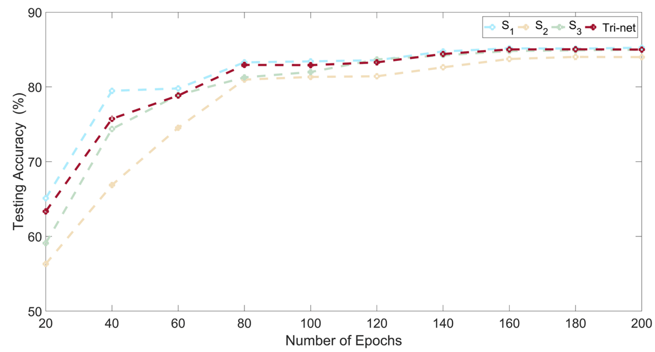


Fig. 5. Results of the proposed model's testing using 20% labeled events.

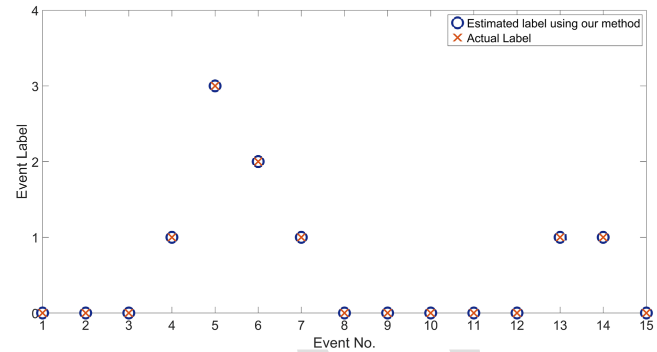


Fig. 6. Comparison of estimated event type and actual event type using the proposed method.

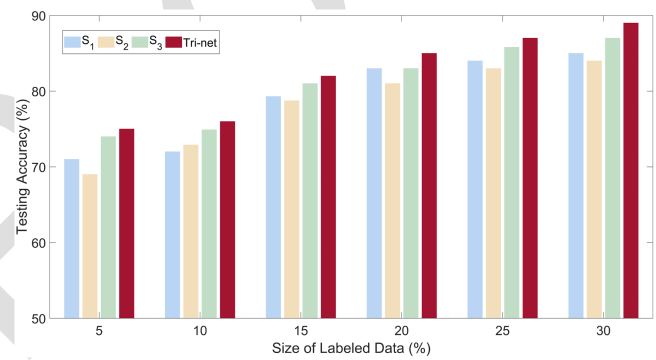


Fig. 7. Results of the sensitivity analysis using the proposed method.

such as line outages, frequency events, and transformer outages. To simulate a situation when the utility only captured a few occurrences, the event labels are kept for 25% of the records after data pre-processing. The remaining 75% of the event labels were regarded as being unidentified. This process is completely random. Considering that this dataset is an imbalanced dataset (i.e., more than 75% of the events are line outages), we randomly select 25% of the data samples for each type of event as labeled data for the purpose of model training and testing, instead of randomly selecting 25% of the data points in the entire dataset. Note that a similar data partitioning strategy is also applied to control the size of the labeled dataset in sensitivity experiments. The available dataset is then evenly divided into k equal folds, taking into account the PMU measurements and associated event labels. In this work, the value of k is selected as 5. Based on these partitioned folds, the proposed model is trained and tested in k iterations. In each iteration, one fold is left for testing and the model is trained on the remaining $k-1$ folds. With this strategy, it is possible to evaluate the performance of the suggested model using all of the available data as unseen data.

A. *Effectiveness of the Proposed Method*

The accuracy achieved from each iteration is averaged to assess the model performance using the k-fold cross validation strategy. The accuracy is calculated as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{14}$$

Where, FP stands for the false positive (e.g., event type is inferred as frequency event while its true state is normal). TN is the true negative (e.g., system state is inferred to be normal while its true state is normal). FN for the false negative (i.e., system state is inferred to be normal while its actual type is frequency event). TP refers to the true positive (e.g., event type is inferred to be a frequency event while its actual type is also frequency event). In Fig. 5, testing results for the three safe event classifiers and the suggested tri-network technique are shown. It can be seen that the single safe classifier has an accuracy range of 84 to 85% and the final testing accuracy of the tri-net method converges to about 85%. This result indicates that the proposed triple net framework is reliable and all classifiers converge to similar accuracy regions. Additionally, Fig. 6 displays the actual and estimated labels for 15 example events. As can be observed, the proposed method successfully categorizes the various event categories. It is noteworthy that these results are obtained with only 25% of the labeled events.

B. *Sensitivity Analysis*

To demonstrate how sensitive the proposed framework is to the number of labeled events, the average accuracy with varied quantities of labeled events is assessed and determined. As a result of the loss of event information, the event classifier's performance is expected to degrade as the volume of labeled data diminishes. In this case study, we gradually increase the number of labeled events from 5% to 30% (i.e., a total of 6 cases). The results are presented in Fig. 7. For each case, testing accuracy is calculated for S_1, S_2, S_3 and tri-network, respectively. As can be seen in the figure, as the percentage of labeled data rises from 5% to 30%, the model's accuracy is gradually improved. When the 30% of labeled events are available, the accuracy of the proposed method is close to 90%. Meanwhile, it is clear that the accuracy of the three modules is different, which proves the effectiveness of our model diversity strategy. By combining these three modules, a better generalization capability can be achieved. Compared to the previous study using the same dataset [13], the proposed method requires only a much smaller labeled dataset to achieve similar accuracy. Thus, the high-value use case of our algorithm is when the utility has only a very small

TABLE I
STATISTICAL ANALYSIS OF EVENT IDENTIFICATION

% of Labeled Data	Recall	Precision	F_1	MCC
5%	0.71	0.65	0.67	0.6
10%	0.76	0.66	0.69	0.64
15%	0.81	0.75	0.76	0.73
20%	0.83	0.79	0.81	0.77
25%	0.85	0.8	0.82	0.78
30%	0.86	0.83	0.84	0.81

number of labeled events (e.g., 5% of the total recorded events), the proposed method can still achieve 75% accuracy and provide meaningful help.

As an imbalanced classification task, it is crucial to show that the proposed method can correctly categorize each event type. Therefore, for each event type, several statistical metrics, including recall, precision, F_1 score, and Matthews correlation coefficient (MCC) are utilized to further evaluate the performance of our method with different amounts of labeled data [35]. Specifically, recall is thought of as the percentage of relevant events that are correctly identified. Its dual metric, precision, is defined as the fraction of identified events that are relevant. F_1 score can be considered as the harmonic average of the precision and recall:

$$F_1 = \frac{(\beta^2 + 1) * Prec * Recall}{(\beta^2 * Prec + Recall)} \tag{15}$$

where, β is the precision weight which is set at 1 in this paper. F_1 score ranges in [0,1], where the maximum is reached when $FN = FP = 0$. F_1 score is not defined based on confusion matrix since it is independent from TN. Meanwhile, it is not symmetric for type swapping. In comparison, MCC is a contingency matrix method of calculating the Pearson productmoment correlation coefficient in terms of the entries of confusion matrix:

$$MCC = \frac{TP{*}TN - FP{*}FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{16}$$

MCC ranges in $[-1, 1]$, where 1 shows a perfect event identification, 0 corresponds to the random identification, and -1 indicates total disagreement between estimated labels and actual labels. The average values of these indexes are presented in Table I. It is clear that the values of all metrics are at the same level. This result shows that the proposed method is able to handle the imbalance of the dataset and obtain stable estimation results for different kinds of events.

C. Performance of the Proposed Method With Class Mismatch Problem

To demonstrate the performance of the proposed method with the class mismatch problem, we assume a special case where the utility never records a certain event type, but this event type appears in large numbers in unlabeled events. Specifically, all events belonging to the line outage are first excluded from the labeled dataset and then added to the unlabeled events proportionally. Only the remaining types of events (i.e., normal

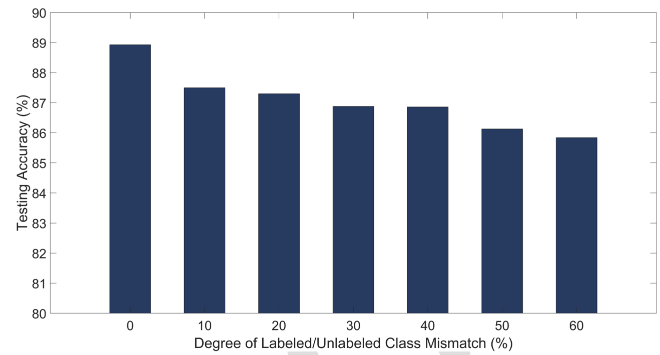


Fig. 8. Event identification accuracy of the proposed method with varying class mismatch degrees.

operation data, XFMR outages, and frequency events) are used for initial model training. As training proceeds, our model is expected to avoid pseudo-labeling hidden line outage events and adding them to the training set, thus preventing performance degradation. Here, we gradually increase the degree of labeled and unlabeled class mismatch degree from 0% to 60% to test the effectiveness of our algorithm, respectively. Note that the degree of labeled/unlabeled class mismatch is obtained by the ratio of the number of line outage events (i.e., unknown events) to the number of other kinds of events (i.e., known events) among the unlabeled events. This degree can be equivalently viewed as the exploration value of unlabeled events. In the extreme case, when this degree is 100%, it means that no unlabeled events should be exploited in model training. The results are presented in Fig. 8. As shown in the figure, it can be found that the accuracy of the algorithm slightly decreases as the degree of class mismatch increases. When unknown events accounted for half of the unlabeled events, the accuracy of our algorithm dropped by roughly 3% (from 89% to 86%). However, in this extreme case, our algorithm still performs better than the supervised learning-based event identification method (i.e., 82%) [13]. These findings corroborate the premise of this study, according to which the performance of the proposed framework diminishes with increasing class distribution mismatch between labeled and unlabeled data but never performs worse than that of the supervised learning method.

D. Method Comparison

Considering that most existing works on event identification rely on unsupervised techniques (i.e., clustering algorithms) to connect unlabeled data and labeled data, We have conducted numerical comparisons with two clustering algorithms (i.e., hierarchical clustering and spectral clustering) previously used for event identification tasks [21], [36], [37]. Moreover, two state-of-the-art semi-supervised classification algorithms, PI model and mean teacher, are included in our comparison experiments to observe whether our models can perform better than previous semi-supervised learning models in the presence of high class mismatch degree [31], [38]. To ensure a fair comparison with unsupervised learning methods, the total number of event types in the set of unlabeled events is unknowable. In other

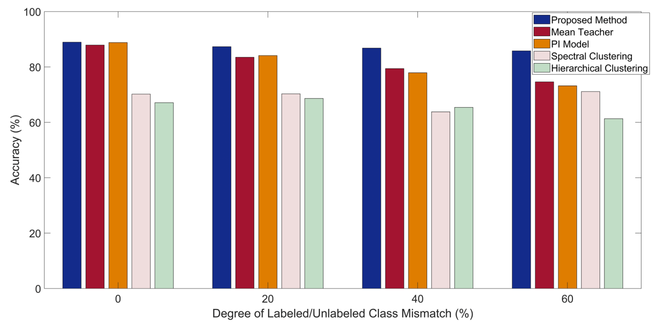


Fig. 9. Comparison results of five event identification methods.

words, the number of clusters is not available. Hence, in the experiments, the Davies-Bouldin validation index is applied to calibrate the unsupervised learning method to find the number of clusters [39]. The identification accuracy is calculated based on the misclassification between the true labels and the clustered labels. Like the last case, all methods are tested with varying class mismatch degrees. The comparison results are demonstrated in Fig. 9. It can be observed that the three semi-supervised learning methods generally outperform unsupervised learning methods, especially in the cases of low mismatch degree. The reason behind this is that the unsupervised learning methods do not use any labeling information, but only the data itself. This makes their results generally poor under the event classification task. Meanwhile, in some tests, we cannot obtain the correct number of clusters in a calibrated manner, which further reduces the accuracy. Among the semi-supervised learning methods, the proposed method performs better than the two state-of-the-art methods, especially when the mismatch degree is high. In some extreme cases (e.g., mismatch degree is 60%), the proposed algorithm still performs better than supervised learning-based methods, but other semi-supervised methods show performance degradation. Note that unsupervised learning models do not suffer from the class mismatch problem, as they do not care about label information.

E. Computational Complexity Analysis

To demonstrate the practical complexity of the proposed algorithm, we conducted the case study on a typical personal computer. Based on our multiple experiments, when the event labels are retained for 25% of the records, the training computation of the proposed model time ranges from 1.7 hours to 1.9 hours. It should be noted that the training time also changes slightly with the volume of labeled data due to the pseudo-labeling process. The proposed method's average test time, based on 1,440 test samples, is roughly 0.8 ms. As a result, in a real grid, our method may deliver estimates in around 0.1 seconds after the PMU measurements arrive at the phase data concentrator after accounting for the communication delay. This is still much faster than the vast majority of heuristic-based methods.

VI. Conclusion

In this paper, we design a novel data-driven method to accurately identify events using a limited number of labeled events and a rich set of unlabeled events. Our approach is built on a semi-supervised learning framework with three event identifiers. By designing a weight function, each classifier can selectively explore unlabeled events to provide additional information about the shape of the decision boundary among different event types. The proposed method can address two main challenges in power system event identification: 1) poor generalization of deep learning models caused by the limited number of labeled events. 2) class distribution mismatch problem between labeled events and unlabeled events caused by event data scarcity. The proposed solution has been successfully tested on an actual Western Interconnection dataset.

Acknowledgment and Disclaimer

This material is based upon work supported by the Department of Energy under Award Number DEOE0000910. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

[1] Y. Ge, A. J. Flueck, D. -K. Kim, J. -B. Ahn, J. D. Lee, and D. -Y. Kwon, "Power system real-time event detection and associated data archival reduction based on synchrophasors," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 2088–2097, Jul. 2015.

[2] D. Kim, T. Y. Chun, S. Yoon, G. Lee, and Y. Shin, "Wavelet-based event detection method using PMU data," *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1154–1162, May 2017.

[3] J. D. L. Ree, V. Centeno, J. S. Thorp, and A. G. Phadke, "Synchronized phasor measurement applications in power systems," *IEEE Trans. Smart Grid*, vol. 1, no. 1, pp. 20–27, Jun. 2010.

[4] G. V. de S. Lopes, G. R. Moraes, D. Issicaba, and D. Dotta, "WAMS-based two-level robust detection methodology of power system events," *Sustain. Energy, Grids Netw.*, vol. 31, 2022, Art. no. 100689.

[5] Y. Liu et al., "Robust event classification using imperfect real-world PMU data," *IEEE Internet Things J.*, early access, doi: 10.1109/JIOT.2022.3177686. Q5

[6] S. Brahma, R. Kavasseri, H. Cao, N. R. Chaudhuri, T. Alexopoulos, and Y. Cui, "Real-time identification of dynamic events in power systems using PMU data, and potential applications–Models, promises, and challenges," *IEEE Trans. Power Del.*, vol. 32, no. 1, pp. 294–301, Feb. 2017.

[7] J. Shi, B. Foggo, and N. Yu, "Power system event identification based on deep neural network with information loading," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5622–5632, Nov. 2021.

[8] M. K. Jena, B. K. Panigrahi, and S. R. Samantaray, "A new approach to power system disturbance assessment using wide-area postdisturbance records," *IEEE Trans. Ind. Informat.*, vol. 14, no. 3, pp. 1253–1261, Mar. 2018.

[9] M. Biswal, S. M. Brahma, and H. Cao, "Supervisory protection and automated event diagnosis using PMU data," *IEEE Trans. Power Del.*, vol. 31, no. 4, pp. 1855–1863, Aug. 2016.

[10] W. Li, M. Wang, and J. H. Chow, "Real-time event identification through low-dimensional subspace characterization of high-dimensional synchrophasor data," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4937–4947, Sep. 2018.

[11] S. S. Negi, N. Kishor, K. Uhlen, and R. Negi, "Event detection and its signal characterization in PMU data stream," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3108–3118, Dec. 2017.

[12] A. Ghasemkhani, Y. Liu, and L. Yang, "Real-time event detection using rank signatures of real-world PMU data," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2022, pp. 1–5.

[13] Y. Yuan, Y. Guo, K. Dehghanpour, Z. Wang, and Y. Wang, "Learning-based real-time event identification using rich real PMU data," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5044–5055, Nov. 2021.

[14] R. Ma, S. Basumallik, and S. Eftekharnejad, "A PMU-based data-driven approach for classifying power system events considering cyberattacks," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3558–3569, Sep. 2020.

[15] Y. Liu et al., "Robust event classification using imperfect real-world PMU data," *IEEE Internet Things J.*, earlt access, doi: 10.1109/JIOT.2022.3177686.

[16] O. P. Dahal and S. M. Brahma, "Preliminary work to classify the disturbance events recorded by phasor measurement units," *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2012, pp. 1–8.

[17] H. Li, Z. Ma, and Y. Weng, "A transfer learning framework for power system event identification," *IEEE Trans. Power Syst.*, vol. 37, no. 6, pp. 4424–4435, Nov. 2022.

[18] S. Liu et al., "Data-driven event detection of power systems based on unequal-interval reduction of PMU data and local outlier factor," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1630–1643, Mar. 2020.

[19] H. Li, Y. Weng, E. Farantatos, and M. Patel, "An unsupervised learning framework for event detection, type identification and localization using pmus without any historical labels," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2019, pp. 1–5.

[20] A. A. Hai et al., "Transfer learning for event detection from PMU measurements with scarce labels," *IEEE Access*, vol. 9, pp. 127420–127432, 2021.

[21] O. P. Dahal, S. M. Brahma, and H. Cao, "Comprehensive clustering of disturbance events recorded by phasor measurement units," *IEEE Trans. Power Del.*, vol. 29, no. 3, pp. 1390–1397, Jun. 2014.

[22] J. Ma, Y. V. Makarov, R. Diao, P. V. Etingov, J. E. Dagle, and E. D. Tuglie, "The characteristic ellipsoid methodology and its application in power systems," *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 2206–2214, Nov. 2012.

[23] H. Li, Z. Ma, Y. Weng, E. Blasch, and S. Santoso, "Structural tensor learning for event identification with limited labels," *IEEE Trans. Power Syst.*, early access, doi: 10.1109/TPWRS.2022.3231262.

[24] Y. Yuan, Z. Wang, and Y. Wang, "Learning latent interactions for event identification via graph neural networks and PMU data," *IEEE Trans. Power Syst.*, vol. 38, no. 1, pp. 617–629, Jan. 2023.

[25] G. Lan-Zhe et al., "Safe deep semi-supervised learning for unseen-class unlabeled data," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 3897–3906.

[26] Z. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005.

[27] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," 2020, *arXiv:2006.05278*.

[28] L. Breiman, "Randomizing outputs to increase prediction accuracy," *Mach. Learn.*, vol. 40, no. 3, pp. 229–242, 2000.

[29] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, *arXiv:1610.02242*.

[30] D. Angluin and P. Laird, "Learning from noisy examples," *Mach. Learn.*, vol. 2, no. 4, pp. 343–370, 1988.

[31] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[32] Y. Fan et al., "Revisiting consistency regularization for semi-supervised learning," in *Proc. DAGM German Conf. Pattern Recognit.*, 2021, pp. 63–78.

[33] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 4331–4340.

[34] A. Martń et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. operating Syst. Des. Implementation*, 2016, pp. 265–283.

[35] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomic.*, vol. 21, pp. 1–13, 2020.

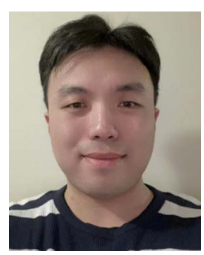
[36] T. Lan, Y. Lin, J. Wang, B. Leao, and D. Fradkin, "Unsupervised power system event detection and classification using unlabeled PMU data," in *Proc. IEEE PES Innov. Smart Grid Technol. Europe*, 2021, pp. 01–05.

[37] Z. Yang, H. Liu, T. Bi, and Q. Yan, "Bad data detection algorithm for PMU based on spectral clustering," *J. Modern Power Syst. Clean Energy*, vol. 8, no. 3, pp. 473–483, May 2020.

[38] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1163–1171.

[39] G. J. Tsekouras, P. B. Kotoulas, C. Tsirekis, E. N. Dialynas, and N. D. Hatziargyriou, "A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers," *Elect. Power Syst. Res.*, vol. 78, pp. 1494–1510, Jun. 2008.

Yuxuan Yuan (Member, IEEE) received the B.S. and Ph.D. degrees in electrical and computer engineering from Iowa State University, Ames, IA, USA, in 2017 and 2022, respectively. He is currently with Electric Power Engineers as s Data Analysis Engineer. His research interests include distribution system condition estimation, PMU and AMI data analysis, and the development and application of machine learning in distribution systems.



Yanchao Wang (Graduate Student Member, IEEE) received the Bachelor of Engineering degree in optical information and technology from the Beijing Institute of Technology, Beijing, China, in 2014. He is currently working toward the Ph.D. degree with Iowa State University, Ames, IA, USA. His research interests include deep learning in power systems, machine learning, and signal processing

Zhaoyu Wang (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, and the M.S. and Ph.D. degrees in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA. He is currently the Northrop Grumman Endowed Associate Professor with Iowa State University, Ames, IA, USA. He is also the Principal Investigator for a multitude of projects funded by the National Science Foundation, Department of Energy, National Laboratories, PSERC, and Iowa Economic Development Authority. His research interests include optimization and data analytics in power distribution systems and microgrids. He was the recipient of the National Science Foundation CAREER Award, Society-Level Outstanding Young Engineer Award from IEEE Power and Energy Society (PES), Northrop Grumman Endowment, College of Engineering's Early Achievement in Research Award, and Harpole-Pentair Young Faculty Award Endowment. He is the Co-TCPC of IEEE PES PSOPE, Chair of IEEE PES PSOPE Award Subcommittee, Vice Chair of PES Distribution System Operation and Planning Subcommittee, and Vice Chair of PES Task Force on Advances in Natural Disaster Mitigation Methods. He is an Associate Editor for IEEE Transactions on Sustainable Energy, IEEE Open Access Journal of Power and Energy, IEEE Power Engineering Letters, and *IET Smart Grid*. He was an Associate Editor for IEEE Transactions on Power Systems and IEEE Transactions on Smart Grid.