Procedure International Journal of Science and Technology
(International Open Access, Peer-reviewed & Refereed Journal)
(Multidisciplinary, Monthly, Multilanguage)
ISSN : 2584-2617 (Online)
Volume- 2, Issue- 1, January 2025
Website- www.pijst.com DOI- https://doi.org/10.62796/pijst.2025

Data-Driven Semi-Supervised Approaches for Event Identification in Large-Scale Power Systems

Anuradha Krishna

Assistant Secretary, SBTE, Department of Science, Technology and Technical Education (Bihar)

ABSTRACT

In this paper, we investigate three traditional SSL techniques for event detection: self-training, transductive support vector machines (TSVM), and graph-based label spreading (LS). Load loss, generation loss, line trip, and bus fault are four important event categories that may be classified using features extracted from synthetic PMU data using modal analysis. By comparing various methods on the South Carolina 500-Bus synthetic network, we find that graph-based LS is the most effective, demonstrating the usefulness of data-driven SSL techniques for detecting events in large-scale power systems. For real-time monitoring and analysis, data-driven methods are becoming crucial due to the growing integration of Phasor Measurement Units (PMUs) and developments in data science. Unfortunately, fully supervised learning methods aren't very effective because it's hard to get enough labeled data, which is a problem because some grid events are rare and unclear. Because of its ability to use both labelled and unlabeled data to enhance performance, semi-supervised learning (SSL) becomes a potent option.

Keywords: Semi-supervised, Event, Identification, Power, Detection.

I.INTRODUCTION

The ever-increasing complexity and interconnection of grid activities in contemporary power systems bring about new possibilities while simultaneously posing formidable obstacles to preserving the stability and dependability of the system. Sensing technologies like Phasor Measurement Units (PMUs) and Supervisory Control and Data Acquisition (SCADA) systems are rapidly expanding, resulting in the continuous generation of massive amounts of data across expansive geographical regions. With this data, we may be able to better understand the grid's operational state and spot abnormalities like faults, load shedding, equipment failures, cyber intrusions, and more. Methods for event detection in the past have mostly used model-based techniques or completely supervised machine learning algorithms, both of which necessitate large, labelled datasets. However, owing to factors such as the transient nature of power systems, the difficulty of manually labeling events, and the rarity of certain events, it is frequently unfeasible to acquire enough labeled event data for real-world applications. The need to find new ways to make use of the mountain of unlabeled data has led to the development of data-driven semi-supervised methods for event recognition.

At the crossroads of supervised and unsupervised learning, semi-supervised learning (SSL) trains predictive models using a small collection of labelled data alongside a big pool of unlabeled data. When it comes to identifying events in power systems, SSL offers a robust foundation for enhancing detection algorithms' accuracy and generalizability, all without depending on massive labeled datasets. This works especially well in systems with a lot of nodes, since it would be impractical to label each occurrence of an event type. Through propagating label information based on the underlying structure in the data distribution, data-driven SSL methods learn representations that can better distinguish between normal and abnormal circumstances. Power system applications are seeing a surge in the use of techniques such graph-based approaches, self-training, co-training, and consistency regularization, which have been implemented in multiple domains.

Due of the spatial and temporal correlations present in grid data, SSL is more useful in power systems. Rapid shifts in voltage, frequency, or current are common indicators of events in power systems, and the effects of these shifts often trickle down through the network. With data-driven SSL approaches, even with few labelled instances, the source and nature of disturbances can be correctly identified by building models that can capture these correlations. With graph-based SSL techniques, for instance, the grid can be represented as a network with nodes standing for measurement locations or substations and edges for electrical connectivity. By propagating labels from a small number of known event locations to nearby nodes, the entire graph can be labeled. Both the precision of event detection and its localization are improved by this, which is essential for prompt operational reactions.

New developments in deep learning have enabled the creation of more complex semi-supervised architectures including contrastive learning frameworks, generative adversarial networks (GANs), and semi-supervised autoencoders, in addition to graph-based models. By training on massive amounts of unlabeled data, these techniques improve the model's discriminatory power against subtle event signatures. One example is the use of semi-supervised autoencoders, which may be trained to mimic typical system behavior and then used to detect abnormalities when they deviate from the norm. A combination of anomaly detection and classification is achieved when a small number of labelled events is used to train the model. This allows the model to learn to identify known event kinds.

The promise of data-driven semi-supervised techniques isn't without its challenges when applied to real-world power systems. Problems with data quality and heterogeneity, imbalance across classes, changing grid topologies, and noise are major obstacles. In addition, operator confidence and acceptance depend on SSL models' interpretability and explainability. Making judgments using machine learning models requires transparency and justification since electricity systems are safety-critical infrastructures. Therefore, there has been a recent uptick in efforts to create SSL models that can be understood by humans and incorporate domain expertise to direct the training process.

The incorporation of SSL techniques into preexisting operational workflows is another critical factor to think about. Accuracy, low latency, and resilience in the face of data loss or connection delays are all essential for real-time event detection. This is why researchers are looking at hybrid frameworks that use semi-supervised learning in conjunction with more conventional rule-based systems or models based on physical principles to guarantee dependability in a wide range of operational scenarios. Also, these models can't be implemented without scalable infrastructures that can handle streaming data and adjust to changing system dynamics.

II.REVIEW OF LITERATURE

Yuan, Yuxuan et al., (2023) In order to build real-time event detection models for transmission networks, this study explores the utilization of phasor measurement unit (PMU) data in conjunction with deep learning approaches. There is a significant chance to achieve decarbonization with the increasing penetration of distributed energy resources, but there are also problems in systematic situational awareness. A large number of state-of-the-art classifiers can tackle the power event identification problem when there is enough manually recorded event labels and high-resolution PMU data. Gathering extremely high-quality event labels, meanwhile, might get pricey in actual grids. It is common for utilities to have a high volume of event records that lack detailed information, sometimes known as unlabeled events. We provide a new approach based on semi-supervised learning to fill this knowledge gap; it uses information from large amounts of unlabeled events to train event classifiers that were previously trained with a small number of labelled events. To rephrase, our method achieves the same level of accuracy with a tiny amount of labeled data as existing data-driven methods, but with far less effort. The performance degradation induced by a mismatch between the training set and real applications in terms of class distribution is discussed and addressed in this paper. In particular, this approach gradually increases the size of the training dataset and probes the worth of unlabeled events using the pseudo-labeling technique. In order to lessen the effects of a mismatch in class distribution and stop performance from dropping, a safe learning mechanism is also created. Our model uses a thorough evaluation approach to selectively incorporate unlabeled events during model training, based on the proposed safe learning mechanism. In order to confirm that the suggested strategy works, numerical investigations were conducted on a large PMU dataset.

Sen, Debarshi et al., (2019) Guided ultrasonic waves (GUWs) have been a prominent approach for SHM of pipelines for more than 30 years. When compared to more conventional vibration-based methods, GUWs excel at detecting cracks and corrosion over a sufficient length of pipeline, which are examples of minor damages. Unfortunately, model-based approaches are computationally too expensive due to the system's highly complicated physics. In these cases, data-driven methods grounded in statistical learning algorithms work far better. We provide two data-driven methods for pipe damage identification in this research, one using a supervised learning approach and the other using a semi-supervised one. In addition to avoiding model-based techniques, the suggested methods help save maintenance expenses by lowering the number of sensors placed. Using an algorithm based on hierarchical clustering, the semi-supervised learning method can identify damage. Within a multinomial logistic regression framework, the damage localization is carried out by the supervised learning-based method. The suggested algorithms are proven correct by collecting guided wave responses from experimental pipes in a pitch-catch arrangement with inexpensive piezoelectric transducers. Using a mix of two sensors, we show that our data-driven methods can reliably identify and pinpoint cracks in two cast-iron pipes of varying lengths.

Zhou, Yuxun et al., (2017) Unconventional sources of energy and loads, like electric vehicles, controlled loads, and distributed renewable resources, have been finding their way into the power grid in growing numbers. In order to diagnose and control the system, specialists in high-resolution monitoring and agile decision-support methods are needed due to the induced dynamic and stochastic power flow. In order to detect events in power distribution networks, this research delves into the use of data from micro-phasor measurement units (ìPMUs). Hidden structure semi-supervised machine (HS3 M) is a new data-driven event detection approach. To bridge the gap between supervised learning, semi-supervised learning, and learning with hidden structures, HS3 M mixes unlabeled and partly labeled data in a broad margin learning objective. It only requires partial expert knowledge. A new global optimization approach, the parametric dual optimization procedure, is defined by its equivalence to a concave programming in order to optimize the non-convex learning objective. Lastly, the suggested approach is tested on a real distribution feeder that has íPMUs installed. The outcome confirms that the learning-based event detection framework is effective and might be used as a fundamental technique for power system dependability and security.

Zhou, Yuxun et al., (2016) Recently, data-driven event detection has proven to be advantageous in handling complex systems, particularly those exhibiting substantial stochastic and dynamic behavior. This is made possible by the advent of data gathering and processing technologies like sensor networks and machine learning. Traditional approaches, on the other hand, rely on supervised learning frameworks and costly, often impracticable, expert labeling throughout the learning phase. Using just partial expert knowledge, we present a new data-driven event detection system called Hidden Structure Semi-Supervised Machine (HS3M). To fill the void between supervised, semi-supervised, and hidden structure learning, the central idea is to merge partially labeled and unlabeled data in a large margin learning objective. The problem becomes non-convex when more learning terms are added, which causes difficulties. We design a new global optimization technique, Parametric Sub-Gradient Descent (PSGD), to maximize the learning objective by demonstrating that the parameterized dual problem has local explicit solutions and that the associated optimality is convex in hidden variables. Power distribution network event detection is the target of the suggested method, and the outcome validates the efficacy of HS3M and the novel global optimization algorithm.

III.SEMI-SUPERVISED EVENT IDENTIFICATION: MODEL LEARNING AND VALIDATION

Our three-step process for testing semi-supervised techniques is as follows: (i) passing off some of the training set's unlabeled samples as labelled ones, using a combination of the two types of samples, $D_{M}^{(s)}$,(ii) determining how well a classifier performs on the validation set after training it with the mixed labeled and pseudo-labeled data, 5_{5I} .

In Figure 1 we can see the big picture of the strategy that has been suggested. The given model is semi-supervised F_1 and a classifier F_2, To begin, we take the labelled samples from the training set's 5Xth fold and 5Pth split. To find the model's hyperparameters, we do grid search using these labelled samples. F_1 and F_2, denoted as θ_1^* and θ_2^*. (Please be aware that these hyperparameters will vary depending on 5X and 5^.) Afterwards, we make use of the event feature matrix and the matching label matrix in the to provide the unlabeled samples pseudo-labels by means of F_1. Making use of the labelled and pseudo-labeled samples that were collected, , we then use model F_2 " {SVMR, SVML, GB, DT, 5>NN} to label the occurrences in the validation dataset

D_{5K}. The models that serve as F_1 in this process will be detailed in the sections that follow.

1. Self-training: Using unlabeled data to train supervised classifiers has been successful with self-training. In self-training, the model is trained repeatedly using unlabeled data that have been pseudo-labeled based on the model's predictions. We train a model F_1 from the labeled data in the 5>NN model space, precisely for every given base classifier. Then using the learned model, we predict the labels for each unlabeled samples to obtain the augmented labeled and pseudo-labeled samples, denoted as Algorithm 1 outlines the steps involved in this procedure. Note that the parameter 5_{5H} in this algorithm specifies the number of unlabeled samples (among thesamples) that will be assigned pseudo-labels in each iteration.

Algorithm 1 Self-Training (for a given k, q, s, and r).

```
Input: D_M^(s)
Output: D̂_M^(s)
Initialize: [f : t] = [1 : δ_U]          ▷ from sample f to sample t
    X̃_L ← X_L, Ỹ_L ← Y_L, X̃_U ← X_U[f : t]
while t ≤ n_U^(s) do
    F_1 : Ỹ_L ← X̃_L                     ▷ Learning the model
    Ŷ_U = F_1(X̃_U)                      ▷ pseudo-labeling
    X̃_L ← [X̃_L^T, X̃_U^T]^T,  Ỹ_L ← [Ỹ_L^T, Ŷ_U^T]^T   ▷ Augmentation
    f ← f + δ_U,   t ← t + δ_U
    if t > n_U^(s):
        t = n_U^(s)
    X̃_U ← X_U[f : t]
end while
Ŷ_M ← Ỹ_L
Return: D̂_M^(s) = {X_M, Ŷ_M}
```

2. TSVM, or Transductive Support Vector Machine, is a variant of SVM that overcomes the problem of sparse labeled data in classification applications. The challenge of optimizing the TSVM is presented by

$$\min_{\mathbf{w},b,\eta,\zeta,z} \quad C\left[\sum_{i\in\mathcal{I}_L}\eta_i + \sum_{j\in\mathcal{I}_U}\min(\zeta_j, z_j)\right] + \|\mathbf{w}\|^2$$

Subject to-

$$y_i(\mathbf{w}^T x_i - b) + \eta_i \geq 1, \quad \eta_i \geq 0, \quad i \in \mathcal{I}_L \qquad (5b)$$
$$\mathbf{w}^T x_i - b + \zeta_j \geq 1, \quad \zeta_j \geq 0, \quad j \in \mathcal{I}_U \qquad (5c)$$
$$-(\mathbf{w}^T x_i - b) + z_j \geq 1, \quad z_j \geq 0, \quad j \in \mathcal{I}_U \qquad (5d)$$

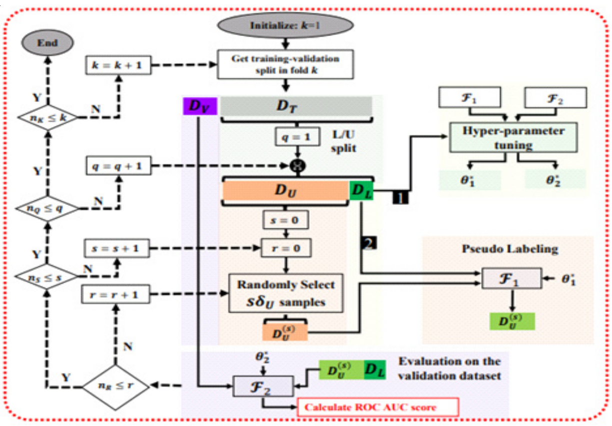

Figure 1: Overview of the proposed semi-supervised pipeline.

The direction of the decision boundary is represented by w " !5Q, while the bias (or intercept) term is denoted by 5O " !. It applies two restrictions (i.e., (5c) and (5d)) to every training dataset sample, treating them as if they were members of either class and computing the misclassification error accordingly. The goal of the objective function is to reduce the minimum of the misclassification errors (z) and maximize the margin (w), while simultaneously minimizing the misclassification error of labeled samples (5). For accurate classification, the TSVM may use both labelled and unlabeled samples. The unlabeled samples are then given pseudo-labels. In order to keep things brief, we direct readers to for more information.

LS, or Label Spreading: Models that are graph-based and semi-supervised (GSSL) include label spreading (LS). Building a network with weighted connections that show similarities and nodes that represent samples is the key to inferring labels for unlabeled data. Think about a graph. 5:5@ = ($V_{5@}$, $W_{5@}$) which is constructed over the combined labeled and unlabeled training set.

Each sample, 5e , "$_{5V}$" $I_{5?}$ *" I_{5H} , can be represented as a node in a graph. For the resulting graph, we define the edge weights matrix as . Defining 57_{5V5W} = $5e_{5V}$ "$5e_{5W}{}^{2}$, the $5V^{th}$ row and $5W^{th}$ column of $W_{5@}$, denoted as $5d_{5V5W}$, can be obtained as $5d_{5V5W}$ = exp("57_{5V5W}"25^{2}) if 5V `" 5W, and $5d_{5V5V}$ = 0. According to this edge weight metric, the weights assigned to adjacent pairs of samples will be higher. By using weighted edges, which convey the idea of similarity, the LS method enables the transmission of labels from labeled to unlabeled samples, capitalizing on the classical intuition that nearby samples often share labels. We lay down the procedures of the LS method in Algorithm 2. Line 7 of Algorithm 2 captures the update rule, which updates the labels of both the labeled and unlabeled samples. Specifically, for the labeled samples, this update incorporates information from the neighbors (first term) while retaining the original label (second term). The relative importance of neighbor-derived information and the original label information from the sample is determined by the parameter 5.

IV.RESULTS AND DISCUSSION

First, we follow the steps in Section II to create eventful synthetic PMU data, and then we test out several semi-supervised learning techniques. The 500-Bus System in South Carolina was the basis for our simulations. For the time being, we let the system run normally. $5a_{5S}$ = 1 second and then we immediately apply a disturbance.

Algorithm 2 Label spreading (for a given k, q, s, and r).

1: **Input:** $G = (\mathcal{V}, \mathcal{W}) \leftarrow \mathbf{D}_M^{(s)} = \{\mathbf{X}_M, \mathbf{Y}_M\}$
2: **Output:** $\widehat{\mathbf{D}}_M^{(s)}$
3: **Compute:** $\mathcal{D}_{ii} = \sum_j w_{ij}, \quad \forall i \in \mathcal{I}_L \cup \mathcal{I}_U$
4: **Compute:** $\mathbf{Z} = \mathcal{D}^{-1/2} \mathcal{W}_M \mathcal{D}^{-1/2}$
5: **Initialize:** $\begin{bmatrix} \mathbf{Y}_L|_{t=0} \\ \mathbf{Y}_U|_{t=0} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{Y}_L \\ \mathbf{Y}_U \end{bmatrix}$
6: **while** $\begin{bmatrix} \mathbf{Y}_L|_t \\ \mathbf{Y}_U|_t \end{bmatrix}$ converges **do** ▷ *Based on some threshold*
7: $\begin{bmatrix} \mathbf{Y}_L|_{t+1} \\ \mathbf{Y}_U|_{t+1} \end{bmatrix} \leftarrow \alpha \mathbf{Z} \begin{bmatrix} \mathbf{Y}_L|_t \\ \mathbf{Y}_U|_t \end{bmatrix} + (1-\alpha) \begin{bmatrix} \mathbf{Y}_L|_{t=0} \\ \mathbf{Y}_U|_{t=0} \end{bmatrix}$
8: $t \leftarrow t + 1$
9: **end while**
10: $\widehat{\mathbf{Y}}_M \leftarrow \begin{bmatrix} \mathbf{Y}_L|_t \\ \mathbf{Y}_U|_t \end{bmatrix}$
11: **Return:** $\widehat{\mathbf{D}}_M^{(s)} = \{\mathbf{X}_M, \widehat{\mathbf{Y}}_M\}$

We then run the simulation for an additional 5a5' = 10 seconds, and record the resulting eventful measurements at the PMU sampling rate of 30 samples/sec. The 5aclr for the BF events is 5 cycles (H" 0.083 seconds). We assume that 95 buses (which are chosen randomly) of the Carolina 500-bus system are equipped with PMU devices and extract features for each such bus from the $5I_{5Z}$, $5I_{5N}$, and 59 channels. We thus collect 5A = 300 samples after the start of an event for each channel. We use the modal analysis methodology as outlined in our recent prior work to extract features using modal analysis. In total, we simulated 1827 events including 500 LL, 500 GL, 500 LT, and 327 BF events.

We use the area under the curve (AUC) of the receiver operator characteristic (ROC) to objectively assess and compare the performance of different semi-supervised learning methods across different situations. The accuracy of categorization for various discriminating thresholds may be characterized using this measure. One way to measure a classifier's performance is by looking at its ROC AUC value, which may be anywhere from 0 to 1. Classification accuracy improves as the AUC gets closer to 1. By measuring the ROC-AUC score for predicting event classes inside the hold-out fold, we may determine how well a certain classifier F2 performs for a given set of parameters 5X, 5P, 5f, and 5S. The model trained from the augmented labeled and pseudo-labeled data, derived using the pseudo-labeling model F1, and is the basis of this assessment.

We compare different semi-supervised models by looking at their average, 5th, and 95th percentile area under the curve (AUC) scores, which are based on how well the assigned pseudo-labels worked on the unlabeled samples. We also measure how well a generalizable model predicted the labels of validation samples after using the assigned pseudo-labels. For robustness, we primarily aim for the 5th percentile of the AUC values as it gives a (almost) worst-case metric across various initial labeled and unlabeled sample choices. In other words, it doesn't matter whether the initial set of labeled and unlabeled samples are unfavorable; a strategy that leads to accurate findings in the 5th percentile is likely to be the best option. We examine two separate methods within this context to guarantee a fair comparison of different transductive and inductive semi-supervised approaches:

- **Approach 1 (Inductive semi-supervised setting):**

$F_1 \in$ {SVMR, SVML, GB, DT, 5>NN} represents the base classifier utilized in self-training for pseudo-labeling, and the same type of classifier will be used as F_2.

Approach 2 (Transductive semi-supervised setting): $F_1 \in$ {TSVM, LS} represents a semi-supervised method used for pseudo-labeling, and $F_2 \in$ {SVMR, SVML, GB, 5>NN}.

As part of our assessment, we use 5[5> = 10 folds and 5[5D = 30 random divides of the training data into subsets with labels and those without labels. Using the labelled training data inside each fold, we hyperparameter tune the models, as mentioned in Sec. IV. You may find the values of the model's hyperparameters and other simulation parameters in Tables 1 and 2, respectively. The 5 value that was acquired from the hyper parameter adjustment of the SVMR model is used for the LS model. Various classifiers' comparative performance is shown in Figure 2 across different semi-supervised models (self-training, TSVM, and LS). These models include SVML, SVMR, 5>NN, DT, and GB. The results show that compared to the results obtained by the self-training and TSVM methods, the incorporation of more unlabeled samples and the use of LS for pseudo-labeling provide better results. On top of that, the LS algorithm reliably makes all classifiers better. Each semi-supervised model's performance is further detailed in the sections that follow.

Approach 1- Inductive semi-supervised setting

Here order to forecast the labels of validation samples, the simulation results for the 5th percentile of the AUC scores of the SVML, SVMR, 5>NN, DT, and GB classifiers are shown here.

Table 1: Parameters used in semi-supervised event identification

Parameter	Description	Value
n_D	Total no. of samples	1827
n_K	No. of folds	10
n_T	No. of training samples	1644
n_V	No. of validation samples	183
n_Q	No. of random splits of training samples into labeled and unlabeled samples	20
$(B_{\min}, B_{\max})$	Class balance range in the labeled samples	(0.2, 0.8)
n_L	No. of labeled samples	24
n_U	No. of Unlabeled samples	1620
δ_U	No. of unlabeled samples in each step	100
nS	Total No. of steps	18
n_R	No. of random selection of $n^{(s)}$ samples at each step U	10

Table 2: Values used for hyperparameter tuning of the models in semi-supervised event identification

Model	Hyperparameter	Values
KNN	No. of neighbors in KNN	2, 4, 6, 8, 10
SVML	Regularization parameter	logspace(10^{-3}, 10^{2}, 10)*
SVMR	γ in RBF kernel	logspace(10^{-3}, 10^{2}, 10)
	Regularization parameter	logspace(10^{-3}, 10^{2}, 10)
DT	Maximum depth	3, 5, 7
GB	No. of estimators (boosting stages)	50, 100, 150, 200
	Maximum depth	3, 5, 7

In Figure 2, it is evident that the self-training approach performs poorly when employing a small number of labeled samples with SVMR, SMVL, and 5>NN base classifiers. Additionally, event identification accuracy is not guaranteed to increase when GB and DT are used as basis classifiers. The difference between the original selection of labeled samples and the pseudo-labels is the primary cause of this. Accumulating mistakes is a possible outcome of training using biased and untrustworthy pseudo-labels. Essentially, this pseudo-label bias becomes worse for classes with worse behavior, such when the distribution of labeled examples doesn't match the distribution of unlabeled data, and it becomes worse as self-training goes on.

There is a striking sensitivity to the distribution of labelled and unlabeled data in self-training that uses SVML or SVMR as classifiers. These methods have difficulty producing reliable pseudo-label assignments because of the restriction of using a small number of labeled samples. Nevertheless, when we increase the amount of unlabeled data, self-training using 5>NN as the basic classifier still performs worse than SVML and SVMR examples. It is clear that self-training with DT and GB basis classifiers does not improve their performance with increasing the amount of unlabeled examples, even if these base classifiers show more robust performance than other kinds.

Approach 2- Transductive semi-supervised setting

As shown in Figure 3, the second technique's simulation results use TSVM as the semi-supervised method for pseudo-labeling. The particulars of the dataset and the method's sensitivity to the distribution of labelled and unlabeled samples may explain TSVM's poor performance. It is possible that the TSVM will have difficulty correctly capturing the distribution if the samples are imbalanced or show complicated patterns. Consequently, it may provide false pseudo-labels. Moreover, it is clear that the TSVM algorithm's integration of pseudo-labels is very sensitive, even though it improves the overall performance of SVML and SVMR compared to the same models using pseudo-labels from the self-training algorithm that incorporates SVMR and SVML. Because of this sensitivity, the accuracy of assigned pseudo-labels is still very dependent on the original distribution of labeled and unlabeled samples, which is especially noticeable when looking at the 5% AUC scores. This issue is also shown in the declining efficiency of the 5>NN, GB, and DT classifiers, which, shockingly, declines even more than when they are used as base classifiers in the self-training system.

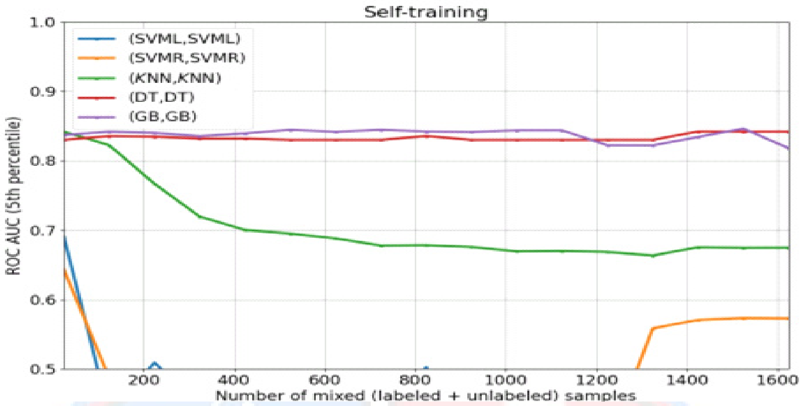

Figure 2: The 5th percentile of AUC scores for different classifiers using pseudo-labels obtained

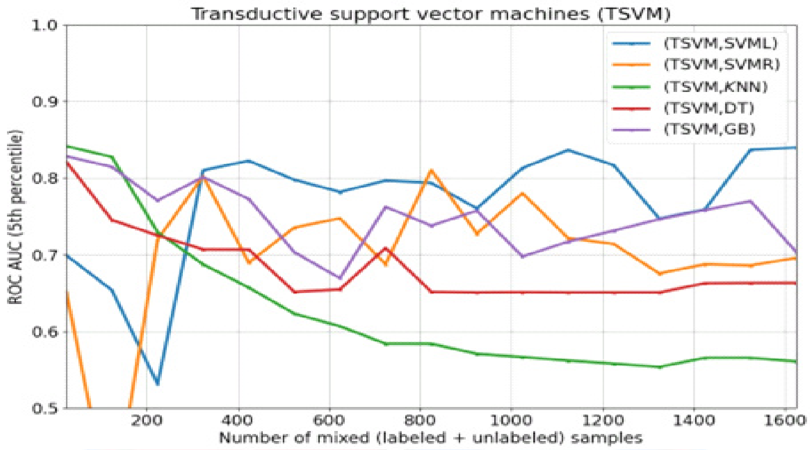

Figure 3: The 5th percentile of AUC scores for different classifiers using pseudo-labels obtained from TSVM,

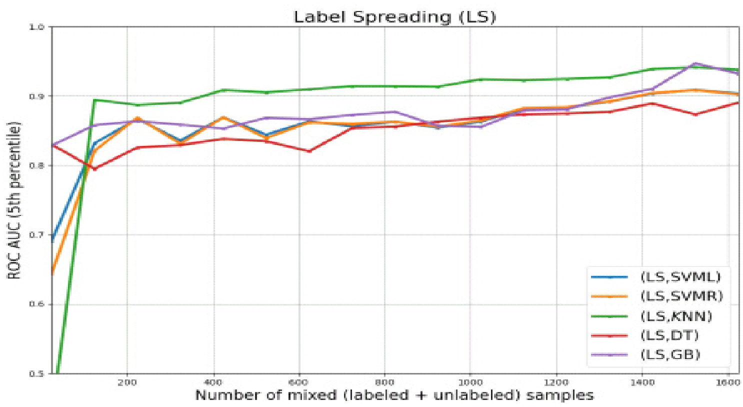

Figure 4: The 5th percentile of AUC scores for different classifiers using pseudo-labels obtained from LS.

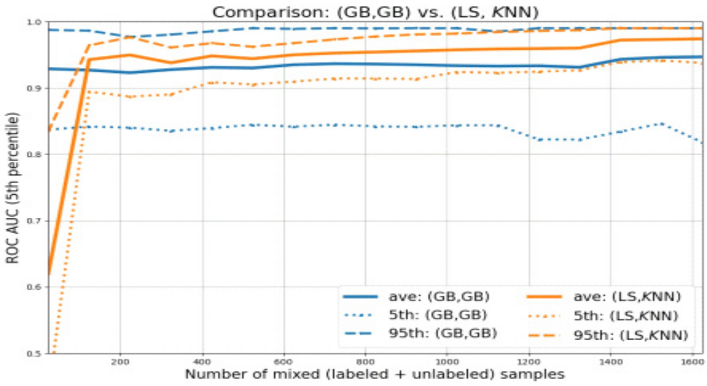

Figure 5: The 5th percentile of AUC scores for different classifiers using pseudo-labels obtained from Comparison between (GB, GB) and (LS, 5>NN) in terms of average, 5 th, and 95th percentile of AUC scores.

Figure 4 shows that even when compared to the self-training and TSVM methods, using the augmented labeled and pseudo-labeled data from LS greatly improves the event recognition performance. In addition, as the amount of unlabeled samples increases, the event identification job performs better. This is important since labeled eventful PMU data is often sparse in real-world scenarios. The LS technique outperforms self-training and TSVM in comparison because it uses labeled and unlabeled samples alike when assigning pseudo-labels, taking use of both types of data. Although the average performance remains relatively same, we see that LS enhances the 5th percentile line with additional unlabeled examples for some classifiers (namely GB and DT). However, as seen in Figure 5, the $5>_{NN}$ classifier exhibits an improvement in the average, 5th, and 95th percentile lines as the number of unlabeled data increases. It seems that LS with $5\sim_{NN}$ is the top classifier overall.

V. CONCLUSION

Finally, event detection in large-scale power systems is heading in a revolutionary new path thanks to data-driven semi-supervised learning techniques. The lack of tagged data is a major obstacle to power system monitoring, but these technologies provide a convincing remedy. Even when minimal supervision is present, semi-supervised models are able to identify and categorize events more accurately and robustly by making good use of large volumes of unlabeled operational data. This is why they are so useful for finding new or unusual incidents that aren't in the historical records. In addition, for better event localization and characterisation, semi-supervised methods can grasp the time- and space-dependent nature of power grid data. Because of their adaptability, they may be used in conjunction with current sensor networks and modified to fit changing grid designs. Unfortunately, these methods won't work in the real world until we fix the problems with interpretability, data quality, and model flexibility. The significance of smart, scalable, and data-efficient solutions is on the increase due to the ever-changing energy environment caused by renewables, dispersed resources, and cyber-physical threats. An important component of smart and resilient power system operations of the future will be semi-supervised learning, which is backed by developments in machine learning and the integration of domain knowledge.

Author's Declaration:

The views and contents expressed in this research article are solely those of the author(s). The publisher, editors, and reviewers shall not be held responsible for any errors, ethical misconduct, copyright infringement, defamation, or any legal consequences arising from the content. All legal and moral responsibilities lie solely with the author(s)

REFERENCES-

1. Yuan, Yuxuan & Wang, Yanchao & Wang, Zhaoyu. (2023). A Data-Driven Framework for Power System Event Type Identification via Safe Semi-Supervised Techniques. IEEE Transactions on Power Systems. PP. 1-12. 10.1109/TPWRS.2023.3266153.
2. Sen, Debarshi & Aghazadeh, Amirali & Mousavi, Ali & Nagarajaiah, Satish & Baraniuk, Richard & Dabak, Anand. (2019). Data-driven semi-supervised and supervised learning algorithms for health monitoring of pipes. Mechanical Systems and Signal Processing. 131. 524-537. 10.1016/j.ymssp.2019.06.003.
3. Zhou, Yuxun & Arghandeh, Reza & Konstantakopoulos, Ioannis & Abdullah, Shayaan & Spanos, Costas. (2016). Data-driven event detection with partial knowledge: A Hidden Structure Semi-Supervised learning method. 5962-5968. 10.1109/ACC.2016.7526605.
4. Zhou, Yuxun & Arghandeh, Reza & Spanos, Costas. (2017). Partial Knowledge Data-Driven Event Detection for Power Distribution Networks. IEEE Transactions on Smart Grid. PP. 1-1. 10.1109/TSG.2017.2681962.
5. Michele Di Santo, Alfredo Vaccaro, Domenico Villacci, and EugenioZimeo. A distributed architecture for online power systems securityanalysis. IEEE Transactions on Industrial Electronics, 51(6):1238–1248,2004
6. Ahmed Naufal AL-Masri, Mohd Zainal Abidin Ab Kadir, Hashim Hizam,and Norman Mariun. A novel implementation for generator rotor anglestability prediction using an adaptive articial neural network applicationfor dynamic security assessment. IEEE Transactions on Power Systems,28(3):2516–2525, 2013
7. Louis Wehenkel. Machine learning approaches to power-system security assessment. IEEE Expert, 12(5):60–72, 1997.
8. Pudi Sekhar and Sanjeeb Mohanty. An online power system static securityassessment module using multi-layer perceptron and radial basis functionnetwork. International Journal of Electrical Power & Energy Systems,76:165–173, 2016
9. C Fatih Kucuktezcan and VM Istemihan Genc. A new dynamic securityenhancement method via genetic algorithms integrated with neural network based tools. Electric Power Systems Research, 83(1):1–8, 2012.
10. Emmanouil M Voumvoulakis and Nikos D Hatziargyriou. A particleswarm optimization method for power system dynamic security control.IEEE Transactions on Power Systems, 25(2):1032–1041, 2009.
11. FengJi Luo, ZhaoYang Dong, Guo Chen, Yan Xu, Ke Meng, YingYingChen, and KitPo Wong. Advanced pattern discovery-based fuzzy clas-sication method for power system dynamic security assessment. IEEETransactions on Industrial Informatics, 11(2):416–426, 2015.
12. Nikita V Tomin, Victor G Kurbatsky, Denis N Sidorov, and Alexey VZhukov. Machine learning techniques for power system security assess-ment. IFAC-PapersOnLine, 49(27):445–450, 2016.
13. Ali Sharifzadeh, Mohammad Taghi Ameli, and Sasan Azad. Powersystem challenges and issues. Application of Machine Learning and DeepLearning Methods to Power System Problems, pages 1–17, 2021.
14. Ali Mollaiee, Mohammad Taghi Ameli, Sasan Azad, Morteza Nazari-Heris, and

Somayeh Asadi. Data-driven power system security assessmentusing high content database during the covid-19 pandemic. InternationalJournal of Electrical Power & Energy Systems, 150:109077, 2023.

15. Jochen L Cremer, Ioannis Konstantelos, Simon H Tindemans, and GoranStrbac. Data-driven power system operation: Exploring the balancebetween cost and risk. IEEE Transactions on Power Systems, 34(1):791–801, 2018.
16. Lipeng Zhu and David J Hill. Data/model jointly driven high-qualitycase generation for power system dynamic stability assessment. IEEETransactions on Industrial Informatics, 18(8):5055–5066, 2021.
17. Lipeng Zhu, David J Hill, and Chao Lu. Semi-supervised ensemble learn-ing framework for accelerating power system transient stability knowledgebase generation. IEEE Transactions on Power Systems, 37(3):2441–2454,2021.
18. Kegang Hao and Qun Wan. Importance sampling based direct maximumlikelihood position determination of multiple emitters using nite mea-surements. Signal Processing, 186:108111, 2021.

Cite this Article-

"Anuradha Krishna", "Data-Driven Semi-Supervised Approaches for Event Identification in Large-Scale Power Systems", Procedure International Journal of Science and Technology (PIJST), ISSN: 2584-2617 (Online), Volume:2, Issue:1, January 2025.

Journal URL- https://www.pijst.com/

DOI- https://doi.org/10.62796/pijst.

Published Date- 02/01/2025