Advanced Semi-Supervised Learning With Uncertainty Estimation for Phase Identification in Distribution Systems

Kundan Kumar, Kumar Utkarsh, Jiyu Wang, and Harsha Vardhana Padullaparti
National Renewable Energy Laboratory, Golden, CO, USA

Background and Motivation

Accurately identifying the ***phase connectivity of customers*** in a distribution system is crucial for system ***efficiency and advanced grid operations***. However, utilities face ***key challenges*** in identifying phase connectivity of customers, They:

- need to send ***field crews to manually check*** for phase connectivity of customers
- need to ***update*** the phase connectivity database after every ***outage*** restoration
- need to ***update*** the phase connectivity database every time a ***new customer*** is added to the system

To overcome these challenges, automated phase identification methods using ***supervised learning*** have been developed. However, their performance typically ***suffers*** because:

- they often require ***substantial labeled ground truth*** training data
- their performance drops significantly with ***limited labeled data***

Proposed SSL Framework

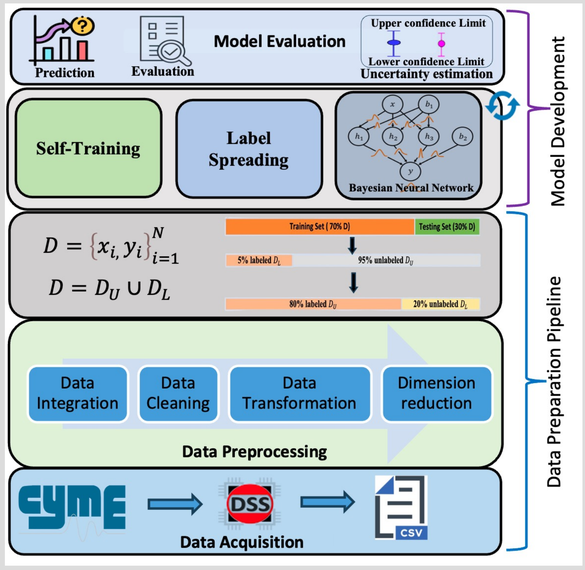

***Fig. 1**: Proposed SSL framework for utility AMI datasets*

Test System & Parameters

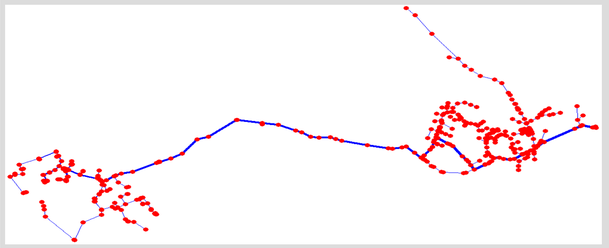

***Fig. 2**: Network topology of the selected distribution feeder.*

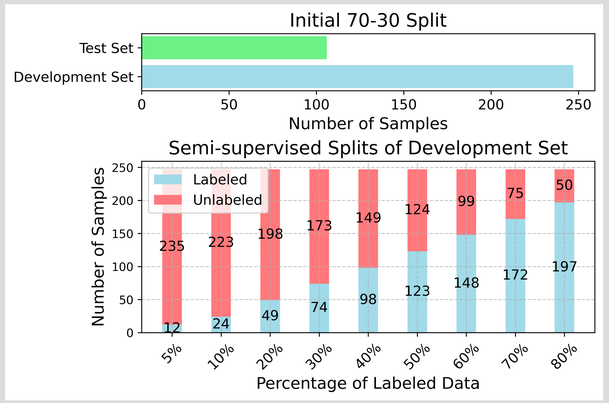

***Fig. 3**: Overview of different data partitions for training and testing.*

Modeling & Implementation of SSL Algorithms

Modeling and implementation require ***preprocessing*** of datasets, ***training*** the SSL algorithms, and making ***predictions*** for the phase identification, along with ***uncertainty estimation***. An overview of the approach is shown in ***Algorithm 1***.

- Using the utility dataset *D*, a filtered dataset *D'* is created by removing missing values, timestamps, and anomalies
- After generating *D'*, we extract the ***feature set*** *F* for training

$$F = \{R_0, X_0, R_1, X_1, P, \text{Max } V, \text{Min } V, \text{AVG } V\}$$

- The dataset *D'* is then ***split*** into 70% for training and 30% for testing
- Within the training set, we create ***two subsets***: D_L with known phase assignments and D_U with unknown assignments. D_L is used for initial model training, while D_L is used to predict pseudo-labels based on the learned model
- The training set is further divided into increments of 5%, 10%, and up to 80% for D_L, with the remaining data forming D_L

Algorithm 1 Semi-Supervised Phase Identification

Input: Dataset $D = (x_i, y_i)_{i=1}^{N}$, Label percentages, P
Output: Accuracies, Predictions, Uncertainties
Filter D to D' and extract features X, labels Y
Split D' into D_{dev}, D_{test} with 70:30 ratio
for $p \in P$ **do**
  Select $n = |D_{dev}| \times (p/100)$ labeled samples
  Form $D_{labeled}, D_{unlabeled}$ from D_{dev}
  Create $X_{semi} = X_{labeled} \cup X_{unlabeled}$
  Set $y_{semi} = [y_{labeled}, -1, ..., -1]$
  Run self-training, label spreading on (X_{semi}, y_{semi})
  Run BNNs on $(X_{labeled}, y_{labeled})$
  Evaluate all methods on D_{test}
end for
return Results for each label percentage P

In ***SSL***, the goal is to ***use both labeled and unlabeled data*** to develop a classifier *f* : → *{A,B,C}* that effectively predicts phase connectivities. The learning ***objective*** is defined as

$$\min_f \left(\frac{1}{n_L} \sum_{i=1}^{n_L} \mathcal{L}(f(\mathbf{x}_i), y_i) + \lambda \cdot \mathcal{R}(f, \mathcal{D}_U) \right)$$

- where $\mathcal{L}$ is the supervised loss (e.g., cross-entropy), $\mathcal{R}$ is the unsupervised regularization term, and *λ* is a hyperparameter that balances the contribution of the supervised and unsupervised components.

We then run ***three SSL algorithms***:

- ***Self-Training With Ensemble Multilayer Perceptron Classifiers*** – an approach to enhance the labeled dataset through iterative pseudo-labeling using an ensemble of MLP classifiers
- ***Label Spreading Classifiers*** – a graph-based SSL technique that spreads labels across similar data points
- ***Bayesian Neural Networks*** – a probabilistic approach to understanding predictions by estimating epistemic and aleatoric uncertainties

Numerical Results

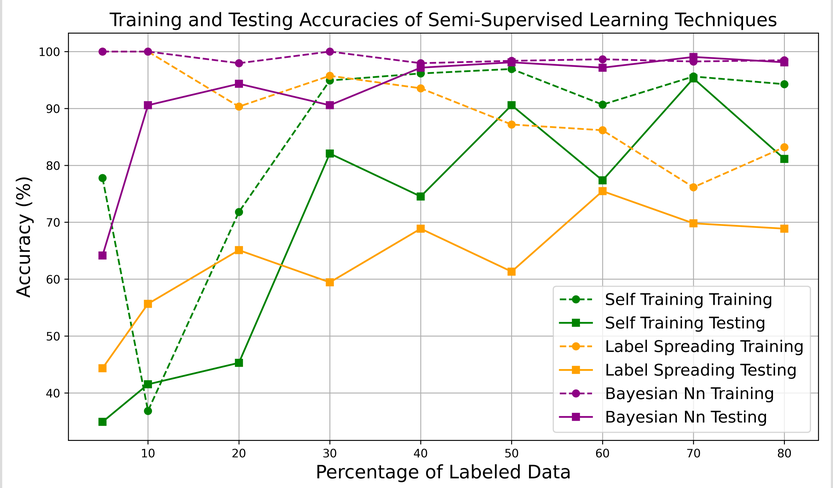

***Fig. 4**: Comparison of different SSL algorithms.*

Ground Truth Percentage	Self Training (Accuracy)	Label Spreading (Accuracy)	BNNs (Accuracy)
5%	34.91 ± 0.11	44.34 ± 0.16	64.15 ± 0.14
10%	41.51 ± 0.12	55.66 ± 0.13	90.57 ± 0.11
20%	45.28 ± 0.11	65.09 ± 0.11	94.34 ± 0.10
30%	82.08 ± 0.12	59.43 ± 0.09	90.57 ± 0.09
40%	74.53 ± 0.11	68.87 ± 0.09	97.17 ± 0.07
50%	90.57 ± 0.13	61.32 ± 0.08	98.11 ± 0.06
60%	77.36 ± 0.12	75.47 ± 0.08	97.17 ± 0.06
70%	95.28 ± 0.10	69.81 ± 0.08	99.06 ± 0.06
80%	81.13 ± 0.10	68.87 ± 0.08	98.11 ± 0.07

***Table 1**: Results of SSL Algorithms With Uncertainty Estimation.*

Conclusions and Future Work

The proposed framework addresses the challenge of ***limited labeled data*** in phase identification using ***SSL*** and ***uncertainty estimation***. By integrating ***Bayesian Neural Networks***, we achieved 98% ± 0.08 accuracy with robust ***uncertainty quantification***. It also provides critical insights into the ***minimum data*** needed for reliable phase identification, aiding future data collection and labeling efforts.

This work was authored in part by the National Renewable Energy Laboratory (NREL), for the U.S. Department of Energy (DOE) under Contract No. DE-AC36- 08GO28308. Funding for the NREL authors was provided by U.S. Department of Energy Office of Electricity under Agreement No. 102400X-FY23 TCF-23-20116. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes

IEEE Power and Energy Society General Meeting 2025
Austin, TX
7/27/2025-7/31/2025
NREL/PO-5D00-95656