MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Phase Identification in Wye-Connected and Delta-Connected Loads Under High PV Penetration

Ye, Zong-Jhen; Sun, Hongbo; Guo, Jianlin; Wang, Ye; Mohsenian-Rad, Hamed

TR2025-113 July 31, 2025

Abstract

This paper provides a data-driven phase identification approach for a power distribution system with high photovoltaic (PV) penetration in both wye-connected and delta- connected loads. The proposed approach is built upon concepts in information theory, followed by a mathematical proof to solve the challenge when the penetration of PV is up to 100%. The proposed method is based on a joint analysis of the event signatures between the smart meter at the customer side and those at the feeder head. The high accuracy of the proposed method is confirmed by conducting case studies on the IEEE 34 bus test system with different load profiles with different PV penetration levels. The maximum distinguishable number of customers for the proposed method is also estimated based on theory of probability.

IEEE PES GM 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Phase Identification in Wye-Connected and Delta-Connected Loads Under High PV Penetration

Zong-Jhen Ye[†], *Student Member, IEEE*, Hongbo Sun[‡], *Senior Member, IEEE*, Jianlin Guo[‡], *Senior Member, IEEE*, Ye Wang[‡], *Senior Member, IEEE* and Hamed Mohsenian-Rad[†], *Fellow, IEEE*
[†] Dept. of Electrical and Computer Engineering, University of California, Riverside, CA, 92521
[‡] Mitsubishi Electric Research Laboratories, Cambridge, MA, 02139
zongjhen.ye@email.ucr.edu, {hongbosun, guo, yewang}@merl.com, hamed@ece.ucr.edu

***Abstract*—This paper provides a data-driven phase identification approach for a power distribution system with high photovoltaic (PV) penetration in both wye-connected and delta-connected loads. The proposed approach is built upon concepts in information theory, followed by a mathematical proof to solve the challenge when the penetration of PV is up to 100%. The proposed method is based on a joint analysis of the event signatures between the smart meter at the customer side and those at the feeder head. The high accuracy of the proposed method is confirmed by conducting case studies on the IEEE 34 bus test system with different load profiles with different PV penetration levels. The maximum distinguishable number of customers for the proposed method is also estimated based on theory of probability.**

***Keywords*— Power distribution systems, phase identification, wye-connected load, delta-connected load, high PV penetration, mutual information, event signature, smart meters.**

I. Introduction

Phase imbalance in power distribution systems can increase power loss, decrease system efficiency, and raise maintenance cost. Load imbalance consistently exists in distribution systems due to factors such as aging and poor maintenance. The growing penetration of distributed renewable energy resources is also aggravating the system imbalance. However, before solving the issue of phase imbalance, it is necessary to first identify the phase for each load.

Broadly speaking, three types of methods have been discussed in the literature to address the phase identification problem [1]: hardware-based, power-based, and voltage-based approaches. Hardware-based methods require installation of probing devices on both the substation and customer sides. These methods ensure accuracy, but the installations can be costly and labor-intensive [2]. Power-based methods utilize the principle of energy conservation between power consumption and generations for each phase [3]. The accuracy of these methods depend on the number of loads, the network's topology, and the coverage of smart meters (SMs). Voltage-based methods assess the correlations between the voltage variations across the SMs to identify which customers are on the same phase [4]. In this paper, a new approach for phase identification is proposed by combining the power-based and voltage-based approaches.

Phase identification poses challenges for several reasons, including: limited SMs coverage, low reporting rates in measurements, insufficient measurement capabilities, variations in infrastructure due to different load transformers, lack of labeled phase data, and inaccuracies in topology information.

Regarding the first challenge above, a good method for phase identification should be independent of the coverage of SMs. For the reporting rates, an advanced sensor (such as phasor measurement unit, PMU) allows comparison of phase angles that can solve the problem easily [5]. However, setting up PMUs on the customer side is too expensive. Furthermore, traditional meters only measure real power consumption. They do not provide reactive power or the magnitude of voltage, which makes phase identification even more challenging [6, Section 2.8.3].

Another key challenge is related to load connections. A load connection could be single-phase wye-connected (i.e. connected between a phase wire and a neutral wire) [1], single-phase delta-connected (i.e. connected between two phase wires) [7], or three-phase [8] (balanced or unbalanced) using either wye or delta configuration. When an event or disturbance happens in a power system, such as a voltage sag, it affects loads differently depending on the types of connection. On the one hand, the event signature is only visible on a wye-connected load when the event occurs on the corresponding phase of the feeder. On the other hand, a delta-connected load can reflect the event signature on two phases simultaneously. The cross interference can hinder phase identification by the traditional methods, such as correlation or regression across the event signatures. Therefore, solving the phase identification problem is very difficult when the load connections are a mix of both wye- and delta-connected loads [9]. In addition, lacking label and topology information can cause difficulties for a machine learning method to solve the phase identification problem [9].

Several studies have concentrated on phase identification using either exclusively wye-connected loads [10] or exclusively delta-connected loads [7]. Few papers addressed phase identification in power systems that involved a combination of wye- and delta-connected loads. Also, few studies considered the impact of PV penetration on phase identification, such as [11], [12]. In [11], the authors proposed a random forest method that showed 86.5% and 94.4% accuracy on phase identification in two real-world feeders with 70% and 24% PV penetration relative to the peak load. However, it was shown that there is potential for overfitting. In [12], two PVs are considered in an IEEE 13 bus test system using a weighted least square optimization technique for phase identification. Although the studies in [11], [12] considered PV in their systems, we are more interested in the phase identification problem with a higher PV penetration percentage and reliable results.

In this research, data collected from the customer-side (i.e. SMs) and the feeder side are used to answer the following questions: **(1)** Can we identify the phases with a mix of loads in single-phase wye-connected, single-phase delta-connected, and unbalanced three-phase in the system? **(2)** Can we identify the phases on the customer side with different penetration of PVs and coverage of SMs? **(3)** Can we identify the phases despite the insufficient information, such as lack of labeled data, line impedance, and topology? The contributions of the proposed method are:

1) Phase identification with high accuracy even when the loads on the feeder are a combination of wye and delta connections.

Zong-Jhen Ye conducted this work during his internship at MERL.

2) Accurate phase identification regardless of the PV penetration and SM coverage.
3) Applied the probability theory to demonstrate the applications of the maximum distinguishable number of customers.
4) The proposed method is low calculation burden and ability to use a relatively small dataset to solve the phase identification problem despite various challenges. The proposed method does not rely on the topology, line impedance, labels, and training.

II. Methodology - A Data-Driven Event-Triggered Approach Based on Information Theory

A. Data Pre-process

Consider a power distribution network where SMs are installed at the load buses. Each SM at customer k can provide three measurements: real power (P_k), reactive power (Q_k), and magnitude of voltage ($|V_k|$). Then, one can estimate the current magnitude on the customer side as:

$$|I_k| = \frac{|S_k|}{|V_k|} \quad \forall k \in [1, c], \tag{1}$$

where $|S_k|$ is the magnitude of the apparent power, which can be found by the real power and the reactive power. The customer k is an integer number in the range of $[1, c]$, where c is the total number of customers in the distribution system.

The data are measured in time series by the SM on the customer side. A set $\mathbb{I}_c$ is used to combine the time series current in the given time n for all the customers as $\mathbb{I}_c = \{|I_1|, ..., |I_c|\} \in \mathbb{R}^{n\times c}$. Another advanced sensor on the feeder head can be used to measure the current in three phases. Note that the three phases are in set $\Phi = \{A, B, C\}$. The magnitude of current at the source side is noted as $|I_s^\Phi|$ for different phases in a set $\mathbb{I}_\Phi$. Given time n, we have $\mathbb{I}_\Phi \in \mathbb{R}^{n\times 3}$. Then, we we create the *differential* set using $\mathbb{I}_c(t) - \mathbb{I}_c(t-1)$ [13], where $\mathbb{I}_c^d \in \mathbb{R}^{(n-1)\times c}$. The same differential process is applied to $\mathbb{I}_\Phi$, which is noted as $\mathbb{I}_\Phi^d \in \mathbb{R}^{(n-1)\times 3}$. Note that the given time is from 1 AM to 7 AM in this research, similar to [14] to avoid noise from PVs.

Then, we use the function $f(.)$ to extract useful information from both the customer side and the feeder side as:

$$\chi = f(\mathbb{I}_k^d) \quad \text{for} \quad \mathbb{I}_k^d \notin [L_k^n, L_k^p], \tag{2}$$

where χ is a set for event $\mathbb{I}_k^d$ if it shows event signature beyond upper threshold L_k^p or lower threshold L_k^n. The threshold L_k^n is defined as $\overline{\mathbb{I}_k^d} - \sigma(\mathbb{I}_k^d)$ and the threshold L_k^p is defined as $\overline{\mathbb{I}_k^d} + \sigma(\mathbb{I}_k^d)$. Note that $\overline{\mathbb{I}_k^d}$ is the mean and $\sigma(\mathbb{I}_k^d)$ is the standard deviation (S.D.) of signal $\mathbb{I}_k^d$. In this research, we select only fluctuations that are beyond the range $[L_k^n, L_k^p]$ as "events" for analysis.

It should be noted that the selection of events is crucial because not every $\mathbb{I}_k^d$ is useful for phase identification. For example, suppose a customer uses an appliance from 6:00 to 6:02 AM. When the appliance is turned on at 6:00 AM, a current increment at the customer side would show between 5:59 and 6:00 AM. Conversely, a current decline at the customer side will show between 6:02 and 6:03 when the customer turns off the appliance. If both the customer and source meters are synchronized on the same phase, the sensor on the source side will also detect the on/off fluctuations.

If the customer does not turn on additional appliances, the PV installed on the customer side has no power supply before sunrise, and the other appliances operate at a constant power, then no fluctuations will be observed on both the customer side and the feeder side during the two-minute period. In such a case, no event signature can be captured even if we compare the same phase from the customer and the source sides. In other words, we can only identify the phase if the event signatures are clear from both the customer and the feeder.

B. Mutual Information among Joint Events

When both the customer and source sides simultaneously observe an event, it does not necessarily mean that they are on the same phase. First, there can be two separate events which occur concurrently on each of the customer and source sides. For example, if an event at customer k is *greater* than L_k^p while an event at phase Φ is *less* than L_k^n, it is highly possible that they are different events and customer k does not belong to phase Φ. Second, even if both events are greater than L_k^p or less than L_k^n, it *does not* guarantee that they are on the same phase. It is possible that different customers have similar load profiles and coincidentally show a similar pattern during the same event. Thus, mutual information is used to solve this issue.

Suppose the number of events m (hereafter m) is chosen between a pair of customer k and phase A in Φ, we can obtain the probability table as follows:

$$p_{A,k}(u, v), \tag{3}$$

where

$$u = \begin{cases} 1 & \text{if } \forall\ m : \mathbb{I}_\Phi^d(m) \geq L_k^p, \\ 2 & \text{if } \forall\ m : \mathbb{I}_\Phi^d(m) \leq L_k^n, \end{cases} \tag{4}$$

$$v = \begin{cases} 1 & \text{if } \forall\ m : \mathbb{I}_k^d(m) \geq L_k^p, \\ 2 & \text{if } \forall\ m : \mathbb{I}_k^d(m) \leq L_k^n. \end{cases} \tag{5}$$

Here, $p_{A,k}(1,1)$ shows that the probability of fluctuations in both source A and customer side are greater than L_k^p; $p_{A,k}(1,2)$ shows that the probability of fluctuations in source A is greater than threshold L_k^p while the fluctuations in the customer k is less than threshold L_k^n; $p_{A,k}(2,1)$ shows the probability of fluctuations in source A is less than threshold L_k^n while the fluctuations in customer k is greater than threshold L_k^p; and $p_{A,k}(2,2)$ is the probability of fluctuation less than L_k^n in both source A and customer side. By using (3), (4), and (5), we can now transform m into a probability table and describe the relationship between phase A at the source side and customer k.

It is possible that a probability table, such as (3), includes bias. It is also possible to mitigate bias in the probability table by increasing the number of events. In other words, the larger of m, the more accurate the results will be. In this paper, we only analyze the load phases when $m \geq 12$ over the period of seven days. Since we utilize the differential time series data, we have 359 = 60 minutes $\times$ 6 hours minus 1 data for a day. Over seven days, this yields a total of 2,513 = 359 $\times$ 7 time series of differential data. Thus, choosing $m \geq 12$ represents close to 0.5% of the total time series data in the given period, which is large enough for our numerical verification. Mathematical verification is provided in section IV.

Note that equation (3) can be expanded into:

$$p_{A,k}(u) \quad \text{and} \quad p_{A,k}(v), \tag{6}$$

where

$$\begin{aligned} p_{A,k}(u) &= \begin{cases} p_{A,k}(1,1) + p_{A,k}(1,2) & \text{if u = 1,} \\ p_{A,k}(2,1) + p_{A,k}(2,2) & \text{if u = 2,} \end{cases} \\ p_{A,k}(v) &= \begin{cases} p_{A,k}(1,1) + p_{A,k}(2,1) & \text{if v = 1,} \\ p_{A,k}(1,2) + p_{A,k}(2,2) & \text{if v = 2.} \end{cases} \end{aligned} \tag{7}$$

Notations $p_{A,k}(u)$ and $p_{A,k}(v)$ in equation (6) are the marginal probability of equation (3); and $p_{A,k}(u=1)$ is the probability that m at A are greater or equal to L_k^p and $p_{A,k}(u=2)$ is the probability that m at A are less or equal to L_k^n. Similarly, $p_{A,k}(v=1)$ and $p_{A,k}(v=2)$ is the probability that m at customer k are greater or equal to L_k^p and less or equal to L_k^n, respectively.

Equations (3) and (6) can be utilized for constructing a mutual information index [15, Ch 4] as:

$$MI(\Phi; k) = \sum_{p_{\Phi,k}(u)} \sum_{p_{\Phi,k}(v)} p_{\Phi,k}(u, v)\ log \frac{p_{\Phi,k}(u, v)}{p_{\Phi,k}(u) p_{\Phi,k}(v)}. \tag{8}$$

When the relationship of phase Φ and customer k is higher, the value of mutual information index is higher in (8), revealing the

answer for phase identification. Therefore, for a wye-connected load, we can use (9) to estimate the load phase $\hat{\Phi}$ as:

$$\hat{\Phi} = \arg\max_{\Phi,k} \ MI(\Phi; k). \tag{9}$$

Note that, $\{\hat{\Phi}\}$ is one phase because customer k is a wye-connected load. The process for identifying a delta-connected load is:

$$\hat{\Phi}' = \arg\min_{\Phi,k} \ MI(\Phi; k), \tag{10}$$

and

$$\hat{\Phi} = \Phi - \hat{\Phi}'. \tag{11}$$

The lowest value in (10) implies the relationship of phase $\hat{\Phi}$ and customer k is low. We can remove the phase $\hat{\Phi}'$ and derive the phase $\hat{\Phi}$ by (11). Note that $\{\hat{\Phi}\}$ from (11) should be two phases because customer k is a delta-connected load between two phases.

Before we end this section, we shall point out a special feature to the mutual information: two different cases could have the same mutual information index in equation (8), even if both cases have opposite probability tables. For example, suppose a probability table U is $p(1,1) = 0.1,\ p(1,2) = 0.4,\ p(2,1) = 0.4,\ p(2,2) = 0.1$, the result will be the same as another table V which is $p(1,1) = 0.4,\ p(1,2) = 0.1,\ p(2,1) = 0.1,\ p(2,2) = 0.4$. Although the results of (8) by different inputs U and V are the same, the Φ-k relationships are different because they are positively correlated in V but not in U.

The same values of (8) between different Φ and k is not an issue if $\{\hat{\Phi}\} = 1$ in (9). However, suppose $\{\hat{\Phi}\} > 1$ in (9), we need to decide which phase is the correct one for k. In this research, we choose the phase based on the highest summation value of $p(1,1)$ and $p(2,2)$ as:

$$\hat{\Phi} = \arg\max_{\hat{\Phi}} \left(p_{\hat{\Phi},k}(1,1) + p_{\hat{\Phi},k}(2,2) \right). \tag{12}$$

The reason is that $p(1,1)$ and $p(2,2)$ imply that phase Φ and k have the same pattern, which also imply that phase Φ and k can be on the same phase if k is a wye-connected load. Note that $\{\hat{\Phi}\} > 2$ in (11) is rare and is not considered in this research.

C. Process with Descending Order of Load Sequence

Due to a heavier load may show event signatures that can be identified easier than a lighter load, a strategy is designed for all the loads in the distribution system. This approach follows the descending order of load sequence in the system, which starts from identifying the phase of heaviest load to the lightest load in the distribution system. Once the phase for each load is identified, we update the current on the source side by subtracting the customer-side current from the corresponding source phase.

Note that the process to subtract the current will be different between a wye-connected load and a delta-connected load. For a wye source and a wye-connected load, the updated current on the source side is:

$$|I_s^{\hat{\Phi}}| = |I_s^{\hat{\Phi}}| - |I_k|. \tag{13}$$

We cannot use (13) for a delta-connected load because the current flow is between two phases. Also, a $\frac{\pi}{6}$ angle difference in radius occurs naturally between a wye source and a delta-connected load. We assign $\hat{\Phi}(1)$ and $\hat{\Phi}(2)$ to denote the input and output phases of the delta-connected load. The updated currents from the source sides are:

$$|I_s^{\hat{\Phi}(1)}| = |I_s^{\hat{\Phi}(1)} - I_k| \quad \text{and} \quad |I_s^{\hat{\Phi}(2)}| = |I_s^{\hat{\Phi}(2)} + I_k|, \tag{14}$$

where

$$I_k = \left(\frac{S_k}{V_k} \right)^* = \frac{P_k - jQ_k}{|V_k| e^{-j\left(\frac{\pi}{6} + \angle V_s^{\hat{\Phi}(1)}\right)}} \tag{15}$$

Note that the SM at the customer side can only measure the real power (P_k), the reactive power (Q_k), and the magnitude of voltage ($|V_k|$). We assume the voltage angle is the same between the source and customer k because the phase angle difference in the distribution system is small. Since we have identified the phase at k, the voltage angle at k can be estimated by the voltage at corresponding phase add $\frac{\pi}{6}$ due to the connection between wye source and delta load. Therefore, we can calculate the current at customer k in a complex number by (15) and update the current on source side by (14).

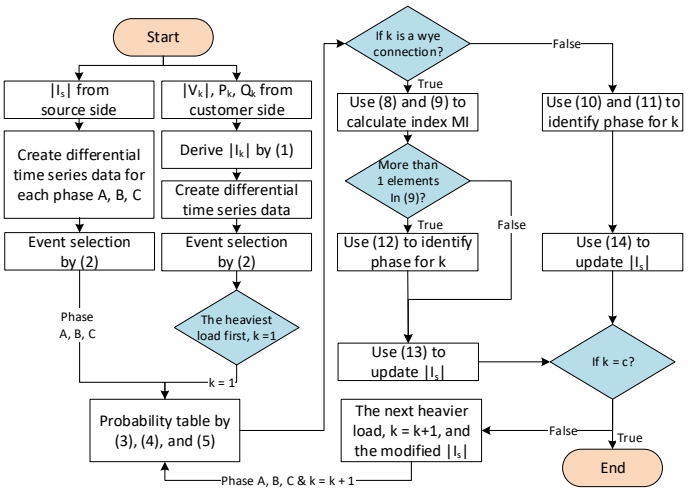

Fig. 1. The flow chart of the proposed method for phase identification in the distribution with c loads.

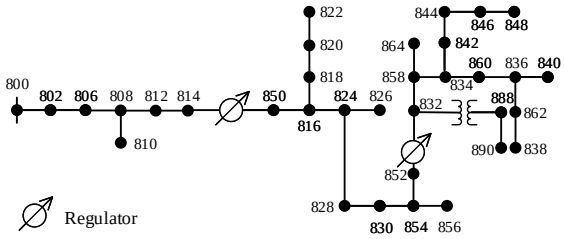

Fig. 2. The IEEE 34 bus test system [18].

Fig. 1 shows the process of the proposed method. Note that although the utility companies may have incorrect phase information of the load, it should have information of power consumption. For example, in the State of California, U.S., the default MWh should be included [16]. Accordingly, the proposed method is available to identify the phase based on the default load sequence in the distribution system.

The cross entropy in the proposed method can perform better than correlation because correlation only consider the linear relationship. However, the relationship between the customer side k and its corresponding substation phase could be positively related but not linearly related. Thus, it is better to use cross entropy to solve the problem; see [17] for the related theoretical discussion.

TABLE I
DEFAULT LOADS IN IEEE 34 BUS SYSTEM.

Phase	Wye Load		Phase	Delta Load	
	kW	kVAr		kW	kVAr
A	351	224	AB	255	133
B	330	210	BC	254	134
C	213	155	CA	366	188
Total	894	589	Total	875	455

III. RESULTS AND PERFORMANCE EVALUATION

The IEEE 34 bus test system (see Fig. 2 for the topology and Table I for the load distribution) is used in this study because default

TABLE II
ACCURACY OF THE PROPOSED METHOD ON DIFFERENT PV PENETRATION RATES IN DIFFERENT PERIODS

PV Penetration Rate (%)	Data Period	
	1 AM to 7 AM	11 AM to 16 PM
0	100.0	100.0
33.3	100.0	86.7
66.7	100.0	60.0
100.0	100.0	70.0

loads are half wye-connected and half delta-connected [18]. The per minute load profiles are from the IEEE European Low Voltage Test Feeder (EuLVTF) [18]. The load profiles are normalized to the [0,1] scale. Then, multiply the default loads in the IEEE 34 system by the normalized load. Since the IEEE 34 test system only has thirty distributed loads while the EuLVTF has a hundred load profiles, the load profiles are shuffled and assigned randomly to the load in the IEEE 34 test system on the same day with the power factor set at 0.95. For this study, daily simulation is used, but there is no relationship between days since the load profiles are assigned randomly.

PVs are added to the IEEE 34 test system with different penetration rates. Note that each PV rated power is 3.5 kW. The irradiation and temperature for PVs are using July 1st, 2021, from NSRDB (National Solar Radiation Database) [19]. Since the database of NSRDB is per-five-minute resolution, we used linear interpolation to treat the data in per-minute resolution. Also, since the IEEE 34 test system is a distribution system, we use the same weather conditions for all PVs. The connection of PV to each load is single-phase and parallel connected. All the PVs are behind-the-mete (i.e. we cannot see the power and current generated by the PVs). All the simulations are conducted using MATLAB and Simulink [20].

A. Different PV Penetration Rates in Different Time Periods

The PV penetration rates of 0%, 33.3%, 66.7%, and 100%, respectively correspond to PVs in 0, 10, 20, and 30 out of all thirty loads in the IEEE 34 bus test system. Table II shows results of the proposed method. The accuracy is 100% for all the PV penetration rates when the proposed method works with the differential time series data from 1 AM to 7 AM. The reason is that PVs have less impacts on the event signature due to less irradiance while the proposed method can still recognize the event signatures between source and customer sides. The results of a different period, 11 AM to 16 PM, are also listed in the same table to show the impact of PVs. Although the proposed method can still work with 100% accuracy when no PVs are installed, the accuracy of the proposed method can decrease to 60% with PVs. The main reason is that the proposed method depends on the event signature, and the contribution of behind-the-meter PVs can bring challenges to phase identification. One may notice that the accuracy decreases to 60% when the PV penetration rate is 66.7% but increases to 70% when the PV penetration rate is 100%. A possible reason is that the simulation in this study assumes all the PVs in the IEEE 34 bus test system have the same irradiation and temperature inputs. This assumption may help the proposed method separate load profiles easily when the PV penetration rate is 100%. For the rest of the analysis, we use the differential time series data from 1 AM to 7 AM.

B. Different Percentages of Wye- and Delta-connected Loads

Although the default loads in IEEE 34 bus system are half wye- and half delta-connected, it is important to verify the proposed method on different percentage of wye- and delta-connected loads. Suppose we keep the same topology but switch all the delta connected loads into wye-connected loads (i.e. switch delta load in phase AB to phase A, phase BC to phase B, and phase CA to phase C), one can then test the proposed method with 100% wye-connected loads. However, we cannot switch all the wye loads into delta connected loads because some wye loads are at single phase. We can only switch the wye-connected load into delta-connected if the load is at the topology with more than one phase. Based on the new settings, the modified IEEE 34 bus system includes 30% wye and 70% delta connected loads. Using different percentage of wye- and delta-connected loads, the proposed method can still reach 100% accuracy.

C. Different SM Coverage

It is possible that not all the customers have SMs installed, which means equations (13) or (14) are unavailable for all customers. Therefore, here we test the performance of the proposed method with different SM coverage. Assuming each customer has only two options, to install an SM or not, we can define SM coverage (%) as:

$$\text{SM coverage } (\%) = \text{Number of SM installed}/c \times 100\%.$$

Since the proposed phase identification method can only be available for customers who installed SM, the accuracy for different SM coverage is calculated as:

$$\text{Average Accuracy } (\%) = \frac{1}{n_M}\sum_{j=1}^{n_M}\left(\frac{1}{c}\sum_{k=1}^{n_{SM}} i \times 100\%\right). \quad (16)$$

The variable n_M is the Monte Carlo testing number, n_{SM} is the total number of SM deployed in the distribution system, i is 1 if $\hat{\Phi}_k = \Phi_k^T$, and i is 0 if $\hat{\Phi}_k \neq \Phi_k^T$, where Φ_k^T is the true phase of the customer k. A Monte Carlo test with $n_M = 100$ is used for testing and the results are summarized in Figure 3 (a), which shows that the proposed method can still maintain a high accuracy level for the customer who installed the SM in the distribution system.

D. Measurement Errors of SMs

Measurement errors are possible for all kind of sensors. Since the SMs can measure the real power, the reactive power, and the magnitude of voltage, we assume the measurement errors caused by the summation of each measurement belong to a normal distribution with zero mean and different standard deviations (S.D.) as 0.01, 0.05, and 0.1. With a hundred Monte Carlo tests (i.e. $n_M = 100$), the average accuracy of the proposed method is summarized as Figure 3 (b). The average accuracy with $n_{SM} = c$ will drop to 97.2%, 90.4%, and 68.9% respectively, corresponding to the settings of S.D.

E. Comparison with Cross Correlation and kNN

Two traditional methods, correlation and k nearest neighbor (kNN), are considered in this section [4], [21], [22]. For correlation, we use the same differential time series data as Section II-A and remove the corresponding current after the phase is recognized. For kNN method, we build the data set using the same differential time series data as Section II-A. By considering the combination of different source Φ and customer k, we labeled the correct combinations as 1 and incorrect ones as 0. The dimension of the data set for kNN is $\in \mathbb{R}^{5027\times 90}$. MinMaxScaler is used for kNN data pre-processing, and the results remain in the same when we switch from MinMaxScaler to StandardScaler. The data set is divided into 80% in a training set and 20% in a testing set. A preliminary search of $k = 3$ is used for kNN. The data set is shuffled every time before running kNN.

The results for all three methods are shown in Figure 3 (c) based on equation (16) with one hundred cases. The proposed method provides a significantly higher accuracy compared other methods. Note that in the kNN process, the current is not removed from the source side once we obtain the phase of customer.

IV. DISTINGUISHABLE NUMBER OF CUSTOMERS

Section III shows that the proposed method can successfully identify the load phase. However, suppose two customers have *exactly* the same load profiles in a given period, the proposed method will encounter challenges in distinguishing their phases. The more customers in the distribution system, the higher possibility that two customers

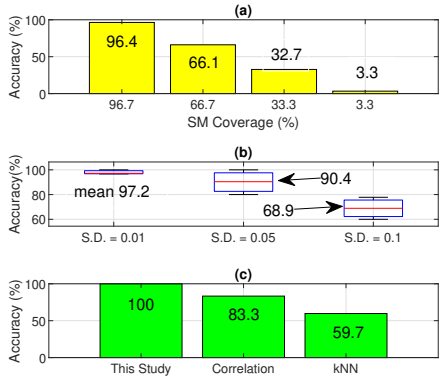

Fig. 3. Figure (a) results of different SM coverage, Figure (b) measurement errors of SMs with 100 tests when the standard deviation (S.D.) is 0.01, 0.05, and 0.1, and Figure (c) different method comparison.

have the same load profiles. We can solve this problem by giving a confidence probability to estimate the maximum distinguishable number of customers, denoted as C.

Suppose $P(.)$ is denoted as the probability of an occurrence and suppose a load changing is a binary variable with 50% for each on/off switching, the probability of any on/off sequence at customer k with m events is $P(\text{customer } k) = (0.5)^m$. Next, the probability that no customer in C has the same on/off sequence as customer k can be denoted as $(1-(0.5)^m)^C$. Suppose the confidence probability is 99%, which means that any other customers in C has the same load profile as customer k is 1%, we can have $1-(1-(0.5)^m)^C = 1\%$ with C as:

$$0.99 = (1-(0.5)^m)^C \Rightarrow C = \frac{log(0.99)}{log\,(1-(0.5)^m)}. \qquad (17)$$

We can solve (17) with $m = 12$ and result in the C in a truncation integer, $C = 41$. Since the feasible number of customer is greater than the customers we tested in IEEE 34, it explains the reason that the proposed method can fully identify the phase of customers in the case study. For a distribution system with more customers than C, three strategies can be applied to the proposed method: (1) increase m for mutual information, (2) split the distribution system into smaller systems, or (3) decrease the probability of the same load profile, such as from 1% to 0.1%.

Note that the proposed method is unable to distinguish customer phases when the load profile and power consumption are identical across all phases. We can use the proposed method to effectively recognize the customer phases for single-phase wye-connected loads, single-phase delta-connected loads, and unbalanced three-phase loads in both wye and delta configurations under a reasonable size of system. However, balanced three-phase loads are not considered in this method because the event signatures remain identical across all three phases.

V. Conclusion and Future Work

Identifying load phases in a distribution system with wye- and delta-connected loads poses significant challenges. The problem escalates with higher PV penetration. This paper used advanced SMs capable of measuring real power, reactive power, and voltage magnitude per minute to address the problem. Based on the technology, we propose an approach for identifying the phase of a wye- or delta-connected load with different PV penetration rates. As a data-driven method, the accuracy of the proposed method depend on the accuracy of measurements but does not require labels, topology nor line impedance in the distribution system. An application approach for feasible number of customers is also discussed in this paper, showing strategies for applying the proposed method to distribution systems with more customers.

Future work can address the phase identification for a balanced three-phase load and PVs, excess PV power generated, and extend the proposed method on a larger system.

References

[1] F. Therrien, L. Blakely, and M. J. Reno, "Assessment of measurement-based phase identification methods," *IEEE Open Access Journal of Power and Energy*, vol. 8, pp. 128–137, 2021.

[2] P. Kulkarni and A. R. Kolwalkar, "Phase identification system and method," Patent 2 562 554.

[3] A. Heidari-Akhijahani, A. Safdarian, and F. Aminifar, "Phase identification of single-phase customers and PV panels via smart meter data," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 4543–4552, 2021.

[4] T. A. Short, "Advanced metering for phase identification, transformer identification, and secondary modeling," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 651–658, 2013.

[5] M. Bariya, D. Deka, and A. von Meier, "Guaranteed phase topology identification in three phase distribution grids," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 3605–3612, 2021.

[6] H. Mohsenian-Rad, *Smart Grid Sensors: Principles and Applications*. Cambridge University Press, UK, Apr. 2022.

[7] K. Matsumoto, Y. Fukuyama, K. Seki, A. Oi, T. Jintsugawa, and H. Fujimoto, "Connection phase estimation of pole mounted distribution transformers by integer form of population based incremental learning considering measurement errors and outliers by correntropy," in *IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Brisbane, Australia, Dec. 2021.

[8] M. Izadi and H. Mohsenian–Rad, "Improving real-world measurement-based phase identification in power distribution feeders with a novel reliability criteria assessment," in *IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, Espoo, Finland, Dec. 2021.

[9] B. Foggo and N. Yu, "Improving supervised phase identification through the theory of information losses," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2337–2346, 2020.

[10] A. Heidari-Akhijahani, A. Safdarian, and F. Aminifar, "Phase identification of single-phase customers and pv panels via smart meter data," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 4543–4552, 2021.

[11] H. Padullaparti, S. Veda, J. Wang, M. Symko-Davies, and T. Bialek, "Phase identification in real distribution networks with high PV penetration using advanced metering infrastructure data," in *IEEE Power Energy Society General Meeting (PESGM)*, Denver, CO, USA, Jul. 2022.

[12] V. D. Krsman and A. T. Sari´c, "Verification and estimation of phase connectivity and power injections in distribution network," *Elect. Power Syst. Res.*, vol. 143, pp. 281–291, 2017.

[13] L. Blakely, M. J. Reno, C. B. Jones, A. Furlani-Bastos, and D. Nordy, "Leveraging additional sensors for phase identification in systems with voltage regulators," in *IEEE Power and Energy Conference at Illinois (PECI)*, Urbana, IL, USA, Apr. 2021.

[14] S. Overington, D. Edwards, P. Trinkl, and A. Buckley, "Application of constrained k-means algorithm for phase identification," in *31st Australasian Universities Power Engineering Conference (AUPEC)*, Perth, Australia, Nov. 2021.

[15] J. V. Stone, *Information Theory: A Tutorial Introduction*. Sebtel Press, 2015.

[16] California Independent System Operator (CASIO), "Reliability coordinator services agreement," https://www.caiso.com/rules/Pages/ContractsAgreements/Default.aspx.

[17] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon, "Mutual information analysis: A comprehensive study," *J. Cryptol.*, vol. 24, no. 2, pp. 269–291, Apr. 2011.

[18] "IEEE PES test feeder," https://cmte.ieee.org/pes-testfeeders/resources/.

[19] "National Renewable Energy Laboratory," https://nsrdb.nrel.gov/.

[20] MATLAB, *version: 9.14.0.2286388 (R2023a)*. Natick, Massachusetts, United States: The MathWorks Inc., 2023.

[21] M. Xu, R. Li, and F. Li, "Phase identification with incomplete data," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2777–2785, 2018.

[22] K. Chen, J. Shi, X. Wei, and S. Cai, "Phase identification with single-phase meter and concentrator based on NMF dimension reduction and label propagation," in *2021 IEEE 11th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems*, 2021, pp. 1–6.