Phase Identification in Real Distribution Networks with High PV Penetration Using Advanced Metering Infrastructure Data

Preprint

Harsha Padullaparti,[1] Santosh Veda,[1] Jiyu Wang,[1]
Martha Symko-Davies,[1] and Tom Bialek[2]

1 National Renewable Energy Laboratory
2 San Diego Gas & Electric Company

Presented at the 2022 IEEE Power and Energy Society General Meeting
Denver, Colorado
July 17–21, 2022

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5D00-81336
July 2022

Phase Identification in Real Distribution Networks with High PV Penetration Using Advanced Metering Infrastructure Data

Preprint

Harsha Padullaparti,[1] Santosh Veda,[1] Jiyu Wang,[1] Martha Symko-Davies,[1] and Tom Bialek[2]

1 National Renewable Energy Laboratory
2 San Diego Gas & Electric Company

Suggested Citation
Padullaparti, Harsha, Santosh Veda, Jiyu Wang, Martha Symko-Davies, and Tom Bialek. 2022. *Phase Identification in Real Distribution Networks with High PV Penetration Using Advanced Metering Infrastructure Data: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-5D00-81336. https://www.nrel.gov/docs/fy22osti/81336.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5D00-81336
July 2022

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Electricity and by California's Electric Program Investment Charge (EPIC) Program in San Diego Gas & Electric Company under a cooperative research-and-development agreement (CRADA) # CRD-17-712. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

Phase Identification in Real Distribution Networks with High PV Penetration Using Advanced Metering Infrastructure Data

Harsha Padullaparti[1], Santosh Veda[1], Jiyu Wang[1], Martha Symko-Davies[1], Tom Bialek[2]
[1]National Renewable Energy Laboratory (NREL), Golden, Colorado.
[2]San Diego Gas & Electric Company (SDG&E), San Diego, California.
HarshaVardhana.Padullaparti@nrel.gov

***Abstract*—Many distribution network monitoring and control applications—including state estimation, Volt/VAr optimization, and network reconfiguration—rely on accurate network models; however, the network models maintained by utilities can become outdated because of restoration activities, network reconfiguration, and missing data. With the widespread deployment of advanced metering infrastructure (AMI), abundant measurement data from low-voltage secondary networks are available. The AMI measurement data can be used for phase identification to improve the network models. Although the existing phase identification techniques work well in passive distribution feeders that do not have photovoltaic (PV) generation, they can fail to accurately identify the phases in the presence of PV. This paper proposes a robust phase identification algorithm based on supervised machine learning that accurately identifies the AMI meter phase connectivity in the presence of significant PV generation. The proposed algorithm does not require network topology information or feeder-head measurement data. The algorithm is validated using the AMI measurement data collected in the field and the field-validated phase connectivity database on two real distribution feeders from San Diego Gas & Electric Company that have significant PV generation.**

Keywords—Advanced metering infrastructure (AMI), distribution network, machine learning, phase identification, photovoltaic systems, smart meter, supervised learning.

I. Introduction

The penetration of distributed energy resources (DERs) has continued to increase in recent years. The integration of DERs helps to meet increasing load demand and reduce the fossil fuels needed for generation. However, the high penetration of DERs also requires electric utilities to have better monitoring applications of their distribution systems because of the uncertainty from some types of DERs, such as photovoltaics (PV) [1]. In the United States, electric utilities have started to deploy advanced metering infrastructure (AMI) as a part of ongoing advancements in their systems [2]. The introduction of AMI smart meters provides opportunities for data analytics because the meters record a large amount of data on the customer side [3]. The AMI data analytics examples include primary voltage estimation [4], phase identification [5], [6], transformer identification [6], topology estimation [7], [8] and load disaggregation [9], [10].

Among these analyses, phase identification is a critical aspect because it gives an accurate distribution network and phase connectivity model. Typically, electric utilities do not have accurate phase connectivity information of their distribution system because this information can keep changing when new customers are connected to the feeder. The phase identification problem is defined as identifying the phase connectivity of each smart meter and structure in the power system network. With phase identification results, electric utilities can update their phase connectivity information in their models so they can make more accurate simulation and control decisions [11].

Several works in the literature have attempted to address the phase identification problem. The techniques reported in [12]-[14] use special equipment to identify the phase connectivity in distribution feeders. Although they provide accurate results, these techniques are time consuming and expensive because they involve manual processes and additional equipment. The data-driven methods that use linear regression and voltage correlation [5], [7] as well as supply and load balancing approaches [15], [16] are applicable to distribution feeders that have only phase-to-neutral connections. These techniques cannot be applied to feeders that have a mix of phase-to-neutral and phase-to-phase connections. In [17]-[20], the phase identification is performed in the feeders having phase-to-neutral connections only. Their application to the feeders having a mix of phase-to-neutral and phase-to-phase connections is not shown. Further, the works [5], [7] assume the availability of reliable substation supervisory control and data acquisition (SCADA) data. The clustering-based phase identification algorithms reported in [21]–[23] require network topology information to define must-link constraints in their algorithms. However, the topology information derived from a geographic information system or a utility planning network model can be inaccurate, which is a primary reason to develop a data-driven phase identification method. The works [17], [24]-[26] used numerical simulations for the validation of phase identification algorithms. While the simulations are a good starting point for the algorithm development, the performance must be validated on the field AMI data for practical application. Further, all these works are validated on passive distribution networks only that do not include PV generation. As shown in [6], the PV generation can adversely impact the accuracy of the phase identification. Thus, the research gaps in the existing phase identification works are: (a) application to passive distribution networks only without PV (b) validation using numerical simulations only without testing on real data (c) application to feeders having phase-to-neutral connections only, and (d) dependency on topology and/or SCADA data.

This paper proposes a robust phase identification algorithm to address the research gaps. The phase identification is treated as a data classification problem, and supervised learning is applied. Specifically, a random forest classifier is used to perform the phase identification. The advantages of a random forest classifier include efficient handling of large data sets, large number of variables, and faster computational time [27]. Since these features are desirable for phase identification application, we used the random forest classifier in this work. The proposed algorithm offers the following advantages over the existing works:

1. The algorithm is robust and provides high phase identification accuracy in the presence of significant PV generation, a mix of phase-to-neutral and phase-to-phase meter connectivity, and overhead/underground lines. To our knowledge, our work is the first to validate such robust performance on real distribution feeders.
2. The algorithm does not require network topology information, SCADA data, or AMI data from all the meters connected to a given secondary. It can accurately determine the phase connectivity even when the AMI data from some customers are missing.
3. The performance of the algorithm is demonstrated using real field AMI data collected from our utility partner.

In the remainder of this paper, Section II describes the distribution feeder characteristics and the AMI data sets used in this study. Section III presents the proposed phase identification algorithm. Section IV discusses the results, and Section V presents the conclusions and future work.

II. Distribution Feeders of SDG&E and Field Data

A. *Feeder Details*

The phase identification is performed on two distribution feeders of San Diego Gas & Electric Company (SDG&E) using the proposed algorithm. The topology of the first feeder, referred to as Feeder 1, is shown in Fig. 1. This is a 12-kV feeder with a peak load of 10.3 MW. The substation transformer is equipped with a load tap changer (LTC). Three capacitor banks are available on the feeder for reactive power support. The feeder serves more than 5,000 customers using 341 service transformers. Most of this feeder comprises overhead lines. Further, this feeder has an existing PV penetration of nearly 70% relative to the peak load in the field. The PV locations are highlighted in Fig. 1. It is observed that the PV systems are present across the feeder and are not confined to any specific feeder neighborhoods.

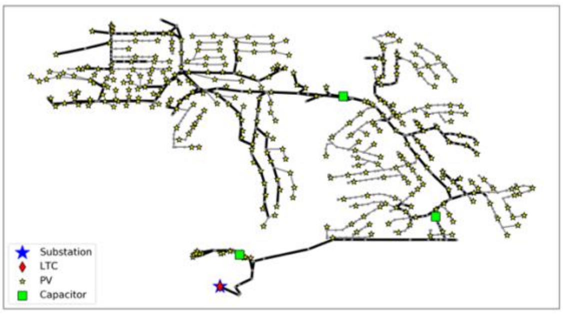

Fig. 1. Topology of the distribution Feeder 1.

The topology of the second feeder (Feeder 2) is shown in Fig. 2. This is a 12-kV feeder with a peak load of 13.3 MW. An LTC and two capacitor banks provide voltage regulation and reactive power support in this feeder. An existing PV generation of 3.14 MW is present in the field, which is nearly 24% relative to the peak load. The customers on this feeder are served using 657 service transformers. Feeder 2 is an underground distribution feeder, i.e., all the customers are served using underground distribution cables.

B. *Field Data*

The AMI data sets collected in the field include the minimum, maximum, and average voltage magnitude time-series data at 5-minute resolution and the net real power consumption data for each AMI meter for the entire year of 2019. Of these data, only the 5-minute average voltage magnitude data are used by the phase identification algorithm. The AMI data for two meters per service transformer are available in the data sets for both feeders. Further, the field-validated phasing information is collected from all the meters through manual field verification. These data are used as the ground truth for validating the phase identification results. The ground truth data represents the actual AMI meter phase connectivity in the field confirmed through the field verification. The phase connectivity distributions of both feeders from the ground truth data are shown in Fig. 3. It is observed that although Feeder 1 has a considerable number of AMI meters associated with all six types of phase connectivity, Feeder 2 has AMI meters primarily associated with phase-to-neutral phase connectivity only. Only 28 of 857 AMI meters in Feeder 2 have phase-to-phase connectivity.

Fig. 2. Topology of the distribution Feeder 2.

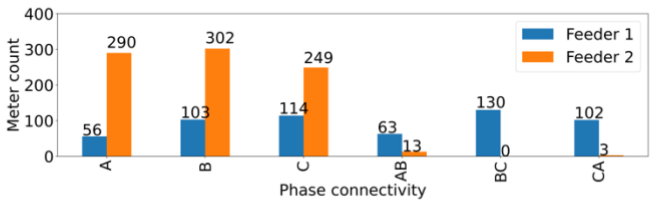

Fig. 3. Phase connectivity distributions of meters from ground truth data.

III. Proposed Phase Identification Algorithm

This section presents a brief overview of the random forest classifier and the proposed phase identification algorithm steps.

A. *Random Forest Classifier*

In recent years, the ensemble method has been widely used for classification and regression among different machine learning methods [28], [29]. The main idea of the ensemble method is to aggregate multiple weighted models to obtain a combined one with improved performance. A random forest is an ensemble learning method for classification consisting of multiple decision trees [30]. It provides an estimation that can capture different levels of relationships among variables.

A random forest generates the training data set by randomly drawing with replacement N examples, where N is the size of the original training data set [31]. These training data sets will be used for the feature and feature combination selection. A random forest usually uses the Gini index as a measure for the best split selection, which measures the impurity of a given element with respect to the rest of the classes. For a given training data set, T, the Gini index can be expressed as [32]:

$$\sum\sum_{j\neq i}\left(\frac{f(C_i,T)}{|T|}\right)\left(\frac{f(C_j,T)}{|T|}\right)$$

where $\frac{f(C_i,T)}{|T|}$ is the probability that the selected case belongs to class C_i. By using a given combination of features, a decision tree is made to increase to its maximum depth. Users will define the maximum depth and the number of trees to be grown. The number of trees will be equal to the number of features because only the selected features will be searched for the best split at each node.

B. Phase Identification Using Random Forest Classifier

The steps in the proposed phase identification algorithm are shown in Algorithm 1. In the first step, the AMI voltage time-series data collected from the meter data management system for a selected feeder and duration are read. Specifically, the average voltage magnitude data at 5-minute resolution are loaded. The data preprocessing is performed in the second step. This includes the data cleanup and data standardization. The meters having missing data for more than half the selected duration or having voltages beyond a reasonable range (50%–150% of nominal voltage) are removed from the data set. Further, the voltage time-series data from all the meters are aligned based on the time stamps. If a data point for a given time stamp is missing from a meter, the corresponding data points of all the meters are removed from the data set for the analysis. After the data cleanup, the AMI data set is standardized by performing mean centering and normalizing by the standard deviation.

In the third step, 30% of the meters selected for the training along with their known phase labels are read. These phase labels can be obtained preferably through field verification. They can also be obtained from an existing connectivity database that is known to be accurate or by applying existing clustering techniques. The higher the training data the higher is the phase identification accuracy. However, this also requires higher time and effort in verifying the phase labels of more AMI meters used in the training dataset. Therefore, a trade-off needs to be made between the efforts required in creating the training dataset and the expected level of phase identification accuracy. We observed the phase identification accuracy levels above 80% with 20% meters and around 90% accuracy with 30% meters in the training dataset. While increasing the meters in the training dataset further can improve the phase identification accuracy, it may not be worth the effort. Thus, we used 30% meters in the training dataset.

In the fourth step, a random forest classifier is constructed using the training data set. This classifier is used to identify the phase connectivity of the meters in the testing data set. In the fifth step, the voltage time-series data in the testing data set are supplied to the random forest classifier to identify the phase connectivity of the meters in this data set. The output of this step is a list of AMI meters and the corresponding phase connectivity labels identified by the algorithm. These results are saved in the last step for post-processing, such as visualization and model corrections.

Algorithm 1 Phase identification using random forest classifier

1: Load AMI data set.

2: Perform data preprocessing: data cleaning and standardization.

3: Load training data, including field-validated phase labels for the AMI meters in the training data.

4: Construct a random forest classifier using the training data.

5: Input the voltage time-series testing data to the random forest classifier and obtain the corresponding phase connectivity.

6: Save the phase identification results for post-processing.

C. Assumptions and Limitations

The proposed phase identification algorithm assumes that the number of phase connections in the feeder is known. This is a reasonable assumption since the utility engineers typically have this information in their database. It is assumed that the training data including the accurate phase labels for the meters in the training dataset are available. The proposed phase identification algorithm uses voltage time series data from the AMI meters. As the conventional meters used for billing only record the power consumption data, this algorithm cannot be applied to the feeders that do not have the AMI meters installed. The training data parameters such as data duration, granularity, number of meters etc. influence the phase identification accuracy.

IV. Results and Discussion

The proposed phase identification algorithm is implemented in Python using scikit-learn package on a Windows machine with i7 processor and 32 GB memory. The algorithm is applied on the AMI data sets collected in the field for two SDG&E feeders. The identified phase connectivity labels from the algorithm are validated against the ground truth phase connectivity data for each AMI meter obtained through field verification. The computation time for both training and testing is less than a minute for the datasets used in this work. The results are discussed in this section.

A. Feeder 1

The phase identification results of Feeder 1 are shown in Fig. 4. For each type of phase connectivity, the number of meters that the algorithm has identified as pertaining to that connectivity is shown against the ground truth in this figure. The results show that the phase identification algorithm can accurately identify all the types of phase connectivity.

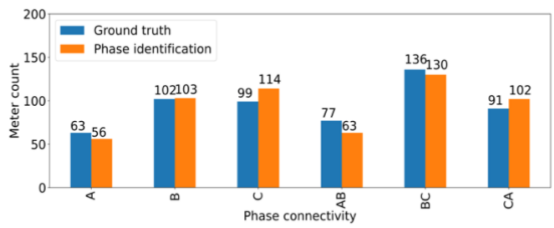

Fig. 4. Phase identification results of Feeder 1.

The detailed breakdown of the AMI meter counts in each of the training, testing, and full data sets of the phase identification is shown in TABLE I. The phase label column indicates whether the AMI meter connectivity labels for the meter counts are obtained from the ground truth or from the phase identification (Correct phase ID) results. For each type of phase connectivity, 30% of the meters are randomly selected along with their ground truth phase connectivity for the training data set. The full data set includes both the training and testing data sets together. With the phase connectivity accurately identified for 347 of 401 meters in the testing set alone, the phase identification accuracy is 86.5% on the testing set. The phase identification accuracy on the training and full data sets are 100% and 90%, respectively. Note that the meter counts in the 'Correct phase ID' row in TABLE I represent the number of meters for which the phase connectivity is identified correctly. These meter counts do not match with the phase identification meter counts in Fig. 4 as the meter counts in Fig. 4 also include meters for which the phase connectivity is identified incorrectly.

TABLE I. PHASE IDENTIFICATION RESULTS OF FEEDER 1.

Data set	Phase label	Phase connectivity						Total
		A	B	C	AB	BC	CA	
Full	Ground truth	63	102	99	77	136	91	568
	Correct phase ID	55	98	98	56	126	81	514
Testing	Ground truth	45	72	70	54	96	64	401
	Correct phase ID	37	68	69	33	86	54	347
Training	Ground truth	18	30	29	23	40	27	167
	Correct phase ID	18	30	29	23	40	27	167

The geographic locations of the AMI meters and the corresponding match/mismatch with the ground truth in the phase identification results are shown in Fig. 5. The meters are distributed all over the feeder; thus, the algorithm can detect the correct phase connectivity in all the neighborhoods of Feeder 1. In Fig. 5, the meter locations for which the phase connectivity is identified incorrectly are highlighted by the purple squares. The locations of the meters for which the phase connectivity is identified correctly are highlighted by the corresponding phase markings in the legend. It is observed that the phase identification errors are not clustered or confined to any specific feeder neighborhoods.

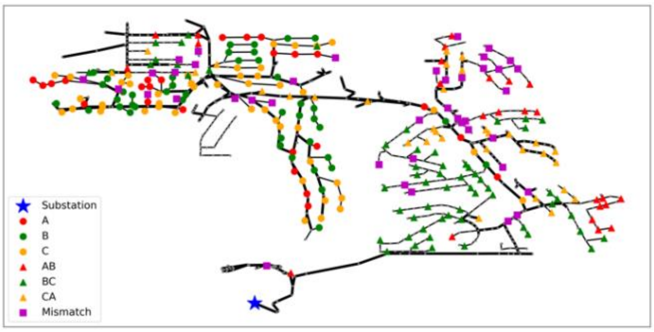

Fig. 5. Geographic distribution of phase identification match/mismatches in Feeder 1.

B. Feeder 2

The phase identification results of Feeder 2 are shown in Fig. 6 and TABLE II. As noted in Section II.B, Feeder 2 has predominantly phase-to-phase meter connectivity. The phase identification accuracy is low for phase connectivity BC and CA. This is because of the lack of sufficient training data for the algorithm with the low number of AMI meters with this connectivity. Further, as observed in TABLE II, the phase connectivity is correctly identified for 809 of 857 meters in the full data set, which includes both training and test data sets, which represents 94.4% accuracy. The phase identification accuracy on the training and full data sets are 100% and 92%, respectively.

The geographic locations of the AMI meters for which the algorithm correctly and incorrectly identified the phase connectivity are shown in Fig. 7. It is evident that the phase connectivity is correctly identified in the entire feeder and that the phase identification errors are not confined to any specific feeder neighborhoods.

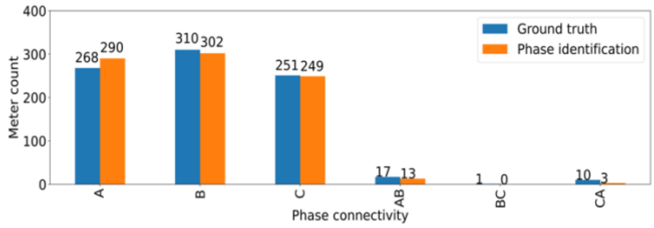

Fig. 6. Phase identification results of Feeder 2.

TABLE II. PHASE IDENTIFICATION RESULTS OF FEEDER 2.

Data set	Phase label	Phase connectivity						Total
		A	B	C	AB	BC	CA	
Full	Ground truth	268	310	251	17	1	10	857
	Correct phase ID	260	293	241	12	0	3	809
Testing	Ground truth	188	217	176	12	1	7	601
	Correct phase ID	180	200	166	7	0	0	553
Training	Ground truth	80	93	75	5	0	3	256
	Correct phase ID	80	93	75	5	0	3	256

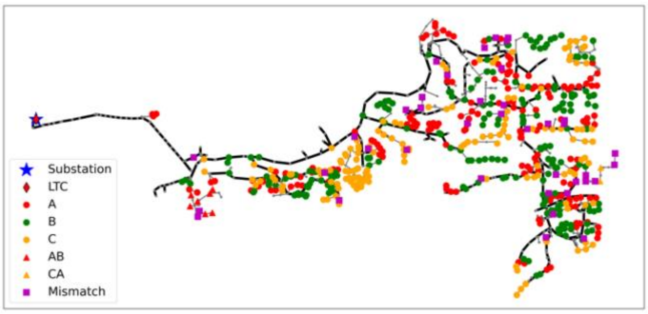

Fig. 7. Geographic distribution of phase identification match/mismatches in Feeder 2.

V. CONCLUSIONS AND FUTURE WORK

This paper proposed a robust phase identification algorithm based on supervised machine learning. The algorithm can be applied to distribution feeders having significant PV generation and a mix of phase-to-neutral and phase-to-phase meter connectivity. The performance of the proposed phase identification algorithm is demonstrated using the AMI data collected in the field from two real distribution feeders of SDG&E having significant PV generation and varied characteristics. The phase identification results are validated against the field-verified phase connectivity data. The results show that the proposed algorithm accurately identifies the phase connectivity, in a high percentage of cases but not universally. In future work, we will explore the sensitivity of the algorithm to different parameters—such as AMI data resolution, duration, PV penetration level—and ways to minimize the training data requirements. We will also consider additional analysis including exploration of other methods and comparison as part of the future work.

ACKNOWLEDGMENT

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Electricity and by California's Electric Program Investment Charge (EPIC) Program in San Diego Gas & Electric Company under a cooperative research-and-development agreement (CRADA) # CRD-17-712. 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.

REFERENCES

[1] L. Liu, Y. Zhao, D. Chang, J. Xie, Z. Ma, Q. Sun, H. Yin, and R. Wennersten, "Prediction of short-term PV power output and uncertainty analysis." in *Applied energy* 228, pp. 700-711, 2018.

[2] "Advanced Metering Infrastructure and Customer Systems," US Department of Energy, Technical Report, Sep. 2016.

[3] J. Peppanen, M. J. Reno, M. Thakkar, S. Grijalva and R. G. Harley, "Leveraging AMI Data for Distribution System Model Calibration and Situational Awareness," in *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 2050-2059, July 2015.

[4] J. Wang, H. Padullaparti, S. Veda, M. Baggu, M. Symko-Davies, A. Salmani, and T. Bialek, "A Machine Learning-based Method to Estimate Transformer Primary-Side Voltages with Limited Customer-Side AMI Measurements," in *2021 IEEE Power & Energy Society General Meeting (PESGM)*, pp. 1-5, 2021.

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

[6] H. Padullaparti, S. Veda, S. Dhulipala, M. Baggu, T. Bialek and M. Symko-Davies, "Considerations for AMI-Based Operations for Distribution Feeders," in *2019 IEEE Power & Energy Society General Meeting (PESGM)*, 2019, pp. 1-5.

[7] W. Luan, J. Peng, M. Maras, B. Harapnuk, and J. Lo, "Smart Meter Data Analytics for Distribution Network Connectivity Verification," IEEE Trans. Smart Grid, vol. 6, p. 1, Jul. 2015.

[8] J. Peppanen, S. Grijalva, M. J. Reno, and R. J. Broderick, "Distribution System Low-Voltage Circuit Topology Estimation using Smart Metering Data," in *IEEE PES T&D Conference and Exposition*, 2016.

[9] J. Wang, X. Zhu, M. Liang, Y. Meng, A. Kling, D. Lubkeman, and N. Lu, "A Data-Driven Pivot-Point-Based Time-Series Feeder Load Disaggregation Method," in *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5396-5406, Nov. 2020.

[10] G. Zhang, G. G. Wang, H. Farhangi, and A. Palizban, "Data Mining of Smart Meters for Load Category Based Disaggregation of Residential Power Consumption," in *Sustainable Energy, Grids and Networks*, vol. 10, pp. 92–103, Jun. 2017.

[11] H. Padullaparti, C. Pisitpol, and S. Surya. "Grid-Edge Voltage Control in Utility Distribution Circuits." in *Power Engineering: Advances and Challenges Part B: Electrical Power*, 251, 2018.

[12] M. H. Wen, R. Arghandeh, A. von Meier, K. Poolla, and V. O. Li, "Phase identification in distribution networks with micro-synchrophasors," in *2015 IEEE Power & Energy Society General Meeting*, 2015, pp. 1–5.

[13] K. J. Caird, "Meter phase identification," Mar. 27 2012, US Patent 8,143,879.

[14] C.S. Chen *et. al.*, "Design of phase identification system to support three-phase loading balance of distribution feeders," in *IEEE Trans. on Industry Applications*, vol. 48, no. 1, pp. 191–198, 2012.

[15] M. Dilek, "Integrated design of electrical distribution systems: Phase balancing and phase prediction case studies," Ph.D. dissertation, Virginia Polytechnic Institute and State University, 2001.

[16] V. Arya et al., "Phase identification in smart grids," in *2011 IEEE Int. Conference on Smart Grid Communications*, Oct. 2011, pp. 25–30.

[17] Z. S. Hosseini, A. Khodaei and A. Paaso, "Machine Learning-Enabled Distribution Network Phase Identification," in *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 842-850, March 2021.

[18] S. Liu *et al.*, "Practical Method for Mitigating Three-Phase Unbalance Based on Data-Driven User Phase Identification," in *IEEE Transactions on Power Systems*, vol. 35, no. 2, pp. 1653-1656, 2020.

[19] F. Therrien, L. Blakely and M. J. Reno, "Assessment of Measurement-Based Phase Identification Methods," in *IEEE Open Access Journal of Power and Energy*, vol. 8, pp. 128-137, 2021.

[20] L. Blakely, M. J. Reno and W. Feng, "Spectral Clustering for Customer Phase Identification Using AMI Voltage Timeseries," in *IEEE Power and Energy Conference at Illinois (PECI)*, 2019, pp. 1-7.

[21] W. Wang and N. Yu, "Advanced metering infrastructure data driven phase identification in smart grid," in *2nd Int. Conf. Green Communications, Computing and Technologies*, 2017, pp. 16–23.

[22] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, "Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 259–265.

[23] F. Olivier, A. Sutera, P. Geurts, R. Fonteneau, and D. Ernst, "Phase Identification of Smart Meters by Clustering Voltage Measurements," in *2018 Power Systems Computation Conference (PSCC)*, 2018.

[24] X. Tang and J. V. Milanovic, "Phase Identification of LV Distribution Network with Smart Meter Data," in *IEEE Power & Energy Society General Meeting (PESGM)*, 2018, pp. 1-5.

[25] W. Wang and N. Yu, "Maximum Marginal Likelihood Estimation of Phase Connections in Power Distribution Systems," in *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 3906-3917, 2020.

[26] Y. Liao *et. al.*, Y. Weng, G. Liu, Z. Zhao *et. al.*, "Unbalanced multi-phase distribution grid topology estimation and bus phase identification," in *IET Smart Grid*, vol. 2, no. 4, pp. 557-570, 2019.

[27] Rodriguez-Galiano, F. Victor, G. Bardan, R. John, C. Mario, and P.R. Juan, "An assessment of the effectiveness of a random forest classifier for land-cover classification." in *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, pp. 93-104, 2012.

[28] G. G. Moisen, "Classification and regression trees," *Encyclopedia Ecology*, vol. 1, pp. 582–588, 2008.

[29] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions [review article]," in *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 41–53, Feb. 2016.

[30] L. Breiman. "Random forests." *Machine learning* 45.1, pp. 5-32, 2001.

[31] M. Pal, "Random forest classifier for remote sensing classification," in *Int. journal of remote sensing*, vol. 26, no. 1, pp. 217-222, 2005.

[32] L. Breiman, H.F. Jerome, A.O. Richard, and J.S. Charles, "Classification and regression trees," CRC Press, 1984.