Voltage-based Machine Learning Algorithm for Distribution of End-users Consumption Among the Phases

Terezija Matijašević, Tomislav Antić, Tomislav Capuder
University of Zagreb Faculty of Electrical Engineering and Computing,
Department of Energy and Power Systems
Zagreb, Croatia
{terezija.matijasevic, tomislav.antic, tomislav.capuder}@fer.hr

***Abstract*—Distribution networks are poorly observable, which is especially evident in different analyses of low voltage (LV) networks, where observability is decreased by the reduced number of smart meters and the lack of network data. Smart meters are in most cases used only for measuring consumption data, while other important information, such as the phase connection of end-users, is not adequately monitored. This aggravates the problem of phase identification for energy utilities, which consequently complicates the numerous calculations required for the smooth operation of the distribution network. In this paper, a comparison of voltage and consumption measurements-based phase identification is presented. Furthermore, a machine learning model based on the voltage measurements is extended to correctly identify the phases of end-users which are three-phase connected to an LV network. The model is tested on a simple 18-node network and the IEEE benchmark network with over 100 nodes and more than 50 end-users. Even though the results show a possibility of using both methods in simpler cases, the voltage measurement-based method is more robust and leads to smaller error in the phase detection problem but also can be extended and used in the case of three-phase connected end-users.**

Keywords*—*end-users consumption, low-voltage networks, machine learning, phase identification

I. INTRODUCTION

The traditional distribution system is characterized by unidirectional power flows and passive end-users at the very end of the electricity supply chain. Due to these characteristic, the aspects of planning and operation of distribution networks did not create serious challenges for Distribution System Operators, analyses were easier and problems were resolved long before they could cause significant issues.

In recent years the transition towards zero-carbon power systems has encouraged the awakening of end-users and their active participation in the operation of distribution networks. Even though in most cases end-users are still placed at the very end of the electricity supply chain, they decide to invest in numerous low-carbon (LC) units, e.g., photovoltaics (PVs), electrified heating, or residential battery storage. Due to the possibility of local production, the traditional consumers turn to prosumers and become active entities in distribution networks. Despite the reduced electricity bills and greenhouse gas emissions, the increased share of distributed energy resources (DERs) creates new technical challenges for DSOs, which need to enable further integration of LC units in order to set the targets defined in different green deals and strategies [1].

Even though the above-mentioned changes create challenges in the power systems planning and operation, the potential of end-users can be encouraged through numerous flexibility and demand response programs in order to mitigate emerging issues. The potential problems that cause insufficient exploitation of the end-users potential are related to the bad observability of distribution networks. Bad observability is caused by the insufficient number of smart meters and the bad quality of collected data. In addition, some other features of smart meters have not yet been fully exploited, including the information of end-user phase connection. However, knowledge of the distribution of consumption among phases, i.e. the affiliation of the end-user to each phase, enables DSOs better observability of network states and near real-time (RT) operation planning [2].

Phase detection is traditionally done by installing additional equipment in the network or by applying mathematical algorithms. Thus, the authors in [3] propose a phase-detection system that measures voltage angles and determines phase connection. Another paper [4] describes an excitation signal injection device for measuring impedances and determining the phase connection. The disadvantage of this method is the requirement for additional devices and communication channels, which represents an expense for DSOs. Another traditional approach involves the implementation of optimization algorithms, such as Mixed Integer Linear Programming (MILP), which aim to minimize the difference between measured values and the estimated

This work has been supported in part by the European Structural and Investment Funds under KK.01.2.1.02.0042 DINGO (Distribution Grid Optimization) and in part by Croatian Science Foundation (HRZZ) and Croatian Distribution System Operator (HEP ODS) under the project IMAGINE – Innovative Modelling and Laboratory Tested Solutions for Next Generation of Distribution Networks (PAR-2018-12).

values obtained from the model [5].

However, the implementation of standard mathematical algorithms is time-complex and error-prone. Due to these problems, a traditional approach could face serious problems related to the computational time and the correctness of the results for large and complex distribution networks. Machine learning (ML) models are the ideal solution for the mitigation of the above-mentioned challenges due to the speed of data processing and the ability to find patterns in the data.

There is a growing interest in the application of ML techniques for detecting challenges in power systems, and the range of applications extends from the general perception of the system to near RT and RT operation [6]. Thus, ML techniques are most often used for various predictions of production of distributed sources [7], consumption [8], prices of electricity but also find their application in other planning and operation aspects [9]–[11].

Many authors have applied ML methods for the phase identification problem in distribution networks. Thus, the authors in [12] apply unconstrained k-means clustering and Principal Component Analysis (PCA) to determine the phase allocation in a given low-voltage (LV) feeder. Voltage measurements of single-phase residential consumers are applied and the algorithm is tested for various PV penetrations in the network. A similar algorithm is developed in [13] where the spectral clustering approach is used to improve the accuracy of the model for large datasets. In addition to voltage measurements, end-user consumption measurements are often applied for the phase identification problem. Thus, in [14] an algorithm based on high-frequency filters and k-means clustering for phase detection of single-phase residential consumers is developed and tested on invalid and inaccurate data. Another work [15] is based on a correlation analysis of variations in end-user consumption, and the amount of data required to obtain accurate results is checked.

However, none of the aforementioned papers deals with the problem of three-phase end-users in an LV distribution network nor tests the algorithm performance on both voltage and consumption measurements. To overcome the identified research gap, an extension of the ML algorithm presented in [14] is developed and presented. The following contributions have been made:

- An extension of the algorithm based on voltage measurements and the comparison of performance with the consumption measurement-based algorithm.
- An extension of the algorithm to solve the problems for three-phase end-users. After the original analysis, which was valid only for single-phase users, the proposed algorithm is tested on networks consisting of several three-phase end-users.
- Testing the sensitivity of an algorithm. The model is investigated in order to determine the resilience to data noise, which can occur due to imperfections in the measuring devices.

The rest of the paper is organized as follows: Section II presents the methodology of the proposed approach. Section III contains case studies used to demonstrate the performance of the proposed approach, while Section IV describes the extension of the algorithm to three-phase end-users and noise-containing data. The conclusions are drawn in Section V.

II. PROPOSED METHODOLOGY

A high number of installed smart meters is a prerequisite for the planning and operation of smart, active distribution networks. Since the role of end-users connected to an LV network is becoming more significant in the energy transition, equipping them with smart meters is especially important. The main function of smart meters is measuring and collecting the energy consumption data, which is then used in numerous power system analyses, but mostly for billing and identification of (non)technical network losses. This means the potential of smart meters often remains unexploited since a lot of other important data are not being observed. One of the missing information is the information about the phase connectivity of the end-users, i.e., phases and devices to which end-users are connected cannot be determined from the initial dataset.

In this paper, an ML-based method is developed for solving the phase identification problem since it successfully mitigates the shortfalls of standard mathematical models, e.g., the large computational time for large networks and large datasets. In order to find patterns in untagged data, it is necessary to implement unsupervised ML, or more precisely data clustering algorithms. The main idea behind such methods is to group objects with similar features into a uniform cluster, while objects with diverse features are classified in different clusters. The goal of clustering algorithms is to increase the distinction between clusters and reduce the difference between the data in a particular cluster.

Popular data clustering methods include the K-Means algorithm, which aims to distribute the objects into K clusters. The first step of this algorithm is to randomly select n data points that will act as the initial clusters' centers, called centroids. Using the appropriate distance formulation, most often the Euclidean distance, the distance between each data point and each of the pre-selected centroids is calculated. The similarity is proportional to the distance obtained, and the point with the smallest distance joins the corresponding cluster. The last step involves changing the centroid by determining the mean of all the associated points which then becomes the new centroid. The steps are repeated until all points are grouped.

Limitations of the traditional K-Means algorithm include the need to specify the K number of clusters and the inability to manually set centroids. Therefore, this paper presents a modified K-Means clustering algorithm,

TABLE I: Modified clustering algorithm

Inputs:
M_i: aggregated consumption (voltage) measurements from transformer substation for phase i P_j: consumption measurements of j^{th} residential consumer V_j: voltage measurements of j^{th} residential consumer
Outputs:
C_k: clusters with associated residential customers
Parameters:
k = 1, 2, 3; index for clusters i = 1, 2, 3; index for aggregated phase measurements at the transformer substation j = 1, 2, ... N; index for residential customers, where N is the total number of customers c_k: centroid of the cluster k
Step 1) Data preprocessing
Step 2) Centroid initialization, $c_k = M_i$
Step 3) Distance calculation, $D_{jk} = f(x_j, c_k)$
Step 4) Minimal distance, $min\ (D_{jk})$
Step 5) Centroid update[a], $if\ i = k \longrightarrow c_k = M_i - P_i$

[a] applicable only in the consumption-based clustering algorithm

which was originally created in [14]. In relation to the mentioned paper, where the algorithm is applied only to consumption measurements, in this paper, the algorithm is modeled to function on consumption measurements as well as on voltage measurements of single-phase end-users. After that, the algorithm is further modified to enable phase-detection in an LV network containing single-phase and three-phase consumers.

At the very core of the algorithm is the fact that the aggregated consumption measurements of the connected phase at the transformer substation are related to the consumption measurements of each end-user and that the voltage measurements of the connected phase at the transformer substation are related to the voltage measurements of each end-user. Such similarity can be presented by the correlation concept (1), where D_{jk} is the distance between j^{th} end-user and k^{th} cluster's centroid and $Corr(x_j, c_k)$ is the Pearson correlation function. This function (2) returns the ordered pair of coefficients between j^{th} measurements x_j and cluster's centroid c_k, where σ_{x_j} and σ_{c_k} are standard deviations of residential customer and cluster's centroid, respectively.

$$D_{jk} = 1 - Corr(x_j, c_k) \tag{1}$$

$$Corr(x_j, c_k) = cov(x_j, c_k)/(\sigma_{x_j} * \sigma_{c_k}) \tag{2}$$

The proposed clustering algorithm consists of four or five steps, depending on whether consumption measurements or voltage measurements are used, and it is presented in the Table I.

A. Consumption-based clustering algorithm

The first step of the proposed algorithm consists of data preprocessing, where excess data and invalid values are removed. Excess data are data that do not affect the algorithm performance (e.g., voltage measurements for the consumption-based clustering algorithm and vice versa), while invalid values most often occur due to errors in measuring devices or extreme network conditions (e.g., outage). This is followed by centroid initialization. In contrast to the traditional K-Means algorithm, where cluster centers are randomly selected, in the proposed modified algorithm initial cluster centers become aggregated consumption measurements from the transformer substation. The third step is determining the distance between each residential customer and each centroid, based on (1), and the result is a distance matrix of size N x 3, where N is the number of residential customers. Then the minimal distance from the given distance matrix is found, which means that the corresponding residential customer can be associated with the proper cluster. In the final step clusters' centers are updated by subtracting the clustered customer's measurements from the previous centroids. Steps 3-5 are continuously repeated until all N customers are clustered.

B. Voltage-based clustering algorithm

The voltage-based clustering algorithm is very similar to the consumption-based algorithm apart from the last step. In this proposed algorithm input dataset is cleaned from outliers and invalid values and excess data are removed. After that, initial centroids become voltage measurements from the transformer substation. Then all customers' distances to each of the centroids are calculated based on equation (1), and when the minimal distance is found, the corresponding customer can be clustered in the appropriate cluster. This is the final step in the voltage-based clustering algorithm, i.e., there is no subtracting measurements and updating centroids. Therefore, initial centroids are also the final centroids, and steps 3 and 4 are being repeated until all residential customers are clustered.

III. CASE STUDIES

Algorithms described in Section 2, are applied to two networks: one is a simple, test network, and one is a large, synthetic benchmark network. In Scenario 1, the algorithm is applied to a smaller synthetic network of five residential consumers, while Scenario 2 considers a modified IEEE-906 LV distribution network [16], presented in Figure 1. This modified network consists of 55 single-phase residential consumers and more than 100 nodes, and is employed for the validation of proposed algorithms on a larger network with a higher number of measurements.

For the input of the model, the consumption dataset of each end-user, voltage measurements of each end-user, and aggregated voltage and consumption measurements collected at the transformer substation are used. Since both networks are synthetic networks, data needed for the validation of the algorithm are obtained using power flow simulations in the *pandapower* programming library [17]. The simulation results are time-series data for two

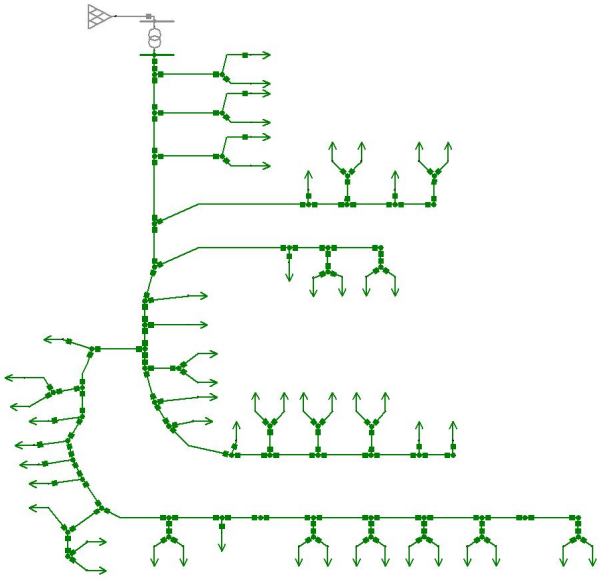

Fig. 1: Modified IEEE-906 LV distribution network

different cases: case A contains a smaller set of measurements, while case B includes a larger set of measurements. The smaller set of measurements involves measurements that correspond to one day in the 1-minute interval, while the larger set of measurements contains measurements of two weeks in the 15-minute interval. Besides, the simulation results (i.e., voltage and consumption measurements) are stored in an external database where the data on the affiliation to a particular phase are not saved, which is an attempt to simulate the actual state of the network.

The smaller network is tested only on a smaller set of measurements, while the larger network is examined on both sets of measurements. Consequently, three different scenarios are created:

- Scenario 1.A: a case of a smaller network and smaller dataset
- Scenario 2.A: a case of a larger network and smaller dataset
- Scenario 2.B: a case of a larger network and larger dataset

A. Scenario 1.A

In order to verify the algorithms, the smaller network with five single-phase end-users and a smaller dataset are used. An example of consumption and voltage measurements for one residential consumer is given in the Figure 2.

Detailed analysis of a given dataset shows that there are no invalid values, and therefore in the first step of the consumption-based clustering algorithm voltage measurement data of the observed consumer are removed. The final results show that the model works properly on such a network and all consumers are correctly classified (Table II).

Similar to the consumption-based clustering algorithm, the voltage-based clustering algorithm removes the consumption data of each end-user. As a result, the model is 100% accurate in determining the phases of given residential consumers (Table II).

B. Scenario 2.A

After the model has been verified on a smaller network, it needs to be tested on a much larger and more realistic network.

The implementation of the consumption-based clustering algorithm described in the Table I shows a significant deviation of the results and real measurements. At the same time, 19 end-users are correctly clustered in phase one (of a total of 21 end-users), 18 end-users are correctly associated with phase 2 (out of 19 end-users), while phase 3 has the worst results with 11 end-users correctly clustered (out of 15 end-users). In percentage terms, phase 1 has an accuracy of 90.5%, phase 2 has an accuracy of 94.7%, and phase 3 has an accuracy of 73.3% (Table II). On the other hand, the voltage-based clustering algorithm provides results with 100% accuracy for all three phases (Table II).

The reason behind these significant discrepancies between voltage-based clustering algorithm and consumption-based clustering algorithm may be in the last step of the algorithm in the Table I. Namely, when updating clusters' centers, the measurements of the selected consumer are subtracted from old clusters' centers and the incorrect consumer clustering will consequently affect the new cluster center.

C. Scenario 2.B

In the last scenario, the larger network is used along with a larger dataset to achieve better model accuracy.

Unfortunately, the consumption-based clustering algorithm gives poorer results, regardless of the size of the used dataset. Thus, 19 out of 21 end-users are correctly clustered in phase 1, 18 out of 19 end-users connected to phase 2 are correctly clustered, while in phase 3 only 8 out of 15 end-users are correctly clustered. Expressed in percentages, the accuracy is 90.5%, 94.7%, and 53.3% for phases 1, 2, and 3, respectively. On the other hand, the voltage-based clustering algorithm results in 100% accuracy for all three phases for this scenario as well (Table II).

These big differences result from the last step of the algorithm (Table I) where the existing cluster centers are updated, as mentioned in Section III.B. Therefore, if one of the consumers is wrongly clustered, it will lead to the wrong centroid value and consequently to wrong clustering.

IV. MODEL EXTENSION

After initial analyses performed on a smaller and a larger network and a smaller and a larger dataset, the

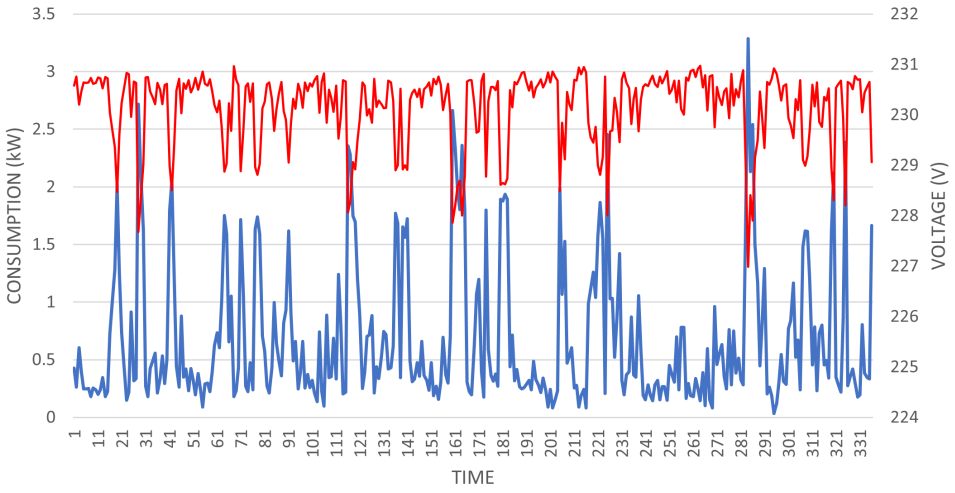

Fig. 2: Consumption and voltage measurements for a single-phase user

TABLE II: Performance of the modified phase detection model

Scenario	Algorithm	End-users	Phase accuracy(%)
scenario 1.A	consumption-based	5 [b]	100%
scenario 1.A	voltage-based	5	100%
scenario 2.A	consumption-based	48 (5, 12, 40, 42, 62, 87, 88) [c]	90.5% 94.7% 73.3%
scenario 2.A	voltage-based	55	100%
scenario 2.B	consumption-based	45 (33, 40, 42, 49, 64, 75, 93, 95, 104, 106)	90.5% 94.7% 53.3%
scenario 2.B	voltage-based	55	100%

[b] green: total number of correctly allocated end-users
[c] red: ID of the incorrectly clustered end-user

model's performance is tested for scenarios that may occur in LV distribution networks, namely noise generated by measuring devices and the appearance of the three-phase residential consumers.

Noises caused by the imperfections of measuring devices are a common occurrence in distribution networks and it is necessary to check the performance of the algorithm on such data. Such imperfections are simulated in the *pandapower* tool where end-user consumption measurements are increased by 1%. The obtained results together with a smaller network are used for the verification of the algorithm (Scenario 3.A), where the correct clustering of all consumers is achieved for both the consumption-based clustering algorithm and the voltage-based clustering algorithm. After that, the noised dataset is applied to a larger network and the voltage-based clustering algorithm (Scenario 3.B). Here, regardless of noises and irregularities in the dataset, accurate clustering of all users by phases is achieved (Table III).

In addition to noises, three-phase end-users are a frequent occurrence in the distribution network, whose consumption is measured at all three phases. It is assumed that smart meters of such consumers do not collect information about the phase connection (i.e., measurements P_{1j}, P_{2j}, and P_{3j} for j^{th} consumer do not have to correspond to phases 1, 2, and 3). Therefore, the original voltage-based clustering algorithm is extended with a step where the phase connection of three-phase end-users is checked using equations (1) and (2) provided that each of the measurements of these users can belong to one of the clusters. The algorithm is tested on a larger network (Scenario 4.A), where several single-phase consumers are randomly selected to become three-phase consumers with a manually set phase connection. This means that voltage measurements of j^{th} end-user, although labeled V_{1j}, V_{2j}, and V_{3j}, do not necessarily correspond to phases 1, 2, and 3. The analysis of the algorithm shows the correct clustering of three-phase and single-phase residential consumers (Table III).

V. Conclusion

In this work, a phase identification approach in LV distribution networks is proposed. A similar method is applied in [14], but the impact of three-phase residential consumers, as well as the performance of the algorithm on consumption and voltage measurements, has not been

TABLE III: Performance of the phase detection model on noisy data and three-phase residential consumers

Scenario	Algorithm	End-users	Phase accuracy(%)
scenario 3.A	consumption-based	5	100%
scenario 3.A	voltage-based	5	100%
scenario 3.B	voltage-based	55	100%
scenario 4.A	voltage-based	55	100%

investigated. Therefore, this paper presents a new, modified algorithm that enables correct phase identification in unbalanced distribution networks with both single-phase and three-phase end-users.

The performance of the approach is demonstrated on two different networks and two datasets. The first network is a test network that contains 5 residential consumers and is applied only for easier understanding of the proposed algorithm. The second network is a larger synthetic benchmark network with over 50 end-users.

In addition, two different time periods of data are used and finally, each scenario is tested with end-user consumption and voltage measurements. Conducted scenario analyses on single-phase residential consumers show the sufficiency of the algorithm for phase identification of consumers using voltage measurements. In these cases, all end-users are properly grouped into the corresponding clusters. On the conrary, the consumption-based clustering algorithm gives significantly worse results (around 50% accuracy for phase 3 in one of the scenarios).

Since residential consumers in real-world distribution networks usually have either single-phase or three-phase connections, the existing algorithm is expanded to allow analysis with three-phase end-users. The algorithm is verified on a larger network containing both types of consumers and satisfactory results are achieved.

The proposed phase identification method can be applied in real-world distribution networks where there is no information on phase connection to improve network observability from available data. Furthermore, with the knowledge of the distribution of residential consumers among phases, i.e. the affiliation of the end-user to each phase, it is possible to perform additional analyses to achieve optimal operation and reduce techno-economic losses.

Further research in this area includes improving the consumption-based clustering algorithm and evaluating the algorithms on LV networks rich in distributed energy sources (PV, electric vehicles, energy storage systems, etc.).

REFERENCES

[1] "Delivering the European Green Deal," jul 2021. [Online]. Available: https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal/delivering-european-green-deal_en

[2] D. K. Chembe, "Reduction of Power Losses Using Phase Load Balancing Method in Power Networks," in *Proceedings of the World Congress on Engineering and Computer Science*, vol. 1, 2009, pp. 20–22.

[3] C.-S. Chen, T.-T. Ku, and C.-H. Lin, "Design of phase identification system to support three-phase loading balance of distribution feeders," in *2011 IEEE Industrial and Commercial Power Systems Technical Conference*. IEEE, 2011, pp. 1–8.

[4] Z. Shen, M. Jaksic, P. Mattavelli, D. Boroyevich, J. Verhulst, and M. Belkhayat, "Three-phase AC system impedance measurement unit (IMU) using chirp signal injection," in *2013 Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*. IEEE, 2013, pp. 2666–2673.

[5] 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.

[6] M. S. Ibrahim, W. Dong, and Q. Yang, "Machine learning driven smart electric power systems: Current trends and new perspectives," *Applied Energy*, vol. 272, p. 115237, 2020.

[7] A. Nespoli, E. Ogliari, S. Leva, A. Massi Pavan, A. Mellit, V. Lughi, and A. Dolara, "Day-ahead photovoltaic forecasting: A comparison of the most effective techniques," *Energies*, vol. 12, no. 9, p. 1621, 2019.

[8] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2017.

[9] S. Motepe, A. N. Hasan, and R. Stopforth, "Improving load forecasting process for a power distribution network using hybrid AI and deep learning algorithms," *IEEE Access*, vol. 7, pp. 82 584–82 598, 2019.

[10] S. Zhang, Y. Wang, M. Liu, and Z. Bao, "Data-based line trip fault prediction in power systems using LSTM networks and SVM," *IEEE Access*, vol. 6, pp. 7675–7686, 2017.

[11] L. He, S. Rong, and C. Liu, "An Intelligent Overcurrent Protection Algorithm of Distribution Systems with Inverter based Distributed Energy Resources," in *2020 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 2020, pp. 2746–2751.

[12] A. Simonovska and L. F. Ochoa, "Phase Grouping in PV-Rich LV Feeders: Smart Meter Data and Unconstrained k-Means," in *2021 IEEE Madrid PowerTech*. IEEE, 2021, pp. 1–6.

[13] L. Blakely, M. J. Reno, and W.-c. Feng, "Spectral clustering for customer phase identification using AMI voltage timeseries," in *2019 IEEE Power and Energy Conference at Illinois (PECI)*. IEEE, 2019, pp. 1–7.

[14] Z. S. Hosseini, A. Khodaei, and A. Paaso, "Machine learning-enabled distribution network phase identification," *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 842–850, 2020.

[15] V. A. Jimenez, A. Will, and S. Rodriguez, "Phase identification and substation detection using data analysis on limited electricity consumption measurements," *Electric Power Systems Research*, vol. 187, p. 106450, 2020.

[16] M. A. Khan and B. Hayes, "A Reduced Electrically-Equivalent Model of the IEEE European Low Voltage Test Feeder," Oct 2021.

[17] L. Thurner, A. Scheidler, F. Schäfer, J.-H. Menke, J. Dollichon, F. Meier, S. Meinecke, and M. Braun, "pandapower—an Open-Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6510–6521, 2018.