Sustainable Energy, Grids and Networks 46 (2026) 102321

Contents lists available at ScienceDirect

Sustainable Energy, Grids and Networks

journal homepage: www.elsevier.com/locate/segan

A data-driven approach for topology correction in low voltage distribution networks with photovoltaics

Dong Liu [a], Sander Timmerman [b], Yu Xiang [b,c], Ensieh Hosseini [b], Peter Palensky [a], Pedro P. Vergara [a,*]

[a] Intelligent Electrical Power Grids (IEPG) Group, Delft University of Technology, Delft, 2628CD, the Netherlands
[b] Alliander N.V., Arnhem, 6812 AH, the Netherlands
[c] Electrical Energy Systems group, Eindhoven University of Technology, Eindhoven, 5612AZ, the Netherlands

ARTICLE INFO

Keywords:
Distribution networks
Robustness
Supervised learning
Topology observability
Unbalanced networks

ABSTRACT

To correct outdated and incomplete topologies in low-voltage distribution networks (LVDNs) using only voltage magnitude measurements, a data-driven approach is developed by integrating machine learning algorithms with correlation analysis. Similar to existing data-driven topology identification and correction methods, the proposed approach exploits smart meter data to infer topology information. However, unlike many conventional approaches that require repeated preprocessing, multiple data sources, or separate procedures for different topology elements, it provides a unified framework that consistently uses the same up-to-date voltage magnitude dataset across all processing stages. Specifically, switch states are identified via supervised learning, while user–feeder connections and customer phase labels are refined using a modified hierarchical clustering algorithm. To address the similarity among smart meter data induced by distributed photovoltaic (PV) systems, a time-based data selection strategy is incorporated into the correlation analysis. The feasibility and robustness of the proposed approach are validated using modified real-world LVDNs and multiple incomplete smart meter datasets collected from customers in the Netherlands. The results demonstrate that the proposed approach can effectively mitigate the impact of PV-induced similarity on phase identification and improve topology correction performance. Although the approach is designed for topology correction rather than full topology reconstruction, the corrected topology improves network observability and supports distribution system operators in load balancing and PV integration.

1. Introduction

1.1. Motivation

Most existing phase balancing and topology reconfiguration problems are formulated as mixed-integer optimization problems that depend on network topologies [1–3]. However, these topologies are often inaccurate and outdated for distribution system operators (DSOs) due to missing recordings, topology maintenance and reconfiguration, such as congestion management [4]. Thus, the topology of the low-voltage distribution network (LVDN) needs to be checked and corrected when it is outdated. The increasing uncertainty of distributed energy resources (DERs), including household photovoltaic (PV), heating pumps, etc., impacts the frequency of topology reconfiguration and challenges the correction of the low-voltage distribution network topology [5–7]. Moreover, the available smart meter (SM) datasets are often limited due to privacy concerns and random communication channel failures, challenging the topology correction [8–10]. Synthetic European networks and benchmark models presented in [11,12] are benchmarks for research but insufficient to represent the diversity of European LVDNs for practical use by DSOs (e.g., state estimation). Thus, practical topology identification and correction approaches are required for real-time topology updating for active management of LVDNs.

1.2. Literature review

The topology identification problem in distribution networks (DNs) could be briefly classified into three categories: (1) switch state identification, (2) connection lines and their parameter identification, and (3) phase identification. The switch state identification is treated as a classification problem in [13], solved by a trained deep neural

* Corresponding author.
Email address: P.P.VergaraBarrios@tudelft.nl (P.P. Vergara).

https://doi.org/10.1016/j.segan.2026.102321
Received 11 November 2025; Received in revised form 13 May 2026; Accepted 15 May 2026
Available online 21 May 2026

Acronyms

ANN	Artificial neural network
CC	Complex feeder from CN1
CC_3_NU	Complex feeder with three-phase non-uniform loads
CC_PV	CN1 with PV at each house
CN1	Network with two complex-structured feeders
CN2	Network with one simple- and one complex-structured feeder
DERs	Distributed energy resources
DNs	Distribution networks
DSO	Distribution system operator
EV	Electric vehicles
GIS	Geographic Information System
GNB	Gaussian Naive Bayes algorithm
HC	Hierarchical clustering algorithm
KNN	K-nearest neighbors algorithm
LVDN	Low-voltage distribution network
MFP	Modified PCC-based Fisher z-transformation
MV/LV	Medium/Low voltage
PCC	Pearson correlation coefficient
PV	Photovoltaic
RF	Random forest algorithm
SA	Saliency analysis
SC	Simple feeder from SN
SC_3_NU	Simple feeder with three-phase non-uniform loads
SC_PV	SN with PV at each house
SM	Smart meter
SN	Network with two simple-structured feeders
SNB	Network with two simple-structured feeders with more branches
SNL	Network with two longer simple-structured feeders
SVM	Support vector machine algorithm

network and providing quick estimations under the selected features. To recover accurate connection lines between transformers and customers, a probabilistic graphical model-based method is developed in [14] by restricting the propagation of SM data errors. In [15], a hybrid data-driven approach integrating a partial correlation analysis strategy and a linear regression model is proposed to estimate topology from limited SM data. Based on an alternating direction method of multipliers, a robust topology identification method is proposed to estimate the line connections and their parameters using μ-PMUs and SM datasets [16]. A two-step approach is introduced to identify the topology of LVDNs with multiple meter types, which provide different resolution datasets, mitigating the dependence on the SMs [17]. Related to phase identification, current approaches rely on voltage magnitude data and demand profiles [18]. Since the correlation of voltage magnitude data from the same phase is stronger than that from different phases, correlation analysis is normally used. Given voltage magnitude datasets, spectral clustering and k-means clustering approaches are adopted to identify customer phase labels [19]. The mixed integer linear programming model in [20] is constructed to identify the phase labels of single-phase customers by minimizing the gap between transformers' measurements and the estimated data (e.g., power and voltage magnitude) in each phase. Saliency analysis (SA) and statistical tests are combined in [21] to guarantee the identification accuracy of the data pairs with weak correlations. Fig. 1 presents several underground feeder deployment patterns in the Netherlands [22], which indicate the challenges of LVDN topology identification and correction in practice, e.g., multiple cables under the same street. In addition, some of the above approaches are sensitive to missing data and measurement errors, e.g., SA-based approaches. Furthermore, the user-feeder connection identification issue is rarely considered in the aforementioned papers due to the assumption that one main feeder connects to the transformer.

To address this issue, based on a modified genetic algorithm, a novel data-driven phase identification approach is introduced to identify the phase labels under incomplete datasets [23]. Based on incomplete and asynchronous SM data, an optimization-based data-driven approach is constructed to recover the approximate topologies in LVDNs with open datasets [24]. Voltage magnitude fluctuations among customers connected to different phases may exhibit similarities due to variations induced by the medium-voltage (MV) network or similar load patterns. To mitigate this effect, a high-pass filter is applied during time-series data pre-processing in the phase identification approach [25]. Additionally, to mitigate the impact of electrical distance on the clustering of SM time-series data, a power-band-based strategy is proposed to select datasets, thereby improving the precision of phase identification and transformer-to-meter matching [26]. An optimization model-based approach is introduced in [27] to estimate customer phase labels and switch states in LVDNs with limited available recordings. SA strategies are integrated to extract underlying features, thereby enhancing the accuracy of phase-label identification [28]. Nevertheless, the impacts of household DERs have not been sufficiently addressed in the above research. With the increasing penetration of household PV installations, the similarity among household voltage magnitudes is expected to increase significantly. Specifically, within the same LVDN, PV systems are exposed to similar solar irradiance conditions, resulting in highly

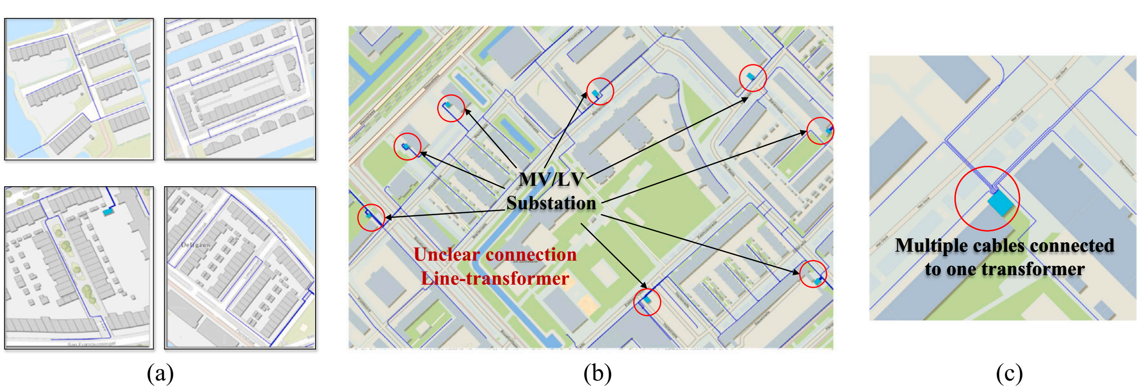

Fig. 1. Deployment patterns of LVNDs: (a) feeder deployment under one street, (b) distribution of MV/LV transformers and (c) underground feeders connected to the same transformer.

correlated PV power generation patterns. This leads to similar voltage magnitude distributions among households, challenging the feasibility of correlation analysis-based approaches.

Topology identification generally aims to determine the structure of an LVDN when the network topology is completely unknown. In this setting, the potentially valuable information contained in previously recorded topologies is typically not utilised. In contrast, topology correction focuses on updating only the incorrect or outdated components of an existing topology by leveraging limited recent data while fully exploiting the information available in historical topology records. While topology identification has attracted significant research attention in recent years, topology correction remains relatively underexplored. To dynamically update the connections and the parameters after topology configuration, a graph fusion network-based approach presented in [29] is used to extract correct connection lines. However, the adjacency matrix is also required. For networks with low observability, a distance matrix is introduced in [30] to correct the topology and is also effective in detecting latent nodes positioned between observed nodes along network paths (i.e., intermediate latent nodes rather than endpoints). A physics-informed neural network is constructed to estimate the topology by representing the nodes without meters, whose parameters are adjusted by a reinforcement learning-based approach [31]. A data-driven approach is introduced in [32] to correct the connection between meters and transformers in GIS systems using voltage magnitude datasets. A similar work in [33] established a data-driven model based on the impedance matrix to correct topology and identify the impedance of connection lines using SM data. A data-driven approach in [34] is introduced to infer latent nodes and line parameters for outdated topologies using power and voltage magnitude profiles. Nevertheless, different approaches often rely on different assumptions and require different types of input SM data. As a result, DSOs need to prepare multiple datasets to perform separate identification tasks.

1.3. Contributions

To overcome these limitations, we propose a data-driven approach for topology correction in LVDNs using incomplete voltage data. Table 1 highlights the key differences between the proposed approach and existing approaches. In this work, voltage magnitude measurements are assumed to be available from household smart meters. By relying solely on voltage magnitude data, this approach preserves privacy and minimizes SM updates, as it eliminates the need for phase angle measurements from households. The main contributions are summarized as follows:

- A systematic data-driven topology correction framework is proposed to update outdated LVDN topology records through three sequential steps: (1) switch state identification, (2) user-feeder connection correction, and (3) phase label identification. The proposed framework relies solely on voltage magnitude measurements from SMs, enabling the identification and correction of multiple topology components using a unified input dataset. Compared with conventional approaches that require complete SM datasets or different types of measurements for separate tasks, the proposed method allows DSOs to correct topology information using limited and noisy SM voltage data, even when household measurements are incomplete.
- A switch state identification strategy is developed by encoding the operational states of switch bars (including multiple switchgears) using label encoding rules. In the topology correction context, historical topology records and network operation logs provide labelled samples corresponding to different switch configurations. This formulation converts the identification of multiple switchgear states into a supervised classification problem that can be efficiently solved using existing machine learning algorithms. Furthermore, a modified distance function is incorporated into traditional hierarchical clustering (HC)–based correlation analysis to enhance the distinction between geographically close loads belonging to different phases or cables, thereby improving the accuracy of user–feeder and phase identification.
- To mitigate the influence of PV power on voltage-based correlation analysis, a time-based SM data selection strategy is proposed to extract voltage measurements during periods when household loads dominate voltage variations, and PV generation remains lower than the local demand. When historical PV generation data are available, the selection can be refined by identifying intervals with low PV injection. Otherwise, voltage datasets after sunset are used for practical utilization. This strategy reduces the distortion of correlation patterns caused by PV generation and improves the accuracy of topology correction without requiring explicit PV modeling.

The remainder of this paper is organised as follows: Section 2 details the framework of the proposed approach, including identification of switch states, user-feeder connections, phase labels, and voltage data

Table 1
Comparison of representative topology identification and correction research in LVDNs.

Method	Ref.	Input Data				Robustness			Output			
		V	P/I	Limited topology	GIS data	DERs impact	SM errors	Missed data	Switch state	user-feeder connections	Phase labels	latent nodes
ANN	[13]	×	✓	×	×	✓	×	×	✓	×	×	×
	[31]	✓	✓	✓	×	×	×	✓	×	✓	×	×
	[29]	✓	✓	✓	×	×	✓	×	×	✓	×	✓
Linear regression	[14]	✓	×	✓	×	✓	✓	×	×	✓	×	×
	[35]	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	×
Optimization based	[24]	✓	✓	×	✓	×	✓	✓	×	✓	×	×
	[20]	✓	✓	✓	×	✓	✓	×	×	×	✓	×
	[34]	✓	✓	✓	×	✓	✓	×	×	✓	✓	×
clustering-based	[19]	✓	×	×	×	×	×	×	×	×	✓	×
	[26]	✓	✓	×	×	×	×	×	×	✓	✓	×
	[32]	✓	✓	×	✓	×	×	×	×	✓	✓	×
	[33]	✓	✓	✓	×	×	✓	×	×	✓	×	×
Heuristic Algorithms	[23]	×	✓	×	×	✓	✓	✓	×	×	✓	×
	[30]	✓	✓	✓	×	×	✓	×	×	✓	×	✓
Our Method		✓	×	✓	×	✓	✓	✓	✓	✓	✓	×

Note: Topology reconfiguration studies are not included in this comparison, as they assume a known network topology and focus on optimal network operation (e.g., loss reduction and voltage regulation), rather than topology identification or correction.

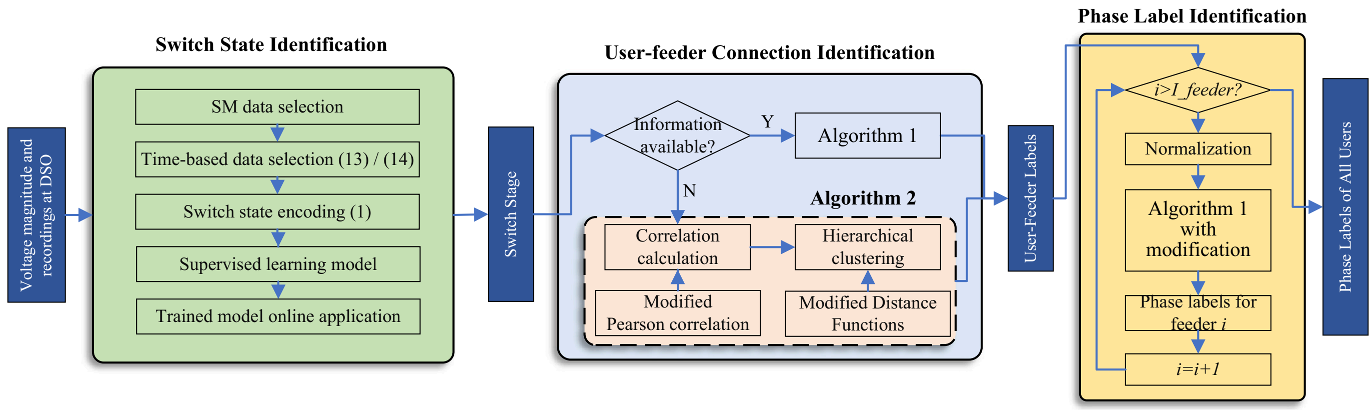

Fig. 2. Proposed topology correction approach composed of three steps: (1) switch state identification, (2) user-feeder connection identification and (3) phase label identification.

pre-processing. Section 3 presents the case studies and results. Finally, Section 4 summarizes the conclusions of this work.

2. Topology correction approach

The proposed topology correction approach is illustrated in Fig. 2, which is composed of three steps: (1) switch state identification, (2) user-feeder connection identification and (3) phase label identification. An illustrative example is presented in Fig. 5, which shows the outputs of each step. The process is triggered when the latest voltage magnitude measurements become available or an outdated topology is detected. The input includes customer voltage magnitude and incomplete recordings of the user-feeder connection, and the output is the corrected topology with accurate customer's user feeder connections and phase labels, including single-phase and three-phase customers.

2.1. Switch state identification

In LVDNs, multiple switch bars exist, each containing a different number of switches. Each switch connects two cables (e.g., A and B), which are linked to different transformers (e.g., AT and BT). Under normal conditions, the switch remains open. When transformer AT becomes overloaded, the switch is closed, shifting the load on cable A to transformer BT. Therefore, the switch state plays a crucial role in maintaining network efficiency and enabling effective load shifting. An example of switch deployment is illustrated in Fig. 3. This non-uniform deployment poses challenges for accurately identifying the state of individual switches using SM data. To address this, a random forest (RF) algorithm is employed, leveraging voltage magnitude datasets to determine the state of switches within each switch bar. First, inspired by Label Encoding rules, the switch states in one bar are encoded into single-digit label formats, i.e., each combination of the switch states in the switch bar is uniquely mapped to a single digit label. The number of switches is assumed to be accessible for the DSO. For example, when there are 3 switches in the switch bar, the switch states SS can be encoded as follows:

$$f(V) = \begin{cases} 0, & \text{if} \quad SS = [1,1,0] \\ 1, & \text{if} \quad SS = [1,0,1] \\ 2, & \text{if} \quad SS = [0,1,1] \end{cases} \tag{1}$$

After the switch states are encoded, the switch state identification problem in each bar is transformed into a standard multi-class classification task. The input for this classification task consists of time-series voltage magnitude datasets. However, the presence of missing data from SMs or unmetered households results in varying numbers of available SMs on each feeder. Feeders with a higher number of available SMs provide more data for the supervised learning process, potentially introducing bias in classification and leading to accuracy loss. Besides, when switch states remain unchanged for extended periods (e.g., a state $[1,0,1]$ resulting in an excessive number of samples labelled as 1), the dataset becomes unbalanced. This imbalance further affects classification accuracy, posing a common challenge in machine learning.

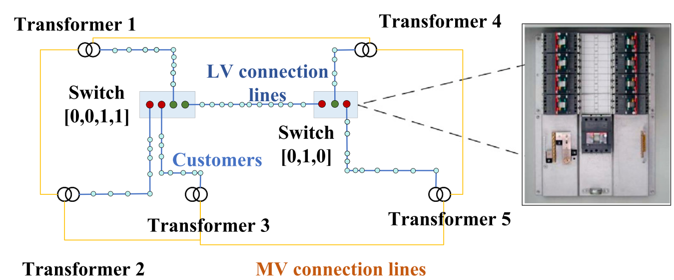

Fig. 3. Illustrative example of switch deployment and states in LVDNs: 1 represents the activated switch, and vice versa [36].

To address the above-mentioned challenges, an SM selection process is employed before switch state identification. This selection process consists of two key steps:

- *Uniform SM number selection across feeders.* M represents the minimum number of available SMs across all feeders. To ensure consistency, the number of SMs used for data extraction from each feeder is fixed at M. Specifically, the M SMs closest to the switch bar are selected as representative features of that feeder.

$$M = \min\{m_1, m_2, \dots, m_K\}, \tag{2}$$

  where m_i is the number of available SMs on feeder i, and K is the total number of feeders that are connected to the same switch bar.
- *Balanced dataset construction.* Given the encoded switch state labels (i.e., 0, 1, 2, …), an equal number of samples is selected for each category. For each encoded class C_i, the same number of samples S is randomly selected.

$$S = \min\{|C_1|, |C_2|, \dots, |C_J|\}, \tag{3}$$

  where, $|C_j|$ represents the number of available samples for class j, and J is the total number of classes.

Through this selection process, the impact of the varying number of available SMs and class imbalance is minimized, leading to a more balanced and effective training dataset for the multi-class classification task. The selected voltage measurements can be represented in a matrix form V_{train}, forming the input for the multi-class classification task.

Each row in matrix V_{train} represents training one sample with an encoded SS label. $v^{s,C_j}_{m_k,M}$ represents the voltage magnitude of the house M connected to the feeder with m_k houses when the switch state belongs to category C_j.

$$V_{train} = \begin{bmatrix} v^{1,C_1}_{m_1,1} & \cdots & v^{1,C_1}_{m_1,M} & \cdots & v^{1,C_1}_{m_k,1} & \cdots & v^{1,C_1}_{m_k,M} \\ \vdots & & \vdots & & \vdots & & \vdots \\ v^{S,C_1}_{m_1,1} & \cdots & v^{S,C_1}_{m_1,M} & \cdots & v^{S,C_1}_{m_k,1} & \cdots & v^{S,C_1}_{m_k,M} \\ \vdots & & \vdots & & \vdots & & \vdots \\ v^{1,C_J}_{m_1,1} & \cdots & v^{1,C_J}_{m_1,M} & \cdots & v^{1,C_J}_{m_k,1} & \cdots & v^{1,C_J}_{m_k,M} \\ \vdots & & \vdots & & \vdots & & \vdots \\ v^{S,C_J}_{m_1,1} & \cdots & v^{S,C_J}_{m_1,M} & \cdots & v^{S,C_J}_{m_k,1} & \cdots & v^{S,C_J}_{m_k,M} \end{bmatrix} \tag{4}$$

The RF algorithm excels in multi-class classification by handling high-dimensional data, capturing complex feature interactions, and maintaining accuracy despite noise. Its ensemble learning approach—building diverse trees with random feature selection—prevents overfitting and enhances generalisation. RF also provides feature importance insights, making it ideal for switch state identification. Given a dataset $D = \{(X_1, y_1), \ldots, (X_n, y_n)\}$ where X_i represents the feature vector (i.e., each row in V_{train}) and y_i is the encoded label by (1), the general process of RF training reads as:

1. Draw B bootstrap samples from the training data V_{train} and the corresponding labels.
2. Grow an unpruned decision tree for each sample by randomly selecting a subset of features at each split.
3. For classification, each tree votes on the most common class label, and the RF outputs the majority vote:

$$\hat{y} = \arg\max_j \sum_{b=1}^{B} \mathbf{I}(h_b(x) = j), \tag{5}$$

where $h_b(x)$ is the prediction of the b-th tree and $\mathbf{I}(\cdot)$ is an indicator function that equals 1 if the condition is met and 0 otherwise. Given a voltage sample V_i, each trained decision tree independently predicts $h_b(x)$ and the final labels of the switch are obtained by the majority voting mechanism.

In the topology correction context, the use of supervised learning for switch state identification is practical. Unlike topology identification problems that assume a completely unknown network topology, topology correction relies on existing topology records and historical operation logs maintained by DSOs. Although these records may contain outdated information, they still provide useful prior knowledge about typical switch configurations. Historical voltage data collected under known operating conditions can therefore be associated with these configurations to form labelled samples for model training. Moreover, the number of feasible switch configurations within a feeder is usually limited. For example, as illustrated in the right part of Fig. 3, there are only seven possible configurations for the switch cabinet with three switches (i.e., (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), and (1,1,1)), excluding the all-open state. Therefore, supervised learning provides a feasible solution for identifying switch states within the proposed topology correction framework.

2.2. User-feeder connections identification

The connection relations between customers and the feeders might be outdated and not always available for DSOs, which is ignored by multiple similar works [24,31], especially for houses and multi-layer buildings that are close to the streets with multiple deployed underground feeders. The practical case shown in Fig. 4 demonstrates the common challenge of ambiguous user-feeder connections and phase labelling, illustrating the physical complexity that makes accurate network topology identification (i.e., step (2) and (3)) difficult. Those unclear

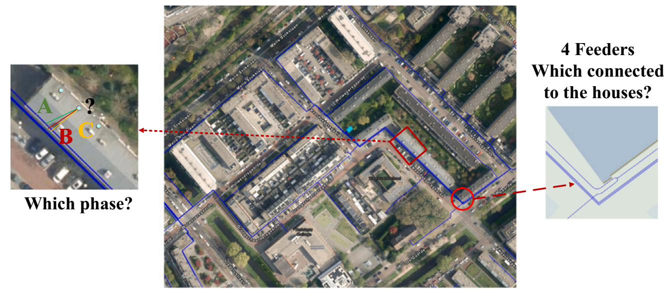

Fig. 4. A practical LVDN with multiple parallel feeders along the same street.

user-feeder connections also impact the phase label identifications. A correlation approach is introduced to reveal the user-feeder connections, which only relies on time-series voltage magnitude datasets. Two practical cases are addressed: 1) user-feeder connection identification without reference and 2) user-feeder connection identification with limited, accurate recordings.

2.2.1. Connection identification without recordings

The proposed user-feeder connection identification algorithm combines the Hierarchical Clustering (HC) algorithm and the correlation analysis based on the modified Pearson Correlation Coefficient (PCC). Algorithm 1 depicts the process of the proposed approach. The voltage magnitude dataset is stored in a matrix V, which is used as the input for the algorithm. Vector $V_n = [v_{n,1}, v_{n,2}, \ldots, v_{n,\mathcal{T}}]^T$ represents the time series voltage magnitude at household n. $v_{n,t}$ represents the voltage magnitude at household n at time t. T is the measurement period of the SM data.

$$PCC(V_n, V_m) = \frac{\sum_{t=1}^{T}(v_{n,t} - \bar{V}_n)(v_{m,t} - \bar{V}_m)}{\sqrt{\sum_{t=1}^{T}(v_{n,t} - \bar{V}_n)^2}\sqrt{\sum_{t=1}^{T}(v_{m,t} - \bar{V}_m)^2}} \tag{6}$$

where $\bar{V}_n$ and $\bar{V}_m$ are the mean values of the samples V_n and V_m, respectively.

Usually, the PCC of voltage magnitudes in the same feeders shows higher values compared to those in different feeders [32]. To amplify the difference between voltage correlations from different feeders, a modified Fisher z-transformation function is employed:

$$FPCC(V_n, V_m) = \ln[\frac{1 + PCC(V_n, V_m)}{1 - PCC(V_n, V_m) + \alpha}], \tag{7}$$

where α added to the denominator is used to avoid an infinite value of $FPCC(\cdot)$ and to control its distribution region.

Algorithm 1: MFP-based User-Feeder Identification without Available Labels.

Input: V, K_s, N
for $n \leq N$ **do**
  for $m \leq N$ **do**
    PCC(V_n, V_m) via Eq. (6)
    FPCC($\rho_{n,m}$) via Eq. (7)
  end
end
for $i \leftarrow 1$ ***to*** N **do**
  for $j \leftarrow i+1$ ***to*** N **do**
    $D_{ij} \leftarrow$ MFP(V_n, V_m) via Eq. (8)
  end
end
Obtain $\mathbf{D} \leftarrow \{D_{ij}\}$
Clusters C: ← Traditional HC and $\mathbf{D}$
labels ← Assigned labels to households based on C
Output: labels, Cluster: C_1, …, C_{K_s}

To amplify the difference between voltages from different feeders, a modified $FPCC(\cdot)$ value (MFP) is employed to replace the Euclidean distance in the traditional HC algorithm.

$$MFP(V_n, V_m) = 1 - \min\left\{\frac{1}{4}\ln(1 + e^{a+4\cdot FPCC(V_n,V_m)}), 1\right\} \quad (8)$$

The second term in (8) is a modified likelihood function $F(\cdot)$ [37], which is used to calculate the $MFP(\cdot)$ between each sample V_n and the rest of the samples. Since the range of the function $F(\cdot)$ is $(0, 1]$, the range of the distance metric $MFP(\cdot)$ is correspondingly $[0, 1)$. A value of $MFP(V_n, V_m)$ closer to 0 indicates a higher likelihood that the two samples V_n and V_m are collected from the same substation.

Given the switch state obtained from Section 2.1, the feeder-transformer connection is available, and the number of feeders connected to the same transformer is assumed to be accessible to the DSO and taken as the number of clusters (i.e., K_s). Lines 1–7 in Algorithm 1 are used to obtain the MFP of the input voltage magnitude dataset. Lines 8–12 are used to find the link matrix of matrix D. From line 13 onward, the input data are clustered into a K_s cluster using the obtained linkage matrix. If two users share the same label, they are connected to the same underground feeder, and vice versa.

2.2.2. Connection identification with recordings

When limited user-feeder connections are known, the user-feeder identification problem becomes more manageable compared to the fully unknown scenario. There are two sources of accurate recordings:

- Historical recordings from the DSO's datasets.
- Identified connections with high confidence from Section 2.2.1.

Those available recordings transform the identification problem from an unsupervised learning problem to a supervised learning problem. Various supervised learning algorithms could be used to identify unknown connections, e.g., ANN, KNN, etc. The traditional KNN algorithm and the simplified MFP-based approach are employed and compared. The KNN algorithm is a widely used supervised learning method for classification tasks. It classifies an unknown sample by considering the labels of its nearest neighbours in the feature space, which is similar to identifying the user-feeder connection based on their location and the neighbours with known labels. Given a set of labelled training samples, the classification of a new sample is determined based on the majority class among its k nearest neighbours. The distance between samples is commonly measured using the Euclidean distance

$$d(V_n, V_m) = \sqrt{\sum_{t=1}^{T}(v_{n,t} - v_{m,t})^2}. \quad (9)$$

Once the distances between the new sample and all labelled samples are computed, the k nearest neighbours are selected. The final

Algorithm 2: MFP-based User-Feeder Identification with Available Labels.

Input: V, N_k, N_{un}, L_{un}, k
for $n \geq N_{un}$ **do**
  for $m \leq N_k$ **do**
    $D_0(n, m) \leftarrow D(V_n, V_m)$ via Eq. (8)
  end
  $\hat{y} = \arg\max_c \sum_{n\in\mathcal{N}_k} \mathbb{I}(y_n = c)\ L_{un}[n] \leftarrow \hat{y}$
end
Output: L_{un} for households with unknown labels

classification is determined using majority voting.

$$\hat{y} = \arg\max_{c} \sum_{n\in\mathcal{N}_k} \mathbf{I}(y_n = c) \quad (10)$$

$\mathcal{N}_k$ denotes the set of k nearest neighbors, y_i represents the label of neighbour i. In user-feeder connection identification, the KNN algorithm assigns users without clear connections to the most probable feeder based on similarity to datasets from users who have available user-feeder connection recordings. Motivated by the traditional KNN approach, the proposed MFP-based approach is simplified to reveal the unknown user-feeder connection in DNs, which is depicted in Algorithm 2. $\mathcal{N}_{un}$ is the number of users with unclear connections, and $\mathcal{L}_{un}$ is their identified label set.

2.3. Phase label identification

After the second step in the proposed framework (i.e., user-feeder connection identification in Section 2.2), the third step is to identify the phase labels of the customer connection for each feeder, which is a typical 3-cluster classification problem. The phase identification approach is introduced by modifying Algorithm 1, i.e., removing Fisher z-transformation and using normalized voltage data V^* as inputs.

$$V = \begin{bmatrix} v_{1,1} & \cdots & v_{1,t} & \cdots & v_{1,T} \\ v_{2,1} & \cdots & v_{2,t} & \cdots & v_{2,T} \\ \vdots & & \vdots & & \vdots \\ v_{N,1} & \cdots & v_{N,t} & \cdots & v_{N,T} \end{bmatrix} \quad (11)$$

$$v^*_{n,t} = \frac{v_{n,t} - \mu_t}{\sigma_t}, \qquad n \in \mathcal{N}, t \in \mathcal{T} \quad (12)$$

$v_{n,t}$ is the value of the t-th feature for the n-th user. μ_t is the mean of the t-th feature, and σ_t is the standard deviation of the t-th feature. Given the normalized voltage dataset V^* and setting K_s as 3, modified Algorithm 1 is adopted to identify the phase labels of each house. The output of each step is depicted in Fig. 5, and the obtained topology presents the clear connections between users, transformer feeders, and phases.

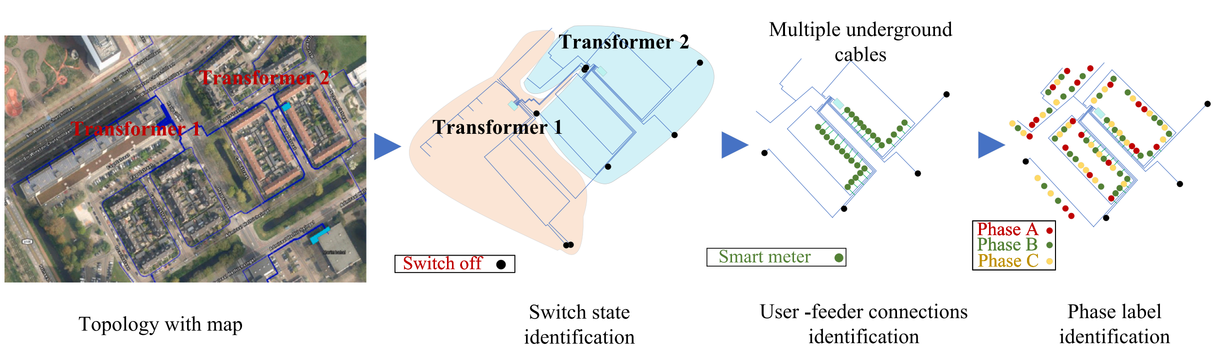

Fig. 5. Illustrative example showing the sequential process and intermediate outputs of the proposed topology correction framework, demonstrating (1) switch states, (2) user-feeder connections and (3) phase labels.

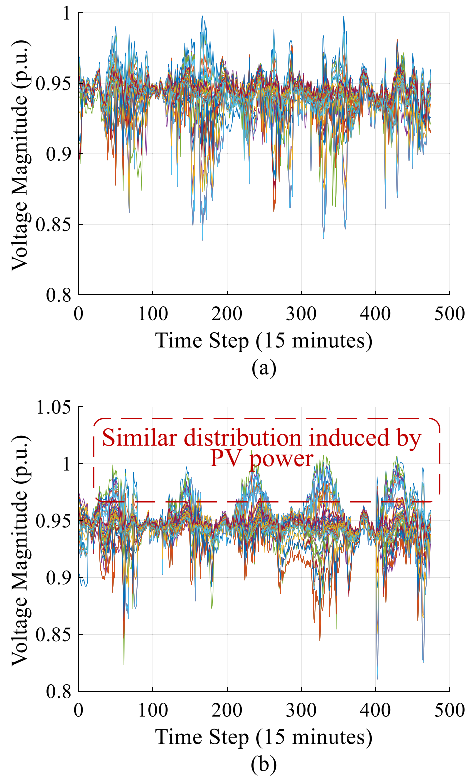

Fig. 6. Distribution of voltage magnitudes in an LVDN:(a) winter datasets and (b) summer datasets.

2.4. Time-based data selection

Compared to traditional LVDNs, the voltage magnitude measurements in modern LVDNs may exhibit higher values due to the integration of PV systems. Within the same distribution network area, environmental factors such as light intensity, temperature, and humidity remain relatively uniform, leading to similar PV generation patterns among users. The voltage magnitudes from the same LVDN are depicted in Fig. 6. This effect is particularly pronounced during summer when high-power PV systems are connected (i.e., the voltage magnitude distribution in Fig. 6(b)). Consequently, users with identical PV capacities but connected to different phases may experience analogous voltage magnitude fluctuations. This phenomenon introduces challenges to the identification steps based on correlation analysis, as the voltage magnitude variations induced by PV generation may obscure the underlying phase-dependent characteristics.

Although various signal processing techniques, such as SA and high-pass filters, can be employed to mitigate high-frequency or low-frequency noise induced by household PV, the determination of appropriate parameters must be tailored to specific datasets. This requirement presents practical challenges in implementing these methods effectively. Inspired by the power-band based strategies in [26], a time-based data selection strategy is introduced to extract voltage measurements during periods when household loads dominate voltage variations, and PV generation remains lower than the local demand. The proposed time-based data selection strategy is easily deployed and fast. Two strategies for selecting relevant data are introduced:

- *Fixed time threshold.* Data is selected from a predetermined time window, such as from 22:00 to 5:00, when PV generation is minimal.

$$\mathcal{T}_d = \{t \mid t_1 \le t \le t_2\}, \tag{13}$$

where t_1 and t_2 define the fixed time window within which the data are extracted.

- *Dynamic time threshold.* Data is selected based on historical PV output values, ensuring that the chosen data corresponds to periods when PV generation is lower than the user load.

$$\mathcal{T}_d = \{t \mid P_{\mathrm{PV}}(t) < P_{\mathrm{load}}(t)\}, \tag{14}$$

where $\mathcal{T}_d$ represents the selected time instances, $P_{\mathrm{PV}}(t)$ is the PV output at time t, and $P_{\mathrm{load}}(t)$ is the user load at time t. Only the data points satisfying this condition are retained for phase identification.

The proposed approach adaptively selects between Eqs. (13) and (14) based on PV data availability. When historical PV generation data are available, Eq. (14) is adopted to refine the selection by identifying intervals with low PV injection. Conversely, in the absence of such data, Eq. (13) serves as an easy and alternative method by utilizing post-sunset voltage measurements, ensuring the model's applicability. By applying these time-based selection strategies, the impact of PV-induced voltage fluctuations can be mitigated, thereby improving the robustness of correlation-based phase identification methods.

For three-phase users, SMs are assumed to provide three independent voltage time-series corresponding to each phase. In the proposed methodology, each phase is modelled as an individual dataset and processed analogously to single-phase users. The algorithm performs independent analysis on the three phase-specific time-series and subsequently assigns an ordered set of phase labels (i.e., final phase labels must remain physically consistent at the user level). Specifically, the three voltage profiles of the same user must be assigned to three distinct phases, including the six feasible phase permutations (A–B–C, A–C–B, B–A–C, B–C–A, C–A–B, C–B–A). This utilization of three-phase datasets enables consistency with the single-phase data processing, avoiding inducing extra calculation to the approach.

3. Case study

To validate the proposed framework, this section conducts comprehensive evaluations across diverse Dutch grid scenarios. Unlike existing methods that often target isolated topology components, our approach provides an integrated solution for correcting switch states, user-feeder connections, and phase labels. Consequently, each step of the algorithm is benchmarked against representative machine learning techniques to demonstrate its feasibility and robustness. Furthermore, the following cases also evaluate the method's practical feasibility for DSO deployment under varying network scales and data distributions.

The code used in this paper is available online[1]. The resolution of the dataset is 15 minutes, resulting in 96 data points per day. The base three-phase voltage level is 0.4 kV, and the loads of three-phase customers are non-uniformly distributed across Phases A, B, and C. The time thresholds t_1 and t_2 are set to 20 and 88. To ensure data privacy, SM datasets provided by the DSO are anonymized before analysis. The original customer identifiers and location information are removed, and the SM data are randomly assigned to end-users within the studied networks. Then, the voltage magnitude profiles are generated using a power flow model, solved by using the Power Grid Model package [38].

3.1. Switch state analysis

The switch bar is connected to three and five transformers, respectively, with all transformers assumed to be connected to the same MV network, which introduces similar low-frequency variations in customer voltage profiles. Each feeder connects to 49 customers. Using the voltage datasets selected through the time-based strategy described in

[1] https://github.com/distributionnetworksTUDelft/Data_Driven_LV_Topology_Correction.

Table 2
Accuracy (%) comparison under multiple scenarios.

Number of Switch	Number of available SM	Methods		
		RF	SVM	GNB
3	2	69.16	63.31	60.39
	10	87.98	88.63	85.71
	20	88.31	88.96	84.74
	All	95.45	96.75	89.61
5	2	56.56	53.59	56.72
	10	91.56	99.38	95.16
	20	92.50	99.21	94.06
	All	98.75	98.59	95.94

Section 2.4, the performance of three models is evaluated under varying numbers of available SMs at each feeder, including RF, Support Vector Machine (SVM), and Gaussian Naive Bayes (GNB).

Table 2 demonstrates that the accuracy of all methods improves as the number of available SM increases. SVM achieves high accuracy (i.e., above 98% for 5 switches scenarios and 88% for 3 switches scenarios) once the number of houses reaches 10, while RF reaches the same level only when all SM data are available. This suggests that SVM is more efficient in learning from limited data, likely due to its ability to find optimal decision boundaries even with fewer samples. For the two SM data available scenarios, the accuracy of RF is about 5% higher than that of SVM in two scenarios (i.e., 3 and 5 switches in the switch bar). Moreover, GNB underperforms compared to RF and SVM, with its accuracy lagging behind when the number of houses is low (lower than 90% and 95% under all the SM data in two cases, respectively).

Given datasets with random missed data points, the accuracy of RF and SVM are assessed, and the average accuracy under five simulations is shown in Fig. 7. In general, the accuracy of the two approaches improves significantly as the number of available SMs increases and slightly decreases as the missing percentage increases, demonstrating the robustness of the proposed method with sufficient data availability. When only 2 SMs data are available, the accuracy remains consistently low (around 55%–60%), indicating the method's limited capability under sparse data conditions. Besides, the sensitivity of SVM accuracy to the available data is less than that of RF, since the accuracy of SVM remains close when the number of available SM is larger than 10. These findings highlight the robustness of SVM and RF for switch state identification tasks in LVDNs.

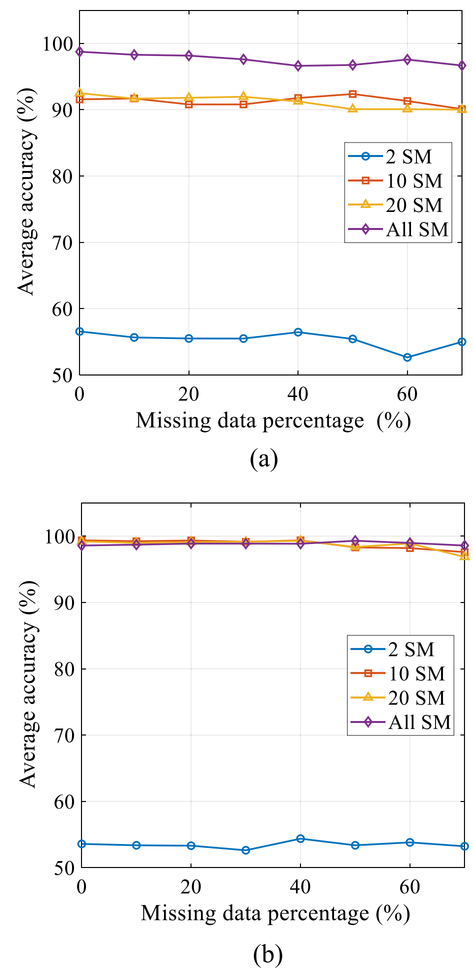

Fig. 7. Switch state identification accuracy under incomplete dataset by: a) RF and b) SVM.

Table 3
Quantitative description of test LVDN topologies.

Network	Feeder number	Nodes per feeder	Structure type
CN1	2	20~26	Complex + Complex
CN2	2	20~26	Simple + Complex
SN	2	20~25	Simple + Simple
SNL	2	20~25	Long Simple + Long Simple
SNB	2	20~25	Branched Simple + Branched Simple
Extended cases	3	20–30	Mixed (Simple/Complex)

3.2. User-feeder identification analysis

The influence of network complexity, phase configuration, and per-node connection on classification accuracy is first analyzed. The per-node connection is the number of households connected to a point, e.g., a multi-unit building. The test networks are used to represent LVDNs with varying structural characteristics in terms of feeder length and branching complexity. Five representative configurations are considered, as summarized in Table 3: CN1, CN2, SN, SNL, and SNB. The topology of CN2 and SNB is shown in Fig. 8.

In this work, feeder structures are classified as simple or complex based on two key topological attributes: feeder length and the number of branches. Specifically, simple-structured feeders are characterised by relatively short radial paths with limited branching, while complex-structured feeders exhibit longer feeder lengths and multiple branching points, resulting in deeper topologies. In Fig. 8, the red diamond markers represent connection nodes, i.e., the electrical connection points between the main feeder and end-users. The SNB represents the SN with a higher number of lateral connections originating from the main cable, i.e., more branch points where users are connected.

Given datasets with SM errors but without available recordings of user-feeder connections, Table 4 presents the accuracy comparison under five modified configurations of LVDNs.

3.2.1. Identification without recordings

The PCC and MFP among customers in CN1 with multiple connections per node are depicted in Fig. 9. As illustrated in Fig. 9(a), the expression in (7) reduces the voltage correlation between users connected to different feeders. This step helps distinguish between different connections more clearly. However, the complexity of the voltage data correlation shown in Fig. 9(b) makes it extremely challenging to simultaneously identify both the phase label and the user-feeder connection using only voltage data. A balanced LVDN represents a scenario where all users are connected to the same phase (i.e., a single-phase LVDN), whereas an unbalanced LVDN corresponds to a three-phase configuration.

Table 4 summarizes the accuracy of user-feeder connection identification across five LVDN scenarios. In general, the proposed approach

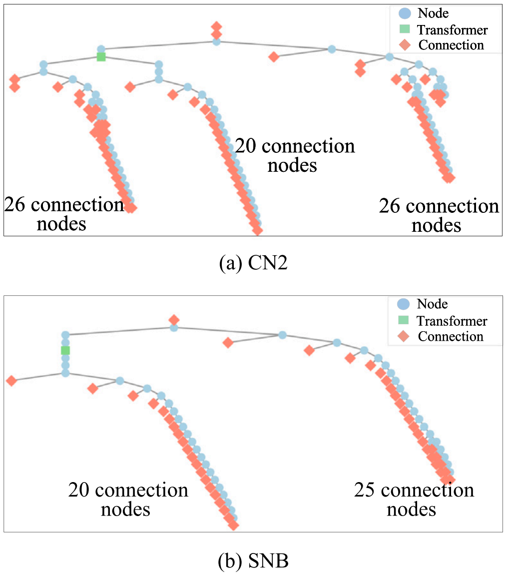

Fig. 8. Topology of CN2 in (a) and SNB in (b).

Table 4
Accuracy (%) comparison under multiple scenarios.

Network Acronyms	Phase	Per-node connection	Two feeders	Three feeders
CN1	1	multiple	100	100
		single	100	96.42
	3	multiple	67.44	55.04
		single	50.00	46.43
CN2	1	multiple	100	98.26
		single	100	100
	3	multiple	42.86	30.43
		single	62.50	35.56
SN	1	multiple	94.56	67.56
		single	75.56	38.46
	3	multiple	55.56	40.54
		single	48.89	24.15
SNL	1	multiple	96.30	97.33
		single	82.22	84.84
	3	multiple	59.26	37.33
		single	46.67	30.30
SNB	1	multiple	95.45	92.50
		single	80.00	88.57
	3	multiple	55.56	43.18
		single	48.49	45.71

demonstrates higher robustness to variations in the number of feeders connected to the same transformer and per-node connections but exhibits lower robustness in three-phase configurations. Specifically, accuracy drops from 100% to 96.43% in CN1 with three feeders and multiple connections per node, and from 94.56% to 55.56% in SN with two feeders and multiple connections per node, highlighting the impact of complex multi-phase settings. In simpler LVDNs, per-node connections significantly affect accuracy. For instance, in SNL with two feeders in a single-phase setting, accuracy declines from 96.3% to 82.22%, while in SNL with two feeders in a three-phase setting, it drops from 59.26% to 46.67%. In contrast, for complex LVDNs, accuracy remains above 95% regardless of single or multiple per-node connections, indicating greater resilience in more intricate network structures. Moreover, accuracy in complex LVDNs is slightly higher than in simpler LVDNs, primarily due to feeder length differences. Simple LVDNs, which are modified from urban areas with high population density, have houses located close to each other (i.e., smaller distances between users

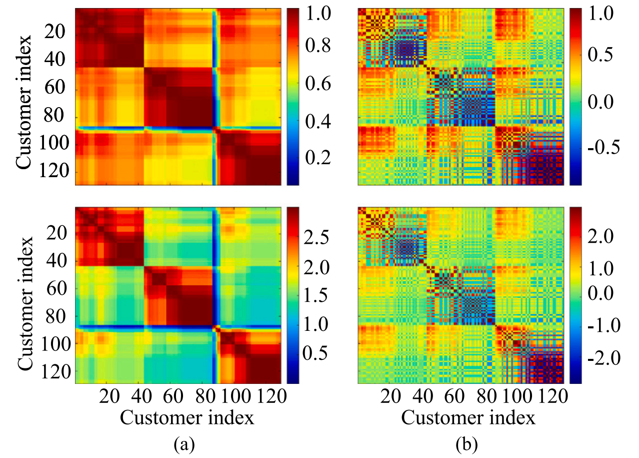

Fig. 9. PCC in the first row and FPCC in the second row among voltage datasets: a) with single-phase customers and b) with three-phase customers.

with different connections). This proximity reduces the difference in the voltage variations between neighbouring houses, leading to greater identification accuracy. However, the length of the feeder and the number of branches have a minor effect on the accuracy, which is verified in the three simple LVDNs (i.e., SN, SNL, and SNB).

It is observed that the identification accuracy of the proposed method declines when applied to three-phase networks featuring three or more outgoing cables. This performance degradation highlights an inherent limitation of relying solely on voltage profiles in scenarios with high branch density and complex coupling. To overcome this shortcoming, incorporating multi-source datasets represents a promising avenue for future enhancement. Leveraging such auxiliary information can expand the feature space, thereby significantly improving the discriminative power and spatial resolution of the algorithm in more sophisticated grid structures.

Since unbalanced LVDNs are more representative of real-world conditions compared to balanced LVDNs, only half of the identified user-feeder connections with high confidence are selected for further processing in the subsequent section. These experimental results highlight the inherent challenge of simultaneously identifying user-feeder connections and accurately determining the phase labels of individual users. The complexity arises from the intertwined nature of voltage variations across different feeders and phases, further complicated by factors such as DERs and seasonal load fluctuations.

3.2.2. Identification with recordings

Given datasets with varying levels of unknown user-feeder connections, Table 5 compares the average accuracy of 20 simulations at each scenario (i.e., each unknown level). The results provide insights into the robustness and effectiveness of both methods across different network configurations. In general, as the percentage of unknown user-feeder connections increases from 10% to 50%, a gradual decline in accuracy is observed across all methods and network configurations. However, both the proposed MFP-based method and KNN (with varying k values) maintain relatively high performance, with accuracy exceeding 91% even at 30% unknown connections.

Table 5 shows that the traditional Euclidean distance-based KNN consistently outperforms the proposed modified KNN when the neighborhood size is small (e.g., $k < 8$), particularly under low levels of missing data. This behavior is expected, as a smaller k relies on highly local information, where the standard distance metric can effectively capture similarity when sufficiently accurate observations are available. However, this advantage diminishes as the proportion of missing data increases and the neighborhood size becomes larger. Under

Table 5
Average accuracy under three-phase datasets and multiple connections at nodes with SM errors.

Network Acronyms	Unknown rate	Modified KNN $k = 1$	$k = 3$	$k = 5$	$k = 8$	Traditional KNN $k = 1$	$k = 3$	$k = 5$	$k = 8$
CN1	10%	98.16%	98.10%	99.02%	99.59%	99.90%	99.79%	99.84%	99.95%
	20%	96.16%	96.49%	97.44%	98.94%	99.84%	99.67%	99.68%	99.51%
	30%	93.73%	95.19%	96.73%	98.65%	99.52%	99.48%	99.08%	98.13%
	40%	91.71%	92.79%	96.13%	97.84%	99.32%	98.75%	97.92%	94.70%
	50%	88.84%	90.63%	95.59%	96.92%	99.03%	98.33%	94.76%	88.51%
SN	10%	97.63%	98.06%	98.41%	98.57%	99.65%	99.65%	99.20%	98.31%
	20%	95.07%	95.80%	97.04%	97.11%	99.24%	99.30%	98.24%	96.98%
	30%	92.19%	93.80%	95.20%	94.72%	98.98%	98.20%	97.00%	92.22%
	40%	90.11%	91.37%	92.50%	92.44%	98.56%	97.48%	93.54%	88.94%
	50%	87.11%	88.70%	89.43%	87.43%	97.72%	92.98%	88.39%	83.50%
SNB	10%	97.43%	98.09%	98.33%	98.43%	99.61%	99.81%	99.48%	98.65%
	20%	94.37%	95.87%	96.56%	96.83%	99.39%	99.31%	98.63%	96.52%
	30%	91.80%	93.19%	94.78%	94.83%	98.81%	98.43%	96.69%	92.54%
	40%	90.35%	91.70%	93.02%	91.87%	98.43%	96.78%	93.57%	89.04%
	50%	87.50%	88.46%	89.31%	87.22%	96.76%	94.15%	88.57%	82.11%

such conditions, the modified KNN demonstrates improved accuracy, particularly at higher unknown rates (e.g., 50%) and larger k, where it achieves accuracy levels exceeding 87%. This indicates that the proposed distance formulation can be more effective in mitigating the distortion introduced by incomplete measurements, thereby helping preserve identification performance when reliable neighbor information becomes sparse.

It is important to note that the selection of k is inherently data-dependent in practical DSO applications, and the number of available labeled users is often not known a priori. In data-sparse scenarios, where only a limited number of neighboring users have reliable phase labels, a small k (e.g., $k = 1$–3) is typically preferred and the traditional KNN may be more suitable. In contrast, in settings with a larger pool of available labeled users, DSOs may rely on a larger neighborhood to improve statistical reliability. In such cases, the proposed modified KNN can provide advantages in terms of robustness and stability. This trend is further confirmed in the SNB network, where increased branching introduces additional structural complexity. While the traditional KNN shows an advantage at low or intermediate unknown rates and small k, the modified KNN achieves comparable or superior performance as missing data increases or when larger neighborhoods are considered. Overall, these results do not suggest that the proposed method universally outperforms the traditional KNN. Rather, they indicate that the two methods may be preferable under different data conditions. In practice, DSOs can select the more appropriate method based on the estimated availability of labeled data, the expected neighborhood size, and other operational information, with Table 5 providing a useful quantitative reference for such selection.

To validate the robustness of the proposed method with respect to network topology, missing data, and measurement noise, a comprehensive set of experiments is conducted under varying scenarios. Specifically, performance is evaluated across different levels of incompleteness of the data, varying proportions of available labeled data, and different noise magnitudes. The missing data ratio varied from 10% to 90%. Regarding SM noise, four classes of SMs are considered according to IEC 62053-21, corresponding to error levels of 0.2%, 0.5%, 1%, and 2% (extended to 5% for extreme scenarios). Besides, to compare the accuracy between the traditional KNN and the improved KNN under different missing data conditions, an *accuracy gap* (Δ accuracy) is introduced, defined as the difference between the average accuracy of the modified KNN algorithm and that of the traditional KNN algorithm. In this analysis, the number of neighbours is fixed at $k = 3$. The corresponding results are illustrated in Fig. 10.

From the line plots in Fig. 10, it can be observed that the accuracy of the proposed method remains relatively stable across different levels

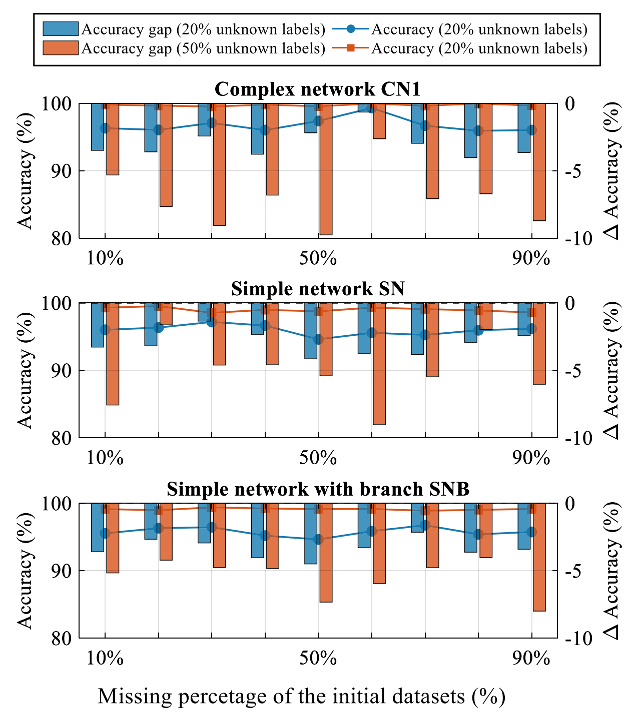

Fig. 10. Mean accuracy and corresponding performance gap under varying levels of data incompleteness ($k = 3$).

of missing data, indicating strong robustness against data incompleteness. Furthermore, the bar charts show that the accuracy gap is generally negative under the considered noise levels when $k = 3$, suggesting that the traditional KNN slightly outperforms the improved method in these scenarios. Nevertheless, the performance difference remains within 10%, which is consistent with the observations reported in Table 4.

Fig. 11 presents the comparison under different noise conditions. As the noise level increases, the accuracy of the traditional KNN gradually decreases, indicating its sensitivity to measurement noise due to its reliance on Euclidean distance. In contrast, the proposed method exhibits a different trend. The accuracy initially slightly increases and then decreases slightly as noise grows, while maintaining comparable overall performance. This is because the proposed method is based on correlation analysis rather than Euclidean distance, making it inherently more robust to noise. Moderate magnitude noise can enhance the

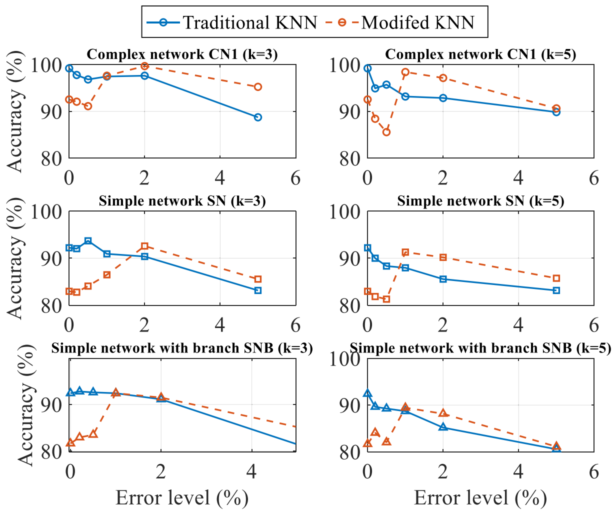

Fig. 11. Mean accuracy across multiple datasets with varying noise levels.

distinguishability between different users due to heterogeneous noise, leading to a temporary improvement in performance. However, as the noise magnitude becomes large, the similarity induced by high-amplitude Gaussian noise impacts correlation coefficient calculation, leading to accuracy reduction.

Based on Table 4 and Figs. 10–11, it can be concluded that the two methods exhibit complementary characteristics under different conditions. For scenarios involving high noise levels or larger k values, the improved KNN tends to achieve better performance. Conversely, under low-noise conditions and smaller k, the traditional KNN can serve as a competitive alternative. Therefore, an adaptive selection between the two methods can be employed to ensure reliable accuracy across diverse operating conditions.

3.3. Phase identification analysis

The influence of feeder complexity, environmental conditions, and load distribution on identification accuracy is analyzed, with results summarized in Tables 6 and 7. CC_PV: represents a complex-structured feeder from CN1, with photovoltaics (PV) installed at each house. SC_PV: corresponds to a simple-structured feeder from SN, also with PV at each house. CC and SC: represent feeders from CN1 and SN with PV, but with datasets preprocessed using the method described in Section 2.4. These four feeders serve single-phase users only. CC_3_NU and SC_3_NU: denote complex and simple feeders equipped with three-phase meters, where the load is non-uniformly distributed among the three phases.

Table 6
Average purity (%) under datasets from feeders with different components.

Network	Setting	Winter Dataset		Summer Dataset	
		single	multiple	single	multiple
CC_PV	–	100	100	78.57	65.12
	+20	100	100	77.08	77.78
SC_PV	–	96.00	97.06	84.00	88.24
	longer feeder	100	94.12	84.00	88.24
CC	–	100	100	100	97.67
	+20	100	100	100	100
SC	–	88.88	88.24	68	82.35
	longer feeder	100	100	68.00	82.35
CC_3_NU	PV	100	100	79.41	83.67
	filtered	100	100	100	100
SC_3_NU	PV	97.14	97.92	71.43	100
	filtered	97.92	94.29	100	95.83

Table 7
Average accuracy (%) under datasets from feeders with different components.

feeder	setting	Winter Dataset		Summer Dataset	
		single	multiple	single	multiple
CC_PV	–	100	100	88.89	72.08
	+20	100	100	76.11	87.88
SC_PV	–	96.97	97.78	87.88	90.30
	longer feeder	100	94.75	87.88	90.30
CC	–	100	100	100	97.78
	+20	100	100	100	100
SC	–	87.45	87.88	71.86	79.22
	longer feeder	100	100	71.86	79.22
CC_3_NU	PV	100	100	86.11	85.19
	filtered	100	100	100	100
SC_3_NU	PV	97.78	97.62	76.97	100
	filtered	97.62	95.56	100	95.24

"+20" indicates an additional 20 customers connected to the feeder. "Long feeder" refers to an extended feeder length compared to the original configuration. Accuracy and clustering purity are used to evaluate the performance of phase identification, as formulated in the expression (15).

$$P_{pu} = \frac{1}{N} \sum_{\mathcal{L}_k \in \mathcal{L}} \max_i |\mathcal{L}_k \cap \mathcal{L}_i| \tag{15}$$

where $\mathcal{L}$ denotes the set of obtained labels for all input data, and $\mathcal{L}_k$ represents the label assigned to the data in cluster C_k. $\mathcal{L}_i$ is the label set of data points that truly belong to class C_i.

Overall, Table 6 indicates consistently higher accuracy in winter datasets compared to summer datasets. This discrepancy is primarily due to voltage variations caused by household photovoltaic (PV) systems, which negatively impact phase identification performance. In the first four cases, where only single-phase or balanced three-phase customers are considered, the influence of PV-induced voltage fluctuations is evident. For instance, in the summer dataset, the CC_PV feeder (a complex feeder from CN1 without PV filtering) achieves 88.89% accuracy with a single connection per node but drops to 72.08% with multiple connections per node. However, when PV-induced variations are filtered out—as in the CC and SC cases—accuracy consistently exceeds 95%, highlighting the critical role of PV filtering in improving phase identification.

These findings demonstrate that phase identification accuracy is highly dependent on seasonal variations and dataset complexity, particularly in the presence of PV-induced voltage fluctuations. Without applying the proposed time-based selection process to mitigate PV impacts, accuracy remains below 90% in most cases. In contrast, once PV-induced variations are filtered, accuracy consistently exceeds 95%, regardless of network complexity or three-phase load distribution. Additionally, factors such as the number of connections per node, feeder length, and the number of connected houses have a relatively minor impact on classification accuracy compared to PV output power fluctuations. This underscores the importance of addressing PV-induced voltage variations to achieve robust and reliable phase identification.

3.4. Limitation analysis

Based on the above discussion and simulation result analysis, it is verified that the proposed method offers an effective solution for DSOs to maintain accurate topology within Digital Twin frameworks via non-intrusive voltage datasets. However, its broad application faces three potential challenges:

- Network topology complexity: The identification performance is increasingly challenged by sophisticated distribution structures, particularly in networks featuring more than three cables (e.g., high branch density), which complicates the correlation patterns.

- Data quality heterogeneity: according to the robustness analysis in Section 3.2, inconsistent SMs precision poses difficulties in tuning the hyperparameter k.
- Data granularity: if users only share the maximum/minimum voltage value [24], the application of the developed approach will be significantly limited.
- Evolving load dynamics: The proliferation of heat pumps may change the distribution of winter data, potentially introducing new patterns of similarity that challenge the feasibility of the proposed approach.
- Long feeders and missing voltage measurements: In long feeders, voltage differences may become less evident, and nearby houses on different phases may show similar voltage fluctuations, which may reduce identification accuracy. If true voltage measurements are unavailable, the affected data can either be discarded or replaced using state estimation.
- Incorrect prior labels: The proposed method assumes that the available labels are correct so that they can effectively support the identification process. Incorrect prior labels may reduce the identification accuracy.

In future work, these limitations will be further addressed. Meanwhile, the ongoing transition toward advanced SMs with enhanced sensing and communication capabilities provides significant opportunities to further improve the accuracy and practical applicability of the proposed approach in increasingly digitised power grids. Another relevant direction for future work is to explore whether the proposed framework can be extended to settings where only power measurements are available.

4. Conclusion

In this paper, we propose a practical data-driven topology correction framework for low-voltage distribution networks, incorporating a time-based smart meter data selection strategy to mitigate the impact of photovoltaic (PV) systems and a balanced dataset construction method to improve the effectiveness of supervised training. Rather than aiming at full topology reconstruction, the proposed framework focuses on the correction of outdated and incomplete topology records using only voltage magnitude measurements within a unified processing framework. The experimental results indicate that the method can achieve robust performance under incomplete data conditions and maintain relatively high accuracy across different levels of random data loss, data dimensionality, and topological complexity. In particular, the correlation-based user–feeder and phase identification components show stronger resilience to data loss, whereas the supervised switch-state identification component is more sensitive to missing and asynchronous measurements. The results also suggest that seasonal variations and PV-induced voltage fluctuations influence the framework's performance, which further underlines the importance of time-based data selection for reliable topology correction. Although the proposed method remains dependent on the quality and representativeness of smart meter measurements, it provides a feasible means of improving network observability and supporting distribution system operators in applications such as load balancing and PV integration. Future work will consider the influence of a broader range of flexible assets and more diverse operating conditions, while also investigating the extension of the framework to settings in which only power measurements are available.

CRediT authorship contribution statement

Dong Liu: Writing – original draft, Software, Methodology, Conceptualization. **Sander Timmerman:** Validation, Software. **Yu Xiang:** Writing – review & editing. **Ensieh Hosseini:** Validation, Software. **Peter Palensky:** Funding acquisition. **Pedro P. Vergara:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] S. Dahale, A. Pahwa, B. Natarajan, Phase identification in unobservable distribution systems, IEEE Trans. on Power Del. 38 (5) (2023) 3067–3075.
[2] H. Peng, K. Liao, J. Yang, B. Pang, Z. He, Deep reinforcement learning based multi-timescale volt/var control in distribution networks considering network reconfiguration, IEEE Trans. Sustain. Energy 16 (4) (2025) 2948–2958.
[3] Y. Cao, B. Zhou, C.Y. Chung, K. Zhou, L. Zhu, Z. Shuai, Resilience-oriented coordinated topology reconfiguration of electricity and drainage networks with distributed mobile emergency resources, IEEE Trans. on Smart Grid 16 (1) (2025) 786–800.
[4] M. Vanin, T. Van Acker, R. D'hulst, D. Van Hertem, Phase identification of distribution system users through a MILP extension of state estimation, Electric Power Syst. Res. 229 (2024) 110107.
[5] P.P. Vergara, J.S. Giraldo, M. Salazar, N.K. Panda, P.H. Nguyen, A mixed-integer linear programming model for defining customer export limit in PV-rich low-voltage distribution networks, J. Mod. Power Syst. Clean Energy 11 (1) (2023) 191–200.
[6] J. Then, A.P. Agalgaonkar, K.M. Muttaqi, Hosting capacity of an Australian low-voltage distribution network for electric vehicle adoption, IEEE Trans. Ind. Appl. 60 (2) (2024) 2601–2610.
[7] R. Zhang, Z. Liu, Z. Bie, B. Hredzak, Dynamic clusters supported bi-level control system with lower communication requirements for a distribution network with distributed battery and photovoltaic systems, IEEE Trans. on Smart Grid 15 (4) (2024) 3824–3838.
[8] H. Zhang, J. Zhao, X. Wang, Y. Xuan, Low-voltage distribution grid topology identification with latent tree model, IEEE Trans. Smart Grid 13 (3) (2022) 2158–2169.
[9] L.F. Costa, J.S. Giraldo, C.A. Castro, Identification and correction of transmission line parameter errors using SCADA and synchrophasor measurements, Int. J. Elect. Power Energy Syst. 135 (2022) 107509.
[10] C.S.C. Dande, N.A. Efkarpidis, M. Christen, M. Ginocchi, A. Monti, Consumer phase identification in distribution grids using graph neural networks based on synthetic and measured power profiles, Energy AI (2025) 100607.
[11] A.B. Birchfield, T. Xu, K.M. Gegner, K.S. Shetye, T.J. Overbye, Grid structural characteristics as validation criteria for synthetic networks, IEEE Trans. on Power Syst. 32 (4) (2017) 3258–3265.
[12] A. Koirala, L. Suárez-Ramón, B. Mohamed, P. Arboleya, Non-synthetic European low voltage test system, Int. J. Elect. Power Energy Syst. 118 (2020) 105712.
[13] B. Azimian, R.S. Biswas, S. Moshtagh, A. Pal, L. Tong, G. Dasarathy, State and topology estimation for unobservable distribution systems using deep neural networks, IEEE Trans. Instrum. Meas. 71 (2022) 1–14.
[14] Y. Liao, Y. Weng, G. Liu, R. Rajagopal, Urban MV and LV distribution grid topology estimation via group lasso, IEEE Trans. Power Syst. 34 (1) (2019) 12–27.
[15] Y. Liu, J. Wang, P. Wang, Hybrid data-driven method for distribution network topology and line parameters joint estimation under small data sets, Int. J. Elect. Power Energy Syst. 145 (2023) 108685.
[16] P. Shah, X. Zhao, Network identification using μ-PMU and smart meter measurements, IEEE Trans Industr Inform 18 (11) (2022) 7572–7586.
[17] A.B. Pengwah, Y.Z. Gerdroodbari, R. Razzaghi, L.L.H. Andrew, Topology identification of distribution networks with partial smart meter coverage, IEEE Trans. on Power Del. 39 (2) (2024) 992–1001.
[18] S. García, M. Fresia, J. Mora-Merchán, A. Carrasco, E. Personal, C. Leon, A data-driven topology identification method for low-voltage distribution networks based on the wavelet transform, Electr. Power Syst. Res. 243 (2025) 111517.
[19] S. Liu, X. Cui, Z. Lin, Z. Lian, Z. Lin, F. Wen, Y. Ding, Q. Wang, L. Yang, R. Jin, H. Qiu, Practical method for mitigating three-phase unbalance based on data-driven user phase identification, IEEE Trans. on Power Syst. 35 (2) (2020) 1653–1656.
[20] A. Heidari-Akhijahani, A. Safdarian, F. Aminifar, Phase identification of single-phase customers and PV panels via smart meter data, IEEE Trans. on Smart Grid 12 (5) (2021) 4543–4552.
[21] V.A. Jimenez, A. Will, S. Rodriguez, Phase identification and substation detection using data analysis on limited electricity consumption measurements, Electric Power Syst. Res. 187 (2020) 106450.
[22] Liggingsdata kabels en leidingen, https://www.stedin.net/zakelijk/open-data/liggingsdata-kabels-en-leidingen., Accessed: Aug. 2023.
[23] V.A. Jimenez, A. Will, A new data-driven method based on niching genetic algorithms for phase and substation identification, Electric Power Syst. Res. 199 (2021) 107434.
[24] D. Liu, J.S. Giraldo, P. Palensky, P.P. Vergara, Topology identification and parameters estimation of LV distribution networks using open GIS data, Int. J. Elect. Power Energy Syst. 164 (2025) 110395.
[25] Z.S. Hosseini, A. Khodaei, A. Paaso, Machine learning-enabled distribution network phase identification, IEEE Trans. on Power Syst. 36 (2) (2020) 842–850.

[26] H.P. Lee, P. Rehm, M. Makdad, E. Miller, N. Lu, A novel power-band based data segmentation method for enhancing meter phase and transformer-meter pairing identification, IEEE Trans. on Power Del. 39 (4) (2024) 2327–2339.

[27] S. de Jongh, F. Mueller, F. Osterberg, C.A. Cañizares, T. Leibfried, K. Bhattacharya, Data-driven topology and parameter identification in distribution systems with limited measurements, IEEE Trans. Power Del. 40 (1) (2025) 249–260.

[28] M. Xu, R. Li, F. Li, Phase identification with incomplete data, IEEE Trans. on Smart Grid 9 (4) (2018) 2777–2785.

[29] Z. Wang, W. Jiang, A. Zhou, M. Xu, Parameter identification of power distribution system based on dynamic graph topology, IEEE Trans. Ind. Inform. 21 (2) (2025) 1309–1318.

[30] N. Feng, Y. Du, Y. Ding, A heuristic-search-based topology identification and parameter estimation method in low voltage distribution grids with low observability, IEEE Trans. on Smart Grid 15 (6) (2024) 5826–5839.

[31] H. Li, Y. Weng, V. Vittal, E. Blasch, Distribution grid topology and parameter estimation using deep-shallow neural network with physical consistency, IEEE Trans. on Smart Grid 15 (1) (2024) 655–666.

[32] W. Luan, J. Peng, M. Maras, J. Lo, B. Harapnuk, Smart meter data analytics for distribution network connectivity verification, IEEE Trans. on Smart Grid 6 (4) (2015) 1964–1971.

[33] X. Wang, Y. Zhao, Y. Zhou, A data-driven topology and parameter joint estimation method in Non-PMU distribution networks, IEEE Trans. Power Syst. 39 (1) (2024) 1681–1692.

[34] L. Guo, G. Yang, Y. Liu, Y. Zhang, Z. Wang, H. Yu, Y. Wang, C. Wang, Topology and parameter identification of three-phase low-voltage distribution network and voltage control strategy based on user smart meter measurements: a data-driven approach, IEEE Trans. Instrum. Meas. 74 (2025) 1–13.

[35] V.C. Cunha, W. Freitas, F.C. Trindade, S. Santoso, Automated determination of topology and line parameters in low voltage systems using smart meters measurements, IEEE Trans. Smart Grid 11 (6) (2020) 5028–5038.

[36] ABB, ABB offerings, 2025, https://new.abb.com/offerings., Accessed: Apr. 16, 2025.

[37] S. García, J.M. Mora-Merchán, D.F. Larios, E. Personal, A. Parejo, C. León, Phase topology identification in low-voltage distribution networks: a Bayesian approach, Int. J. Elect. Power Energy Syst. 144 (2023) 108525.

[38] Alliander, Power grid model: an open-source power grid calculation package, 2024. https://power-grid-model.readthedocs.io/en/stable/., Accessed: Mar. 25, 2025.