A siamese neural network model for phase identification in distribution networks

Liu, Dong; Giraldo, Juan S.; Palensky, Peter; Vergara, Pedro P.

DOI
10.1016/j.ijepes.2025.110718

Publication date
2025

Document Version
Final published version

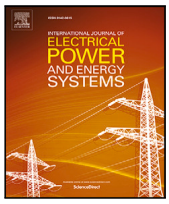
Published in
International Journal of Electrical Power and Energy Systems

Electrical Power and Energy Systems 169 (2025) 110718

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

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



A siamese neural network model for phase identification in distribution networks

Dong Liu [a], Juan S. Giraldo [b], Peter Palensky [a], Pedro P. Vergara [a,*]

[a] Intelligent Electrical Power Grids (IEPG) Group, Delft University of Technology, 2628CD, The Netherlands
[b] Energy Transition Studies Group, Netherlands Organisation for Applied Scientific Research, 2595 DA, The Netherlands

ARTICLE INFO

Keywords:
Distribution systems
Deep learning
Machine learning
Feature extraction
Smart meter data

ABSTRACT

Distribution system operators (DSOs) often lack high-quality data on low-voltage distribution networks (LVDNs), including the topology and the phase connection of residential customers. The phase connection is essential for phase balancing assessment and distributed energy resources (DERs) integration. The existing load profiles-based approaches rely on stepwise subtraction of the identified customers in a step-by-step identification procedure, while the accuracy of each step is not guaranteed. This paper introduces a siamese neural network model to identify single-phase connections without requiring stepwise subtraction. It comprises self-taught learning (STT) and a phase-label identification strategy. The introduced self-taught learning enables DSOs to train a recurrent neural network-based Siamese network (RSN) only relying on an unlabelled dataset. Besides, the siamese network (SN) is robust to noise and fluctuations in the data to a certain extent, making the proposed method robust to measurement errors. A Kendall correlation-based phase modification strategy is introduced to modified phase labels with lower confidence, aiming to mitigate the accuracy loss induced by the limited generalization of SN. The proposed approach is tested on the IEEE European low voltage test feeder and a residential network in the Netherlands Simulation results illustrate the feasibility and robustness of the proposed approach on incomplete datasets. The accuracy exceeded 83% and 90%, respectively, when using datasets of less than 20 days with and without measurement errors.

1. Introduction

Phase connection of customers in distribution networks (DNs) is crucial for distribution system operators (DSOs) to perform active management, e.g., load balancing, congestion management and distributed energy resources (DERs) integration [1–3]. However, this information might be incomplete due to the missed and uninformed phase switching. Compared to three-phase customers, the large amount of single-phase customers in LVDNs, especially in European LVDNs [4], imposes pressure on the timely updating of phase connectivity. Moreover, the uncertainty of DERs impacts the variation of load and the correlations between measurements, challenging the phase identification [5,6]. Approaches relying on phasor measurement units (PMU) might be infeasible since there is a lack of PMU in LVDNs [7]. Although smart meters (SM) are gradually installed in LVDNs, privacy issues and communication errors make it hard to obtain a complete dataset [8]. Thus, flexible phase identification approaches for single-phase customers are needed.

Traditional approaches, such as manual phase identification, are rarely used due to their high cost and low efficiency [9]. Instead, state-of-the-art research focuses on data-driven approaches, which rely on increasing amounts of time-series SM data and machine learning (ML) methods. According to the utilized data, the SM data-based approach could be divided into voltage-based and active power-based approaches. Voltage magnitude of the customers that are connected to the same phase shows similar variations when the load changes, indicating a higher correlation among voltage magnitude within the same phase compared to different phases [10,11]. However, voltage correlations on the same phase are also impacted by other factors, such as customer electrical distance. For instance, the voltage correlation between customers located far apart may be lower than the voltage correlation between customers in closer proximity to each other but connected to different phases. Meanwhile, adequate time-series voltage datasets are not commonly available in LVDNs, which are required by most of the approaches based on voltage [12]. Conversely, time-series load profiles are normally recorded and stored for billing [13].

* Corresponding author.
E-mail address: P.P.VergaraBarrios@tudelft.nl (P.P. Vergara).

https://doi.org/10.1016/j.ijepes.2025.110718
Received 5 August 2024; Received in revised form 24 October 2024; Accepted 30 April 2025
Available online 19 May 2025

Compared to voltage profiles, there is no strong correlation among load profiles in the same phase since the electricity consumption patterns vary across households. However, there exists a correlation between the customer and transformer data. Similarities of paired data samples (i.e., customer data and the corresponding phase data) should be larger than those of unpaired data samples. Thus, phase identification using load profiles could be taken as a time-series data pairing problem, illustrated in Fig. 1. The general process of data pairing is summarized into four steps: (1) calculate the correlations between households and the three phases. (2) if the highest correlation coefficient of a household is higher than the pre-set threshold, the phase label of the household is set as the corresponding phase. (3) The dataset of households with assigned phase labels is removed. (4) Repeat steps 2 to 3 until all the households are assigned one label. To reveal the similarities between time-series data, feature extraction can be integrated into the pairing strategies. Wavelet decomposition (WD) was adopted to extract expressive features to enhance the identification accuracy of power cable faults in [14]. Saliency analysis (SA) was integrated to pre-process the raw time-series data, and correlation analysis was used to determine the phase label [15]. Based on SA, the approach in [16] integrated statistical tests to guarantee the identification accuracy of the data pairs with weak correlations. A clustering approach with a high-pass filter was introduced in [13] to identify the phase connections under random and consecutive incompleteness datasets. The above approaches are similar to hard classification, i.e., labelling the customers as 0/1. Besides, phase identification could be taken as a soft classification problem. A Bayesian-based fuzzy phase identification method was proposed to assign three probabilities to label customers, with the sum of probabilities equalling one [4]. The common step in the above approaches is the stepwise subtraction of the identified customers, which enhances the correlations between transformers and the customers who are located far away from the substation. However, the accuracy of the current step is subject to the accuracy of the previous step, leading to error accumulation. To address this challenge, a data-driven approach in [17] was constructed based on a genetic algorithm and correlation analysis to identify the customer phase under incomplete period datasets. Nevertheless, most of the above approaches require a complete time-series dataset, and some are sensitive to missing data points and measurement errors. [18].

Siamese network (SN) is a common structure network in meta-learning and it is commonly used in similarity analysis of images [19] and visual objection tracking [20]. The main advantage of SNs is that their training does not require a large number of samples from the same category, which alleviates the workload of collecting labelled training datasets. Thus, SN is a promising solution for analysing the similarity between the transformer and customer datasets. For phase identification, there are only limited or no labelled customer datasets for training an SN. To address this issue, self-supervised learning (SSL) [21] and self-taught (ST) process [22] are potential solutions, originally proposed to learn transferable and representative knowledge from unlabelled data or the generated pseudo data. ST aims to complete the tasks only relying on given datasets by the combination of unsupervised and supervised learning schemes [22,23]. A deep SN was constructed to identify the electricity theft behaviour in [24] and a robust classification approach was proposed based on recurrent neural networks (RNN) to identify the grid disturbances [25]. To locate the fault in DNs without requiring real-word labelled datasets, an SN was designed based on Transformers Neural Network (TNN) [26]. Based on a modified TNN, an SSL-based load forecasting approach was proposed to predict the DERs power [21]. In [27], a feature extraction strategy was integrated into the SSL to assess the reliability of the power systems. Although SN and ST are used to analyse time-series measurements in power systems, they have not been used to address phase identification. Moreover, there is no ST technique for SN training, hindering the application of SN for the phase identification problem. It is important to highlight that the similarity between transformer and customer data is impacted

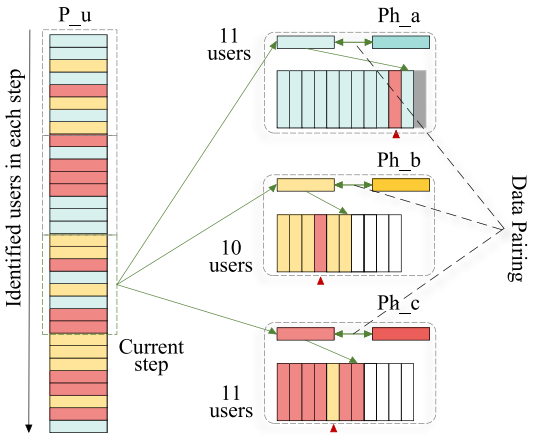


Fig. 1. Phase identification process: blue, orange, and red blocks represent customer data in phases a, b, and c, respectively. Darker blocks represent transformer data. Blocks with a red triangle indicate wrongly identified customers.

by multiple factors, including measurement errors, line loss, and the number of customers connected to the same phase, which makes similarities analysis in phase identification more complex and difficult than that of images. Table 1 summarizes the approaches discussed in the aforementioned papers and the proposed approach.

This paper proposes a siamese neural network model to identify the phase of single-phase customers in LVDNs, which is not subjected to stepwise subtraction while showing strong robustness to measurement errors. The proposed approach consists of two stages: self-taught training (STT) and phase label identification. The first stage is composed of pseudo-data generation, feature extraction and recurrent neural network-based SN (RSN) training procedures, deployed to prepare the input dataset using unlabelled SM data for training RSN. In the second stage, a sliding window strategy is adopted to calculate the probability phase labels by aggregating the output of the trained RSN in each window. Finally, a Kendall correlation-based phase modification strategy is proposed to determine the final phase labels. The main contributions of this paper are summarized as follows:

- An RSN is constructed to analyse the similarity between high-dimension transformer and customer load profiles, which is used to calculate the probability of phase label by each customer. The obtained similarity score and phase identification accuracy are not subject to the stepwise subtraction of the identified customers.
- To train the constructed RSN without requiring labelled datasets, a self-taught training strategy is proposed. WD was first uniquely adopted to extract features for phase identification. By leveraging pseudo data, the STT strategy enables the trained RSN to effectively calculate similarity in time-series data, even in the absence of labels, ensuring accurate phase identification.
- A Kendall correlation coefficient-based phase label modification is introduced to determine the final phase label by modifying the probability phase labels obtained through the trained RSN. This step aims to mitigate the accuracy loss induced by the inherent limitation of neural networks (e.g., limited generalization) by revising labels with less confidence.

The remainder of this paper is organized as follows: Section 2 illustrates the framework of phase identification, the process of self-taught and the phase identification based on the trained RSN. Section 3 describes the case of studies and results. Section 4 presents the conclusions.

Nomenclature

Acronyms

LVDNs	Low-voltage distribution networks
DNs	Distribution networks
DSOs	Distribution system operators
DERs	Distributed energy resources
STT	Self-taught training
SN	Siamese Network
RNN	Recurrent neural network
RSN	Recurrent neural network based SN
SM	Smart meter
PMU	Phasor measurement units
SA	Saliency analysis
WD	Wavelet decomposition
ST	Self-taught
SSL	Self-supervised Learning
TNN	Transformers Neural Network

Index/Set

$n/\mathcal{N}$	Index/set of customers in the networks
$t/\mathcal{T}$	Index/set of time step
$\psi/\mathcal{F}$	Index/Set of the phases
$m/\mathcal{M}$	Index/set of the windows
$\mathcal{I}$	Set of customers whose labels with lower confidence

Parameters

P_{uu}/P_p	Data from N customers/transformer
$\dot{P}_{uu}/\dot{P}_p$	Reconstructed data from customers/transformer
$P^*_{u,n}/P^*_{\psi}$	Split data from each customer/each phase
P^*_{uu}/P^*_p	Split data from all customers/transformer
P/Y	Input dataset/label for proposed approach
$y^m_{\psi,n}$	Label of the sample in the mth window that consists of the pseudo transformer data P^m_{ψ} and the nth customer data $P^m_{u,n}$
P_{trian}	Training dataset for RSN
$P^m_{\psi}/P^m_{u,n}$	The mth row in matrix $P^*_{\psi}/P^*_{u,n}$
S^+	Dataset consists of samples with label 1
E	Error Matrix
γ	Pre-set margin for calculating loss
ϵ	Threshold for $\Delta\tau^*_n$
$\Delta P_{\psi}/\Delta\widetilde{P}_{\psi}$	Variation of real/identified phase data
$p^t_{\psi}/\widetilde{p}^t_{\psi}$	Real/identified phase data at time t
N/N_{ψ}	Number of customers in LVDN/each phase
$\widetilde{N}/\widetilde{N}_{\psi}$	Number of customers whose labels are correctly identified in three phases/each phase
N^*_{ψ}	Number of customers whose labels are labelled as phase ψ

Variables

$f_{\phi}()$	Output of the RSN
$Loss_{SN}$	Loss of the RSN
$d^m_{n,\psi}$	Euclidean distance between customer n and three phases
$\iota^0_{n,\psi}/\iota_{n,\psi}$	Aggregated/normalized distance
$\iota'_{n,\psi}$	Binary phase label of customer n between customer n and three phases
L_{ψ}	Identified probability phase labels
L_{ψ}	Identified hard phase labels
τ	Kendall correlation coefficient
$K^*_{\psi}/K_{\psi}/K'_{\psi}$	Kendall correlation coefficient vectors between transformer data/variation across the same phases
ΔK_{ψ}	Residual vectors of Kendall correlation coefficients between adjacent iterations
$\Delta\tau^*_n$	Maximum difference in the three aggregated distances for customers n
G	Vector of $\Delta\tau^*_n$

Table 1
Summary of literature of approaches for phase identification.

Method	Ref.	Feature extraction	Incomplete dataset	Measurement error	No stepwise subtraction	Probability label
Voltage-based	[10]	✗	✓	✓	✗	✗
	[11]	✗	✓	✗	✗	✗
	[12]	✗	✗	✗	✓	✓
Active power-based	[4]	SA	✓	✓	✗	✓
	[13]	High-pass filter	✓	✗	✗	✗
	[15]	SA	✓	✗	✗	✗
	[16]	SA+selection	✓	✗	✗	✗
	[17]	SA	✓	✗	✗	✗
Proposed method		SA+WD	✓	✓	✓	✓

2. Phase identification framework

Phase Identification based on time-series load profiles could be taken as a data-pairing process, incorporating feature extraction, similarity calculation, etc. The framework of the proposed phase identification based on RSN is depicted in Fig. 2. The proposed phase identification approach consists of two stages: self-taught training (i.e., the lower level in Fig. 2) and phase label identification (i.e., the upper level in Fig. 2).

In the first stage, the proposed STT strategy consists of three steps: (1) pseudo-data generation, (2) feature extraction and (3) RSN training. The first step is to generate a pseudo transformer dataset, and the pseudo-phase labels of customers are derived from the pseudo-phase data to which the customer is grouped. The second step is feature extraction based on WD, which is used to construct the input dataset for RSN training. The third step is to train the RSN on the reconstructed dataset. The second stage consisted of phase label estimation and modification. The reconstructed datasets of initial SM data are first obtained by the same process in the first stage. The phase labels of customers are estimated by aggregating and normalizing the output of trained RSN in each window. Finally, a phase label modification strategy based on the Kendall correlation coefficient is introduced to assess and obtain phase labels. Next, a detailed explanation of each of the stages is presented.

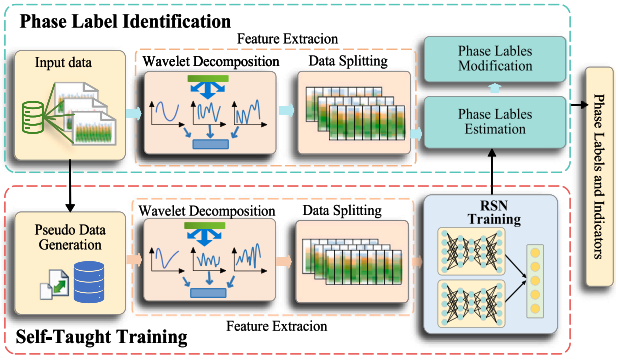


Fig. 2. Framework of the proposed phase identification approach: the self-taught training process of RSN in the lower level and the phase label identification in the upper level.

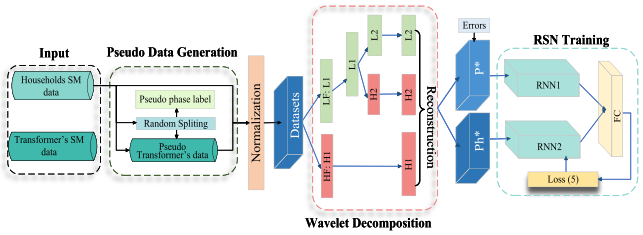


Fig. 3. Framework of self-taught training of RSN: Step I: pseudo data generation, Step II: feature extraction and Step III: RSN training.

2.1. Self-taught training

The training of the RSN requires labelled datasets (i.e., the phase labels), which are not available and are the destination of our approach. To address this issue, a self-taught training strategy is proposed using pseudo datasets. Fig. 3 illustrates the detailed framework of the STT strategy, which corresponds to the lower level in Fig. 2. The STT strategy consists of pseudo-data generation, feature extraction and RSN training. The constructed RSN is introduced first.

2.1.1. Recurrent neural network-based siamese network

In meta learning, SN is a common structure for analysing the similarity among datasets, including images and time-series data, which does not depend on the amount of training samples from the same distribution. SN is normally used in image identification, e.g., objection detection in videos, image classification and faulty identification of bearing. As shown in Fig. 2, the first half of the SN network is two networks with the same structure and shared parameters, indicating that two inputs are required for the training of SN and two feature embeddings are obtained after the feature extraction networks. The latter part of the SN network typically consists of a single-layer fully connected network, providing the distance or similarity of the two inputs.

Compared to image classification, self-correlation and dependencies exist in time-series data. Specifically, the data at the present time step shows a stronger correlation with the data preceding it but exhibits relatively weak correlations with the data from several weeks earlier. RNN is designed to deal with sequence datasets, which are capable of grasping the dynamic features and transferring them to the subsequent neurons through the recurrent neuron [28]. Thus, compared to feedforward neural networks and 1-dimensional convolutional neural networks (CNNs), RNN is more suitable for time-series data preprocessing and is taken as the siamese part of SN in our problem. Besides, the final layer of our network also employs a fully connected neural network, similar to the common structure.

2.1.2. Pseudo data generation

The pseudo-data generation aims to randomly divide the customer load profiles into three sub-datasets, which is consistent with the goal of phase identification. Given the T dimension customer data P_{uu} of N customers (as expressed as (1)), the nth row $P_{u,n}$ represents the data of the nth customer. The process of pseudo-data generation is summarized into two steps:

(1) Shuffle the time-series dataset P_{uu} and randomly divide it into three balance sub-datasets, which contain comparable numbers of customers.

(2) Obtain the pseudo transformer data P_p (as expressed as (2)) by summing the customer data in each sub-dataset.

$$P_{uu} = \begin{bmatrix} p_1^1 & p_1^2 & \cdots & p_1^T \\ p_2^1 & p_2^2 & \cdots & p_2^T \\ \vdots & \vdots & \ddots & \vdots \\ p_N^1 & p_N^2 & \cdots & p_N^T \end{bmatrix} = \begin{bmatrix} P_{u,1} \\ P_{u,2} \\ \vdots \\ P_{u,N} \end{bmatrix} \tag{1}$$

$$P_p = \begin{bmatrix} p_a^1 & p_a^2 & \cdots & p_a^T \\ p_b^1 & p_b^2 & \cdots & p_b^T \\ p_c^1 & p_c^2 & \cdots & p_c^T \end{bmatrix} = \begin{bmatrix} P_a \\ P_b \\ P_c \end{bmatrix} \tag{2}$$

(3) Pseudo label generation: based on the clustering results in step (1), the pseudo-phase label of households is determined according to the sub-datasets to which the customers are assigned. For example, if the dataset of household n is put into sub-dataset A (i.e., pseudo transformer dataset for phase A), the pseudo phase label of household n is set as A.

2.1.3. Feature extraction

There are three sub-steps in feature extraction: normalization, Wavelet decomposition and reconstruction. In order to alleviate the impact of SM data amplitudes on phase identification, the dataset P_{uu} and P_p are first normalized. Then, to reveal the similarity between transformer and customer data, Wavelet decomposition, as an efficient technique in time-series data analysis, is adopted to decompose the time-series datasets P_{uu} and P_p into multiple high-frequency H and low-frequency L components. An illustrated example of two-level Wavelet decomposition of load profile P_0 is formulated in expressions (3) - (6).

$$P_0 = L_1 \oplus H_1 \tag{3}$$

$$L_1 = L_2 \oplus H_2 \tag{4}$$

$$P_0 = L_2 \oplus H_2 \oplus H_1 \tag{5}$$

$$\dot{P}_0 = [L_2, H_2, H_1] \tag{6}$$

where $\oplus$ represents the Wavelet reconstruction. The subscripts of L and H represent the level of Wavelet decomposition. For instance, H_1 and H_2 represent the high-frequency components extracted through Wavelet decomposition at the first and second levels, respectively.

Given the extracted components, a reconstructed dataset $\dot{P}_0$ is obtained according to (6), which has the same dimension as the initial data. The reconstructed datasets of the initial datasets P_{uu} and P_p are represented by $\dot{P}_{uu}$ and $\dot{P}_p$. In the phase identification problem, we take the time-series SM data from the transformer and customers as the two inputs for the RSN, respectively. To define the dimension of the input layer of the RSN, the two time-series datasets $\dot{P}_{uu}$ and $\dot{P}_p$ are split into short-term data based on a pre-set window width w.

$$P_\psi^* = \begin{bmatrix} \dot{p}_\psi^1 & \dot{p}_\psi^2 & \cdots & \dot{p}_\psi^w \\ \dot{p}_\psi^{w+1} & \dot{p}_\psi^{w+2} & \cdots & \dot{p}_\psi^{2w} \\ \vdots & \vdots & \ddots & \vdots \\ \dot{p}_\psi^{(M-1)w} & \dot{p}_\psi^{(M-1)w+1} & \cdots & \dot{p}_\psi^{Mw} \end{bmatrix}, \forall \psi \in \mathcal{F} \tag{7}$$

$$P^*_{uu} = \begin{bmatrix} P^*_{u,1} \\ P^*_{u,2}, \\ \vdots \\ P^*_{u,N} \end{bmatrix} \tag{8}$$

where ψ and $\mathcal{F}$ represent the phase index and its set (i.e., $\mathcal{F} = \{a, b, c\}$), and M represents the total number of windows. The split data $P^*_{u,n}$ of customers has the same structure as P^*_{ψ}. Parameters with star superscripts represent the split data.

The width w of the window is taken as the input dimension of the RSN. The vectors P^m_{ψ} and $P^m_{u,n}$ represent the mth row in matrix P^*_{ψ} and $P^*_{u,n}$, respectively. In our approach, the P^*_h and P^*_{uu} are concatenated to construct the training datasets (i.e., each row in the matrix P in (9)). The first and the second columns of P are the transformer and customer data, respectively. Each row in P represents a sample, and the two columns are used as the two inputs for SN. This dataset is unlabelled data since it is assumed that there are no available phase labels and pre-knowledge of the topology.

According to the obtained pseudo phase labels in Section 2.1.2, the labels of the samples in dataset P are obtained. Specifically, if the nth customer belongs to the sub-dataset $\mathcal{N}_a$ (i.e., the index set of the customers who are connected to the pseudo phase a), the sample $[P^m_a, P^m_{u,n}]$ is taken as positive sample and its label is set as 1. The samples $[P^m_b, P^m_{u,n}]$ and $[P^m_c, P^m_{u,n}]$ are taken as negative samples and their labels are set as 0. $y^m_{a,n}$ represents the label of the sample in the mth window that consist of the pseudo transformer data P^m_a and the nth customer data $P^m_{u,n}$, which is defined by (10). The labels of samples, comprising customer data belonging to pseudo phases b and c, are defined using the same procedure.

$$P = \begin{bmatrix} P^1_a & P^1_{u,1} \\ P^1_b & P^1_{u,1} \\ P^1_c & P^1_{u,1} \\ \vdots & \vdots \\ P^M_a & P^M_{u,N} \\ P^M_b & P^M_{u,N} \\ P^M_c & P^M_{u,N} \end{bmatrix} \tag{9}$$

$$y^m_{\psi,n} = \begin{cases} 1, & \psi = a, \quad \forall n \in \mathcal{N}_a, \forall m \in \mathcal{M} \\ 0, & \psi \in \{b, c\}, \quad \forall n \notin \mathcal{N}_a, \forall m \in \mathcal{M} \end{cases} \tag{10}$$

where $\mathcal{M}$ is the index set of the windows and $\mathcal{N}$ is the index set of the customers in the LVDN. Y is the dataset of labels, i.e., consisting of all labels for each row in P.

A balanced training dataset contributes to enhancing the performance of the training of RSN, the positive samples S^+ (i.e., the samples with pseudo label 1 in P) are duplicated since the rate of negative samples to positive samples is 2 in the dataset P. Moreover, a normally distributed error matrix E with the same dimensions are added to the pre-processed customer datasets, preventing identical samples and over-fitting. Thus, the training dataset P_{train} for training RSN is obtained by Eq. (11).

$$P_{train} = \begin{bmatrix} P \\ S^+ \end{bmatrix} + E \tag{11}$$

2.1.4. RSN training

The final step of the STT strategy is to train RSN with given hyperparameters, e.g., learning rate, batch size, the maximum iteration, the optimizer, etc. The loss function of the constructed RSN is formulated as Eq. (13) [29].

$$D(P^m_{\psi}, P^m_{u,n}) = \|f_{\phi}(P^m_{\psi}) - f_{\phi}(P^m_{u,n})\| \tag{12}$$

$$Loss_{SN} = \frac{1}{N} \sum_{n}^{N} \sum_{m}^{M} (y^m_{\psi,n} \cdot (D(P^m_{\psi}, P^m_{u,n}))^2$$

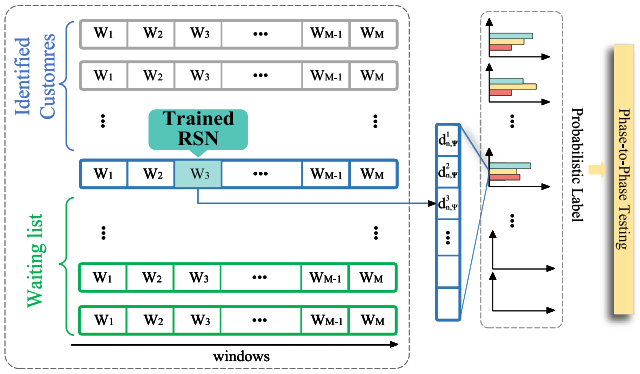


Fig. 4. Framework of probability phase label estimation based on trained RSN.

$$+ (1 - y^m_{\psi,n}) \cdot (max(\gamma - D(P^m_{\psi}, P^m_{u,n}), 0))^2) \tag{13}$$

where $D(\cdot, \cdot)$ represents the Euclidean distance and $f_{\phi}(\cdot)$ represents the output of the RSN. Hyperparameter γ is the margin parameter for the negative and positive samples.

The loss function $Loss_{SN}$ is used to measure the dissimilarity between the two outputs $f_{\phi}(P^m_{\psi})$ and $f_{\phi}(P^m_{u,n})$ for each sample and each window m. It consists of two terms. For positive samples, the loss is calculated as the squared Euclidean distance between the two outputs $f_{\phi}(P^m_{\psi})$ and $f_{\phi}(P^m_{u,n})$, denoted by $(D(P^m_{\psi}, P^m_{u,n}))^2$. This term aims to minimize the distance between the two columns of positive samples. For negative samples, the loss involves the term $(\max(\gamma - D(P^m_{\psi}, P^m_{u,n}), 0))^2$. This term is zero when the $D(P^m_{\psi}, P^m_{u,n})$ is larger than the margin γ. Otherwise, it is $(\gamma - D(P^m_{\psi}, P^m_{u,n}))^2$. This term aims to enlarge the distance between the two columns of negative positives and to enforce a distinct separation between positive and negative samples within the feature space. By minimizing the loss function in (13) during training, the model aims to identify the phase labels while maintaining a margin of separation between positive and negative samples, improving robustness in phase identification.

2.2. Phase label identification

A phase label identification strategy is proposed to identify the phase label using the output of the trained RSN and phase-to-phase correlation coefficients, which consists of probability phase label estimation and modification. The reconstructed datasets of initial SM data (i.e., the transformer and customer data) are first obtained by the same process as the second step in the lower level (i.e., Section 2.1.3), which has the same format as in (9).

2.2.1. Phase label estimation

The framework of the proposed phase label estimation based on RSN is demonstrated in Fig. 4. The data in each window W_m are the processed initial SM data (i.e., the high-frequency and low-frequency components of the data). Given the trained RSN, the distance $d^m_{n,\psi}$ between customers and the three phases are obtained in each window. The probability phase labels L_{ψ} of customers are calculated by aggregating the distance in each window, as formulated in Eq. (14)–(16).

$$\iota^0_{n,\psi} = \sum_{m}^{M} d^m_{n,\psi}, \quad \forall n \in \mathcal{N}, \forall \psi \in \mathcal{F} \tag{14}$$

$$\iota_{n,\psi} = \frac{\iota^0_{n,\psi}}{\sum_{\psi}^{\mathcal{F}} \iota^0_{n,\psi}}, \quad \forall n \in \mathcal{N}, \forall \psi \in \mathcal{F} \tag{15}$$

$$L_{\psi} = \begin{bmatrix} \iota_{1,a}, & \iota_{1,b}, & \iota_{1,c} \\ \vdots, & \vdots, & \vdots \\ \iota_{N,a}, & \iota_{N,b}, & \iota_{N,c} \end{bmatrix} \tag{16}$$

Algorithm 1: Phase Label Identification

Input: Trained RSN $f_\phi(\cdot)$, Reconstructed data P^{t*}, Thresholds ϵ
for $i \le N$ **do**
  $\iota^0_{n,\psi} = \sum_m^M \|f_\phi(P^w_{\psi,m}) - f_\phi(P^w_{u,n,m})\|$
end
Calculate L_ψ using Eq. (14) and (15)
$\widetilde{P}_\psi \leftarrow$ aggregation based on L_ψ
Obtain K_ψ by Eq. (17) to (20)
Calculate G and L'_ψ using L_ψ
Obtain index vector I_p based ϵ and G
for $i \in \mathcal{I}$ **do**
  for $j \le 2$ **do**
    Exchange label in the row i in L'_ψ
    Obtain $\Delta K_{i,\psi}$ by (20) and (24)
  end
  Obtain position index j^*
  $L'_\psi \leftarrow$ Update L'_ψ based on j^*
end
$K^*_\psi \leftarrow$ Update K_ψ based on L'_ψ
Output: Phase label matrix L'_ψ and matrix K^*_ψ

The aggregated distance $\iota^0_{n,\psi} \in [0, \infty]$ and probability phase labels $\iota_{n,\psi} \in [0, 1]$. The smaller the value of $\iota^0_{n,a}$, the greater the similarity between the nth customer data and transformer data in phase a, and vice versa. Thus, the position index corresponding to the minimum value in each row of $L_{n,\psi}$ is considered as the index of the phase to which the customer is most likely connected.

2.2.2. Phase label modification

As the number of correctly aggregated load profiles within the same phase increases, the phase-to-phase correlation coefficients exhibit a larger disparity [16]. Meanwhile, the growth rate of the correlation coefficients between different phases with the aggregation of load profiles is significantly lower than that of the correlation coefficients between the same phases. Thus, a Kendall correlation coefficients-based phase label modification strategy is introduced to evaluate the obtained phase labels (i.e., L_ψ) and modify the labels with lower confidence. Specifically, the Kendall correlation coefficient is employed to evaluate the correlation between variations in transformer data across the same phases (i.e., the real data P_ψ and identified data $\widetilde{P}_\psi$).

$$\Delta P_\psi = [p^2_\psi - p^1_\psi, p^3_\psi - p^2_\psi, \dots, p^T_\psi - p^{(T-1)}_\psi], \quad \forall \psi \in \mathcal{F} \tag{17}$$

$$\Delta \widetilde{P}_\psi = [\widetilde{p}^2_\psi - \widetilde{p}^1_\psi, \widetilde{p}^3_\psi - \widetilde{p}^2_\psi, \dots, \widetilde{p}^T_\psi - \widetilde{p}^{(T-1)}_\psi], \quad \forall \psi \in \mathcal{F} \tag{18}$$

$$\tau(\Delta P_\psi, \Delta \widetilde{P}_\psi) = F_K(\Delta P_\psi, \Delta \widetilde{P}_\psi), \quad \forall \psi \in \mathcal{F} \tag{19}$$

$$K_\psi = [\tau(\Delta P_a, \Delta \widetilde{P}_a), \tau(\Delta P_b, \Delta \widetilde{P}_b), \tau(\Delta P_c, \Delta \widetilde{P}_c)], \tag{20}$$

$$K^*_\psi = [\tau(P_a, \widetilde{P}_a), \tau(P_b, \widetilde{P}_b), \tau(P_c, \widetilde{P}_c)], \tag{21}$$

where $\widetilde{p}^t_\psi$ and $\Delta \widetilde{P}_\psi$ represent identified transformer data and the variations, respectively. The function $F_K(\cdot)$ represent the function that is used to calculated Kendall correlation coefficient [30].

Meanwhile, the Kendall correlation coefficients K^*_ψ between transformer data across the same phases are used as the indicator to show the credibility of the final phase labels in each phase while revealing the recall and precision of phase identification, which is similar to the purity in the classification field. The closer the value of $\tau(\cdot,\cdot)$ is to 1, the higher the precision and recall rate of the phase identification. The process of phase label identification based on RSN is shown in Algorithm 1.

The maximum difference $\Delta\iota_n$ among the three aggregated distances $[\iota^0_{\cdot,a}, \iota^0_{\cdot,b}, \iota^0_{\cdot,c}]$ of each customer is calculated by (22) and stored in vector G. Matrix L_ψ is converted to hard phase labels L'_ψ by converting the smallest value in each row to 1, the other values are replaced by 0. The hard labels are subjected to the constraint (23). If the maximum difference $\Delta\iota_n$ of customer i is smaller than the threshold ϵ, the phase labels of the customer i are taken as a label with lower confidence. The indexes of these customers are stored in set $\mathcal{I}$. Then, the position of label 1 in row i is exchanged with the other two, respectively. For instance, $[1,0,0]$ are replaced by $[0,1,0]$ and $[0,0,1]$, respectively. The residual phase-to-phase correlation coefficients before and after the label modification are calculated by (24). If the residual ΔK_ψ of phase-to-phase correlation is positive, the swapped position j^* that leads to the most significant increase is taken as a correlate one and the corresponding results are taken as the modified phase label of customer i. After modifying all phase labels with lower confidence, the phase-to-phase correlation coefficients K^*_ψ obtained under the final phase labels are taken as the credibility of labels in each phase.

$$\Delta\iota_n = max\{|\iota_{n,a} - \iota_{n,b}|, |\iota_{a,n} - \iota_{n,c}|, |\iota_{n,b} - \iota_{n,c}|\}, \quad \forall n \in \mathcal{N} \tag{22}$$

$$\sum_\psi^{\mathcal{F}} \iota'_{n,\psi} = 1, \quad \forall n \in \mathcal{N} \tag{23}$$

$$\iota'_{n,a}, \quad \iota'_{n,b}, \quad \iota'_{n,c} \in [0,1]$$

$$\Delta K_\psi = K'_\psi - K_\psi \tag{24}$$

where K'_ψ represents the phase-to-phase correlation coefficients during the phase label modification process, which will be identical to K^*_ψ after the final step.

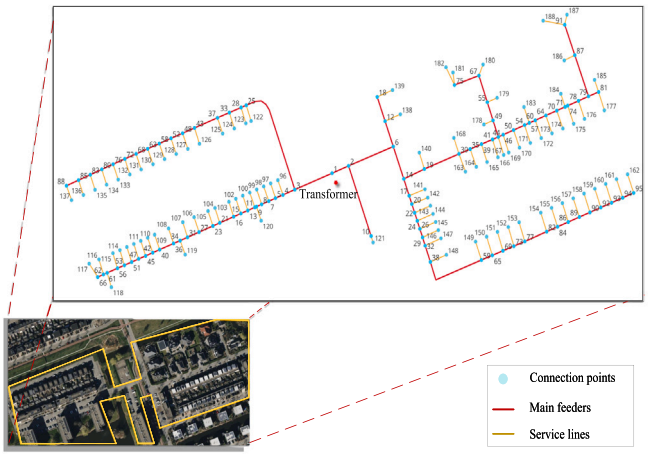


Fig. 5. Topology of the 188-bus LVDN in the Netherlands.

3. Case of study

In this section, the feasibility and accuracy of the proposed phase identification approach are verified on the IEEE 116-bus test feeder case (denoted as LV-116) and an 188-bus LVDN (denoted as LV-188) in the Netherlands with different types of cables. These two LVDNs are obtained from [31] and [32]. The topology of the 188-bus LVDN is illustrated in Fig. 5. The base three-phase voltage is 0.4 kV. Three datasets are used:

(1) Dataset I: time-series profiles for each household with 15 min time resolution are selected and scaled from [33].

(2) Dataset II: time-series profiles with 1 h time resolution are collected from the SM in the Netherlands.

(3) Dataset III: a synthetic dataset is generated from a chi-square distribution, with Gaussian errors introduced simultaneously.

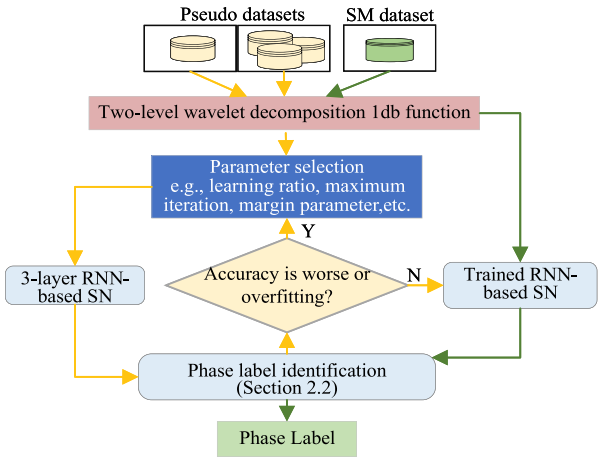


Fig. 6. Framework of parameters tuning and online application represented by yellow lines and green lines, respectively.

Table 2
Summary of key parameters in the proposed approach.

Parameter	Values
Power factor ($\cos\theta$)	0.95
WD base function (Haar wavelet)	db1
Margin parameter (γ)	1
Number of neurons in hidden layer	500
Learning rate	0.005
Activation function	ELU

The power factor $cos\theta$ is set at 0.95 for each household, which is a common value for customers. The time-series load profiles of transformers are generated by a power flow model [34]. For the training of RSN, the learning rate is 0.005, the maximum iteration is set at 100 and the margin parameter γ is set at 1. The input dimension of the RNN is 96 (i.e., the same as the dimension of a one-day sample with a 15 min time resolution) and the number of neurons in the hidden layer is 500, which is also the dimension of the fully connected layer. The activation function is ELU. The SGD optimizer is used to train the RSN. The Haar wavelet and the db1 base function are adopted in the Wavelet decomposition. Besides, the threshold ϵ is set the same as the margin parameter γ (i.e., 1). These parameters were chosen as a result of the cross-validation method, aiming to ensure the performance of the proposed approach. For instance, the pre-set maximum iteration is used to avoid overfitting while ensuring the accuracy of identification. The general process of the parameter tuning is depicted as the yellow lines in Fig. 6, and the parameters are summarized in Table 2.

3.1. Performance evaluation

Two LVDNs with similar load profiles but different topologies were used to test the feasibility and accuracy of the proposed phase identification strategy. The load profiles in both two LVDNs were selected and scaled from Dataset I. This case aims to analyse the impact of topology and power loss on the performance of our approach. Moreover, there are measurement errors in SM data induced by SM and communication issues, which is a common phenomenon in DNs. Thus, Gaussian noise was generated according to the SM class and added to the initial dataset to emulate the measurement errors. According to the IEC 62053–21 [4], four classes of SM were considered in this case, including 0%, 0.5%, 1% and 2%. Meanwhile, the amount of available time-series data (i.e., the value T) will impact the aggregated distance matrix L_ψ and correlation matrix K_ψ, influencing the accuracy of phase identification. The amount of available data represents the number of days during which households' data were measured and collected. Thus, the datasets with

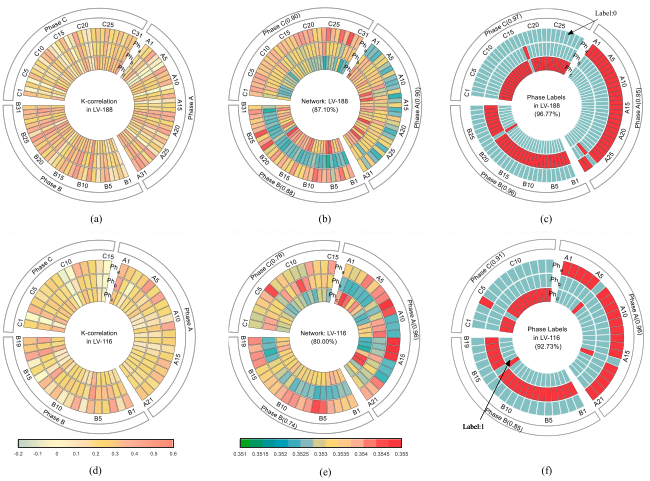


Fig. 7. The correlation coefficients in (a) and (d); the normalized distance matrix L_ψ in (b) and (e) and the phase label in (c) and (f). The correlation coefficients between identified and real phase data are shown in brackets.

Table 3
Accuracy of proposed approach (%) under dataset I with multiple measurement error.

DNs	Class of SM	The amount of SM Data (Day)			
		5	10	15	20
LV-188	0	92.90 ± 2.2	97.42 ± 2.2	97.85 ± 1.5	99.35 ± 0.6
	0.5%	93.55 ± 3.6	96.99 ± 1.4	97.63 ± 1.4	99.57 ± 0.6
	1%	90.75 ± 3.3	95.48 ± 1.2	96.77 ± 0.8	99.57 ± 0.6
	2%	91.82 ± 1.2	94.41 ± 1.8	95.91 ± 1.4	99.14 ± 1.2
LV-116	0	88.00 ± 2.4	89.82 ± 4.9	92.73 ± 7.4	93.46 ± 2.8
	0.5%	86.54 ± 2.8	91.27 ± 1.5	89.45 ± 2.0	92.73 ± 4.5
	1%	86.54 ± 1.0	87.27 ± 1.3	89.09 ± 6.7	93.09 ± 4.5
	2%	85.45 ± 4.1	86.55 ± 6.0	89.45 ± 2.7	90.18 ± 2.8

5, 10, 15 and 20 days under the above four types of errors were used to evaluate our approach. On the other hand, the quality of the generated pseudo dataset, as the training datasets, influences the training of RSN, which therefore impacts the calculation of distances $\iota_{n,\psi}$. The approach was executed twenty times, and the average accuracy of five solutions with large K_ψ and the corresponding standard deviation were recorded, which are summarized in Table 3 and Fig. 7.

As shown in Table 3, given more than 10 days of the Dataset I with measurement errors, the proposed approach correctly identified phase labels for at least 90% of customers in both networks. As expected, with the increasing measurement error magnitudes, there is a slight decrease in phase identification accuracy in both LVDNs, specifically ranging from 2% to 4%. With the increase in the amount of available SM data, the phase identification accuracy increases. Compared to the negative impact of measurement error, the positive impact of the amount of SM data is more significant, indicating more SM data could mitigate the negative effect induced by the measurement error of SM on the phase identification accuracy. Besides, there exists a discrepancy of approximately 5%–10% in the phase identification accuracy between the two LVDNs, showing that power losses and errors impact the phase identification accuracy.

Given the 15-day dataset with 1% measurement error, the Kendall correlation coefficients, normalized distance (i.e., the matrix L_ψ) and the modified phase labels (i.e., the matrix L'_ψ) are demonstrated in Fig. 7. From Fig. 7(a) and (d), it is hard to directly identify the phase labels by the correlations among the initial SM data. However, the normalized distance L_ψ in Fig. 7(b) and (e) show clearer boundaries between positive and negative samples. After the phase label modification, the hard phase label in Fig. 7(c) and (f) described the output of the proposed approach, which reveals the majority of the true labels

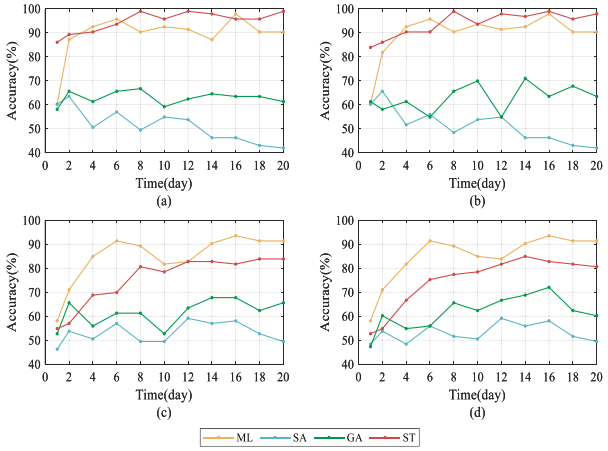


Fig. 8. Accuracy comparison under Dataset I in (a) and (b) and Dataset II in (c) and (d) without and with a 2% measurement error, respectively.

of customers. The phase-to-phase correlation coefficients K_{ψ}^{*} in the brackets also depicted the purity of the phase labels in each phase.

3.2. Method comparison

The accuracy, recall, and precision of the proposed phase identification approach were compared to that of similar approaches (i.e., the approaches based on load profiles) under multiple scenarios, including the ML-based clustering approach [13], the saliency analysis (SA)-based approach [15], and the genetic algorithm(GA)-based approach [17]. These three methods represent three typical phase identification approaches: (1) phase identification using classical machine learning techniques, such as clustering algorithms; (2) approaches that rely heavily on the saliency or variability of time-series data; and (3) optimization-based solutions for phase identification. Dataset I and Dataset II with and without 2% measurement error were used to evaluate the performance of the above approaches in LV-188. The accuracy is depicted in Fig. 8, and the recall and precision are summarized in Table 4.

$$Accuracy = \frac{\widetilde{N}}{N} \times 100\% \tag{25}$$

$$Recall = \frac{\widetilde{N}_{\psi}}{N_{\psi}}, \qquad \psi \in \mathcal{F} \tag{26}$$

$$Precision = \frac{\widetilde{N}_{\psi}}{N_{\psi}^{*}}, \qquad \psi \in \mathcal{F} \tag{27}$$

where $\widetilde{N}$ and $\widetilde{N}_{\psi}$ are the number of households whose labels are correctly identified in three phases and each phase, respectively. N_{ψ} is the true number of households in phase ψ. N_{ψ}^{*} represents the number of households whose labels are labelled as phase ψ.

As depicted in Fig. 8, the accuracy of the SA-based and GA-based approach is lower than 70% and does not increase when given more SM data, which is also not sensitive to time resolution and measurement errors. The performance of the proposed approach is similar to that of the ML-based clustering approach, i.e., above 80% with higher amounts of SM data and not sensitive to measurement errors. Given more than 10 days of data, the accuracy of the proposed approach reached above 93% and 82% under the two datasets with 2% measurement errors, respectively, indicating that the larger time resolution decreases the accuracy of the approach. The ML-based clustering approach demonstrates robust performance across variations in the time resolution of SM data, effectively managing datasets with both regular and irregular temporal intervals. However, the proposed approach exhibits superior capabilities in capturing the variations within the lower-time resolution

Table 4
Recall (%) and precision comparison (%) under dataset with measurement errors.

Approach	Phase	Dataset I		Dataset II	
		Recall	Precision	Recall	Precision
ML	a	96.77	85.71	90.32	87.50
	b	100	96.88	87.10	90.00
	c	80.65	96.15	93.55	93.55
SA	a	35.48	44.00	48.39	55.56
	b	48.39	51.72	41.94	56.52
	c	54.84	43.59	77.42	55.81
GA	a	67.74	72.41	74.19	76.67
	b	77.42	66.67	67.74	60.00
	c	67.74	75.00	64.52	71.43
ST	a	93.55	100	87.10	87.10
	b	96.77	96.77	80.65	86.21
	c	100	93.64	87.10	81.82

data, making it particularly effective in scenarios characterized by irregular consumption patterns or datasets with fine-grained temporal dependencies. This ability to model complex relationships highlights the versatility of the proposed method in handling real-world SM data, where such irregular variations in the time-series dataset are often encountered. Moreover, the proposed method offers distinct advantages beyond accuracy. In addition to outperforming some existing approaches, it provides not only probabilistic and hard labels but also indicators of phase purity. These outputs offer deeper insights into the phase identification process and can be instrumental for applications such as load balancing in DNs, where understanding the confidence level of accurate labels in each phase and its distribution is critical for operational efficiency.

As shown in Table 4, the recall and precision metrics follow a trend similar to the accuracy results observed in Fig. 8. Notably, the recall and precision across all phases for the proposed method are more consistent and balanced compared to the other approaches. This consistency is crucial for ensuring reliable phase identification across the three phases, minimizing the risk of phase misclassification, which provides a reference for load balancing in DNs.

3.3. Impact of incomplete data

On the other hand, the collected load profiles might be incomplete due to communication issues. The incompleteness of SM data point is random or continuous [13]. To evaluate the impact of incomplete datasets on the accuracy of the proposed approach, two scenarios under Dataset I and Dataset III were considered in this case: datasets with random missed data points and datasets with unmetered customers. The incomplete percentage was set between 0% and 20% in the first scenario and the incomplete ratio represents the proportion of missing data points relative to the dimension of the input data (i.e., T). In the second case, the incomplete ratio was set between 0% and 40%, which represents the ratio between the number of unmetered households and the total number of households. The missed data points and unmetered households were removed from the initial data, and the remaining data were used as the input data for the approach. The proposed approach was executed twenty times, and the five solutions with the correlation coefficients K_{ψ} closest to 1 were saved. To ensure a diverse range of missing data scenarios and maintain statistical significance, the positions of missing data points and the identities of unmetered households were randomly selected for each simulation, following a uniform distribution. The average accuracy under the above two scenarios is depicted in Fig. 9.

As shown in Fig. 9(a)–(c), the average accuracy decreases with the increase in incomplete ratio while increasing with the increasing amount of available SM data. The removed missing data points impact the variation of time-series data, influencing the extracted features

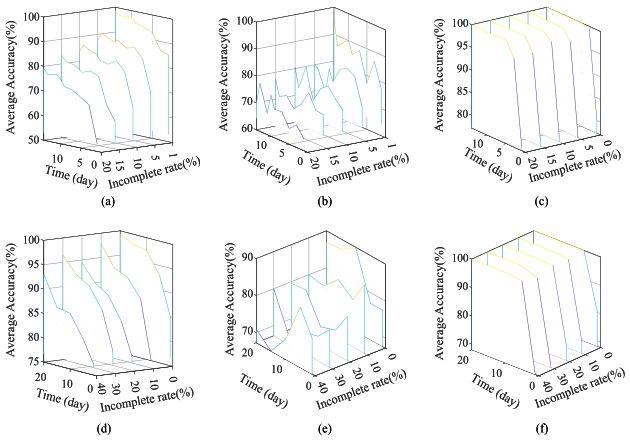


Fig. 9. Average accuracy under Dataset I in (a,d) in LV-188 and in (b, e) in LV-116, and Dataset III in (c) and (f) with random missing data and unmetered households, respectively.

and the similarity calculation, which impacts the accuracy of phase identification. Given 10-day datasets with 5% missing data points, the proposed approach accurately identifies at least 85% of the customer phase labels in the LV-188 network and 80% in the LV-116 network. When the incompleteness ratio increases to 20%, the accuracy drops to 77% in the LV-188 network and 75% in the LV-116 network. However, as shown in Fig. 9(c), when the available simulation data are less than 5 days, the phase identification accuracy declines below 80%. Conversely, when the data exceeds 5 days with varying incompleteness, the phase identification accuracy remains above 99%, indicating that the impact of missing data is negligible in such scenarios.

Fig. 9(d)–(f) depict the relationships between accuracy, the number of unmetered households, and the amount of available SM data. The average accuracy decreases as the percentage of unmetered households increases, attributable to the limited available SM data resulting in a limited training dataset. Additionally, the accuracies exhibit a rapid decline with the increasing incompleteness ratio, particularly when 5-day datasets are available. For datasets containing more than 10-day time-series data, the accuracy under the scenarios decreased by 13%, 20%, and 2% as the percentage of unmetered households increased from 0% to 40%, respectively. On the other hand, compared to measurement errors in SM data and the presence of unmetered households, the negative impact of missing data points in time-series data on phase identification is more pronounced, resulting in lower accuracy.

4. Conclusion

A meta learning based self-taught phase identification approach was proposed based on SM data in this paper. Compared to existing methods and conventional neural networks training, the proposed STT strategy enables the RSN to identify phase labels without requiring extensive months of labelled datasets. This strategy reduces the dependency on large-scale, labelled data, making the approach more efficient and practical in real-world applications where data is limited or incomplete. The feasibility and accuracy of the proposed approach were evaluated on three datasets and multiple scenarios, including datasets with missing data points, unmetered households and multiple dimensions. Not only the probability and hard phase labels of customers but also the purity of each phase were provided by the proposed approach. The results showed that the probability phase labels could be represented by the distance matrix obtained from the trained RSN, and the Kendall correlation coefficients were validated to assess the purity of the phase labels in each phase. Furthermore, the experimental results demonstrated that the proposed approach outperformed existing methods that rely on stepwise subtraction of identified users' data from transformer data in terms of accuracy while requiring a smaller amount of SM data. The results also indicated that the proposed approach is more robust to unmetered houses and measurement errors in comparison to missing data points. On the other hand, more available data alleviated the negative impact caused by missing data on the phase identification accuracy.

CRediT authorship contribution statement

Dong Liu: Writing – original draft, Validation, Software, Methodology, Conceptualization. **Juan S. Giraldo:** Writing – review & editing. **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.

Acknowledgments

This research is supported by China Scholarship Council (CSC) (Grant No. 202206130017).

Data availability

Data will be made available on request.

References

[1] González-Cagigal MÁ, Rosendo-Macías JA, Gómez-Expósito A. Identification of the phase connectivity in distribution systems through constrained least squares and confidence-based sequential assignment. Int J Electr Power Energy Syst 2022;143:108445.
[2] Wang W, Yu N. Maximum marginal likelihood estimation of phase connections in power distribution systems. IEEE Trans Power Syst 2020;35(5):3906–17.
[3] Shi Z, Xu Q, Liu Y, Wu C, Yang Y. Line parameter, topology and phase estimation in three-phase distribution networks with non-μPMUs. Int J Electr Power Energy Syst 2024;155:109658.
[4] García S, Mora-Merchán JM, Larios DF, Personal E, Parejo A, León C. Phase topology identification in low-voltage distribution networks: A Bayesian approach. Int J Electr Power Energy Syst 2023;144:108525.
[5] Hoogsteyn A, Vanin M, Koirala A, Van Hertem D. Low voltage customer phase identification methods based on smart meter data. Electr Power Syst Res 2022;212:108524.
[6] Li N, Hakvoort RA, Lukszo Z. Cost allocation in integrated community energy systems-a review. Renew Sustain Energy Rev 2021;144:111001.
[7] Wen MH, Arghandeh R, von Meier A, Poolla K, Li VO. Phase identification in distribution networks with micro-synchrophasors. In: 2015 IEEE power & energy society general meeting. IEEE; 2015, p. 1–5.
[8] Luan W, Peng J, Maras M, Lo J, Harapnuk B. Smart meter data analytics for distribution network connectivity verification. IEEE Trans Smart Grid 2015;6(4):1964–71.
[9] Vycital V, Ptacek M, Janik D, Toman P. Voltage based phase identification method, robustness and validation. In: 2022 22nd international scientific conference on electric power engineering. EPE, IEEE; 2022, p. 1–6.
[10] Dahale S, Pahwa A, Natarajan B. Phase identification in unobservable distribution systems. IEEE Trans Power Deliv 2023;38(5):3067–75.
[11] Zhou L, Li Q, Zhang Y, Chen J, Yi Y, Liu S. Consumer phase identification under incomplete data condition with dimensional calibration. Int J Electr Power Energy Syst 2021;129:106851.
[12] Peña BD, Blakely L, Reno MJ. Online data-driven detection of phase changes in evolving distribution systems. In: 2023 IEEE power & energy society innovative smart grid technologies conference. ISGT, IEEE; 2023, p. 1–5.
[13] Hosseini ZS, Khodaei A, Paaso A. Machine learning-enabled distribution network phase identification. IEEE Trans Power Syst 2020;36(2):842–50.
[14] Wang M-H, Lu S-D, Liao R-M. Fault diagnosis for power cables based on convolutional neural network with chaotic system and discrete wavelet transform. IEEE Trans Power Deliv 2022;37(1):582–90.
[15] Xu M, Li R, Li F. Phase identification with incomplete data. IEEE Trans Smart Grid 2018;9(4):2777–85.

[16] Jimenez VA, Will A, Rodriguez S. Phase identification and substation detection using data analysis on limited electricity consumption measurements. Electr Power Syst Res 2020;187:106450.
[17] Jimenez VA, Will A. A new data-driven method based on niching genetic algorithms for phase and substation identification. Electr Power Syst Res 2021;199:107434.
[18] Therrien F, Blakely L, Reno MJ. Assessment of measurement-based phase identification methods. IEEE Open Access J Power Energy 2021;8:128–37.
[19] Chicco D. Siamese neural networks: An overview. Artif Neural Netw 2021;73–94.
[20] Lu Z, Bian Y, Yang T, Ge Q, Wang Y. A new siamese heterogeneous convolutional neural networks based on attention mechanism and feature pyramid. IEEE Trans Cybern 2024;54(1):13–24.
[21] Liu J, Fu Y. Renewable energy forecasting: A self-supervised learning-based transformer variant. Energy 2023;284:128730.
[22] Ren Z, Luo W, Yan J, Liao W, Yang X, Yuille A, Zha H. STFlow: Self-taught optical flow estimation using pseudo labels. IEEE Trans Image Process 2020;29:9113–24.
[23] Chen X, Li B, Proietti R, Zhu Z, Yoo SJB. Self-taught anomaly detection with hybrid unsupervised/supervised machine learning in optical networks. J Lightwave Technol 2019;37(7):1742–9.
[24] Javaid N, Jan N, Javed MU. An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids. J Parallel Distrib Comput 2021;153:44–52.
[25] Kummerow A, Monsalve C, Bretschneider P. Siamese recurrent neural networks for the robust classification of grid disturbances in transmission power systems considering unknown events. IET Smart Grid 2022;5(1):51–61.
[26] Yu M, Wang B, Lu L, Bao Z, Qi D. Non-intrusive adaptive load identification based on siamese network. IEEE Access 2022;10:11564–73.
[27] Dong Z, Hou K, Liu Z, Yu X, Jia H, Tang P, Pei W. A fast reliability assessment method for power system using self-supervised learning and feature reconstruction. Energy Rep 2023;9:980–6.
[28] Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. 2015, arXiv preprint arXiv:1506.00019.
[29] Kalita I, Roy M. Class-wise subspace alignment-based unsupervised adaptive land cover classification in scene-level using deep siamese network. IEEE Trans Neural Netw Learn Syst 2023;34(7):3323–34.
[30] Zhang B, Liu S, Dong H, Zheng S, Zhao L, Zhu R, Zhao L, Lin Z, Yang L, Wang Q. Data-driven abnormity assessment for low-voltage power consumption and supplies based on CRITIC and improved radar chart algorithms. IEEE Access 2020;8:27139–51.
[31] Khan MA, Hayes BP. A reduced electrically-equivalent model of the IEEE European low voltage test feeder. In: 2022 IEEE power & energy society general meeting. PESGM, IEEE; 2022, p. 1–5.
[32] Liu D, Giraldo JS, Palensky P, Vergara PP. Topology identification and parameters estimation of lv distribution networks using open gis data. Int J Electr Power Energy Syst 2025;164:110395.
[33] Schneider KP, Mather B, Pal BC, Ten C-W, Shirek GJ, Zhu H, Fuller JC, Pereira JLR, Ochoa LF, de Araujo LR, et al. Analytic considerations and design basis for the IEEE distribution test feeders. IEEE Trans Power Syst 2017;33(3):3181–8.
[34] Vergara PP, López JC, Rider MJ, Da Silva LC. Optimal operation of unbalanced three-phase islanded droop-based microgrids. IEEE Trans Smart Grid 2017;10(1):928–40.