ORIGINAL RESEARCH **Open access**

Robust Consumer Phase Identification in Distribution Networks under Uncertainty Using Graph Neural Networks

Arjun Nair[1], Kavita Rao[1*], Sameer Patel[2], Rohan Bhat[2]

Abstract

A large number of distribution system operators struggle to correctly document or sustain up-to-date phase connections among the many LV customers. Numerous consumer phase identification (CPI) techniques using voltages, powers, or further measurements appear in scholarly works. Because gathering voltage data presents notable technical hurdles, approaches focused on power measurements tend to be favored. Thus, this paper introduces a new power-driven CPI method that makes use of Graph Neural Networks (GNNs). The proposed CPI process fabricates synthetic transformer power profiles by applying random phase arrangements to the recorded load profiles, and these are jointly utilized for training the GNN model. Afterwards, the GNN model undergoes evaluation with real measured transformer and load power profiles. Assessment of the method's capability takes place inside a test low voltage grid featuring 55 loads exposed to several scenarios: Photovoltaic penetration, Photovoltaics under maintenance, measurement errors, unmetered consumption, uncertain grid asset parameters, and wrong phase connections. Extra validation performed on a genuine low voltage grid of 111 loads verifies how well the method scales. Findings demonstrate that the GNN model secures precision levels above 90% across most situations, exceeding multiple leading contemporary techniques.

Keywords Consumer phase identification, Graph neural network, Noise level, PV penetration, Unmetered consumption

*Correspondence:
Kavita Rao
kavita.rao@iitm.ac.in
[1] Department of AI for Sustainable Energy, Faculty of Electrical Engineering, IIT Madras, Chennai, India
[2] Department of Intelligent Power Networks, Faculty of Engineering, IIT Bombay, Mumbai, India

Introduction

Background and motivations

Electric power networks normally receive detailed documentation and oversight at the high-voltage (HV) and medium-voltage (MV) levels, supported by proper automation, control mechanisms, and geographic information systems (GIS) [1]. On the other hand, low-voltage (LV) distribution networks seldom include fully integrated systems, restricting monitoring mainly to automation distributed at the transformer substation level and, more recently, to smart metering infrastructure (SMI) [2]. In detail, distribution system operators (DSOs) routinely maintain connectivity records for every customer across transmission and primary distribution networks [3]. Conversely, the connectivity model outlining downstream customer linkages from the substation—along with their phase allocation details—frequently proves imprecise or completely unknown in LV distribution grids [4].

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

LV distribution grids incorporate three-, two-, and single-phase lines and contain a variety of multi- and single-phase loads or distributed energy resources (DERs). Even so, residential customers in many countries consist chiefly of single-phase loads [1, 5]. Given that the actual phase connection for LV customers remains unknown, incorrect, or mismatched [6], DSOs commonly find it difficult to properly log or revise lateral phasing details for the vast number of LV customers [1, 7]. Usually, this data gets entered by hand through field teams when customers first link to the grid. In addition, GIS-provided phase connection records can lag behind reality under many circumstances, including phase wiring misconfiguration, registration errors, grid reconfiguration, fault restoration, repairs or maintenance operations, load transfer, grid upgrades, potential cyberattacks, and new connections [7-9]. What is more, end-customer phase connections may remain entirely unrecorded in aging networks, distant distribution transformers, or bundled conductor arrangements [8]. Also, the SMI cannot establish phase connectivity because the required capabilities are not present in current smart meters (SMs) [10].

Today, determining and entering phase connectivity into the GIS system is done manually, creating major obstacles for DSOs due to the sheer number of end-user meters and complex wiring layouts [11]. Chiefly, confirmation of phase data relies on drawn-out, high-cost, and error-prone site visits that require significant labor, large staffing levels, and substantial material investment [12]. Aside from these manual procedures, different CPI techniques are obtainable. Approaches that rely on signal injection are used in compact electrical grids; however, applying them at every LV connection point is inefficient and increases capital and ongoing maintenance costs [7]. Hardware-based methods require precise phasor measurement units (PMUs) and reliable communication between the customer site and the transformer. Such solutions remain rare at LV levels, however, because Supervisory Control and Data Acquisition (SCADA) systems are scarce and overall costs for materials, duration, and personnel are very high [13, 14].

On top of that, implementing hardware techniques may trigger short-term power outages, disrupting steady electricity delivery and causing financial harm to both DSOs and end users. Further options pair SMs together with Power Line Communication (PLC) lines, in which the meter—triggered by an internal feature—dispatches signals over the phases so the substation receiver can match them [4]. Different strategies depend on SM voltage readings and require an additional device to identify phases by analyzing the phase difference between the substation reference and the SM voltage phase [15]. Nonetheless, techniques using SMs require access to customer sites, additional staff, higher device costs, and remain vulnerable to operator errors. Moreover, the SMI needs to exit normal service and function via a PLC, despite other protocols such as General Packet Radio Service (GPRS), Wi-Fi, and ZigBee also being widely deployed [15-17].

Reliable awareness of phase connectivity is essential for operating, monitoring, designing, and maintaining LV grids. Even though proper grid functioning depends on various activities (e.g., power flow optimization, grid rebalancing, capacity planning of assets, power quality control, loss localization, state estimation, fault detection and reconfiguration/restoration analysis), these often face restrictions at the LV level because phase allocation stays hidden [3, 7]. Increasing DER adoption creates extra difficulties for DSOs who must calculate hosting capacity at each link; yet accurate computations remain impossible without known phase connectivity [13, 15]. Customer-phase connectivity knowledge is a vital element in many distribution grid management activities. As a result, DSOs pursue cheaper, faster alternatives to standard practices, shifting their focus to data-driven solutions.

Table 1. Overview of existing power-based CPI approaches.

Ref.	Multi-phase	DER	DER	No prior	Noise	Unmetered	Parametric	Uncertain grid	Sca
Empty Cell	loads/DERs	penetration	maintenance	knowledge	impact	consumption	analysis	parameters	ar
[1]	✓	–	–	✓	–	–	–	–	
[2]	✓	–	–	✓	–	–	–	–	

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

[7]	–	–	–	✓	✓	–	–	–	
[8]	–	–	–	✓	✓	–	–	–	
[9]	–	–	–	✓	–	✓	–	–	
[12]	–	–	–	✓	–	✓	–	–	
[13]	–	–	–	✓	✓	✓	–	–	
[17]	–	–	–	–	–	✓	–	–	
[18]	✓	✓	–	✓	✓	✓	✓	–	
[19]	✓	–	–	✓	–	–	–	–	
[20]	✓	–	–	✓	–	–	–	–	
[21]	✓	–	–	✓	–	–	✓	–	
[22]	✓	–	–	✓	✓	–	–	–	
[23]	–	–	–	✓	✓	✓	–	–	
[24]	–	–	–	✓	–	–	–	–	
[25]	–	–	–	✓	–	–	–	–	
[26]	✓	–	–	✓	–	–	–	–	
This paper	✓	✓	✓	✓	✓	✓	✓	✓	

Literature review on data-driven methods

The broad adoption of SMs has drawn attention to data-driven CPI techniques that rely on inputs from the current SMI and transformer substation readings. Such data-driven techniques determine phase connectivity for LV end-customers using records of energy usage, voltage and current readings, or additional details supplied by SMs and the GIS platform. These data-driven CPI techniques divide into three core groups based on the chosen measurements: (i) voltage-based [3, 5, 6, 13, 15, 27-30], (ii) power-based [1, 2, 7, 8, 9, 12, 13, 17-26], and (iii) hybrid when integrating multiple data types [11, 14, 31-33].

Voltage-based methods require highly precise, time-aligned voltage data. Individual instruments may exhibit timing offsets, leading to recordings at mismatched times [8, 9, 18]. Besides, wide time spans make it hard to spot resemblances between customer and transformer patterns [8, 18]. On top of this, SMs do not support steady synchronized voltage collection at each LV customer because their role centers on checking service quality during defined incidents [8, 12, 18, 25]. Limited historical voltage records are commonly the biggest barrier when using voltage-based techniques [1, 9]. Certain hybrid methods require voltage and power readings [11, 14, 31, 33] or voltage and current readings from each end customer [32]. Still, the daily collection of voltage time-series is seldom conducted due to limited storage capacity. Therefore, the present study centers on power measurements and accordingly reviews the available power-based approaches.

Regarding power-based methods, multiple strategies are used in research that rely on active power readings. Techniques such as k-NN clustering and correlation analysis can match transformer and SM power patterns using various correlation metrics, including Pearson's and Kendall's rank correlation [1, 9]. Hosseini *et al*. [1] recommend an initial filtering phase for power patterns to notably increase the success rate of phase detection. Correlation checks might cover full-power pattern time series or target key consumption shifts through salient or frequency review [9, 12, 13]. Xu *et al*. [12] start by

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

performing a spectral review using the Discrete Fourier Transformation (DFT) to eliminate low-frequency components of load patterns, then extract features only from the high-frequency remnants that capture the key power shifts. Pappu *et al*. [2] and Jayadev *et al.* [20] employ graph theory to construct trees using Principal Component Analysis (PCA) applied to energy usage readings from SMs. Comparison between the PCA technique and Multivariate Linear Regression (MLR) appears in Santos [23]. Yi *et al*. [26] begin with an MLR model, then use stepwise regression to identify phases, and finally apply a significance adjustment factor to correct the outputs. Problems centered on optimization are commonly built around energy balance rules for active and reactive power needs per phase [7, 18, 19, 21]. Typically, Mixed Integer Programming (MIP) or Quadratic Programming (QP) is used, with defined simplifications, for clean or disturbed environments. LASSO-based data-driven strategies also emerge, focused chiefly on reducing the gap between the summed load patterns and the transformer pattern [13, 22]. Additional strategies draw on probability estimates via Bayesian inference [8], Grey Relational Analysis (GRA) based on shape similarity [24], and Wavelet analysis to extract features [25]. At last, Brint *et al*. [17] assign phase labels from variance found by contrasting actual substation currents to forecasted ones across clusters of SMs mixing phases.

An examination of current power-based CPI approaches is presented in **Table 1**, using particular judgment standards. Virtually all current methods operate without requiring upfront phase-label data to deliver solid outcomes. In cases where scalability checks are performed [1, 2, 7, 8, 12, 13, 18, 24], the suggested techniques show no decline in precision. Mixtures of single- and multi-phase loads or DERs might skew algorithm results, as noted in various papers [1, 2, 18-22, 26]. Along with those, noise extent or reading inaccuracy linked to SM accuracy class is frequently reported in assessments of CPI techniques [7, 8, 13, 18, 22, 23]. Unmetered usage stemming from absent meters or power theft may similarly alter model precision [9, 12, 13, 17, 18, 23]. In addition, only a small set of works run parameter studies to refine settings inside their specific process [18, 21]. Except for Kumar *et al*. [18], the remaining power-based strategies disregard DER penetration effects on their processes, while DERs under service and grid parameter uncertainty remain unaddressed across existing works. In general, the literature review stresses that no power-based CPI technique currently provides reliable, high-level results while covering all noted factors.

Approach and main contributions

This work takes an additional step. It proposes an updated CPI process rooted in Graph Neural Networks (GNNs) that leverages the electrical grid's representation, including its layout and element specifications, along with active power readings from end-customers and the secondary distribution transformer. As with additional ML strategies, the GNN model requires historical data for training. In this setup, synthetic transformer power pattern information is generated by running multiple power flow simulations using a Monte Carlo procedure. Concretely, during each power flow run, each recorded power pattern within the grid is assigned a random phase.

This proposed CPI process underwent checks covering every shortcoming identified in the literature and outlined in **Table 1.** Beyond that, if phase connectivity details are available, the GNN model trains solely on synthetic records, so erroneous labels have no sway on outcomes. Since it qualifies as an ML strategy, fine-tuning of chosen hyperparameters through proper parameter review is anticipated. Assessments of the GNN model's durability and expandability cover two LV grids with differing scales. Each grid requires its own GNN model trained to fit synthetic records, as grids may differ in layout, components, loads, and generators. Importantly, running Monte Carlo cases for each grid secures ample synthetic records, a key factor behind the GNN model's strong precision. As far as the authors know, no prior power-based CPI strategy has delivered durability alongside strong outcomes across the described settings.

The main contributions of this paper are as follows:

1. A GNN-based methodology is developed to identify phase connectivity in distribution grids. The methodology can effectively deal with various conditions of PV penetration, PVs under maintenance, measurement errors, unmetered consumption, uncertain grid asset parameters, namely, transformer and line parameters and inaccurate phase connection;

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

2. The scalability and performance of the methodology are evaluated for various grid sizes and with a mixture of single- and multi-phase DERs and loads;

3. The performance of the methodology is compared with other CPI approaches (namely, k-NN and LASSO) subject to considerable measurement errors or unmetered consumption.

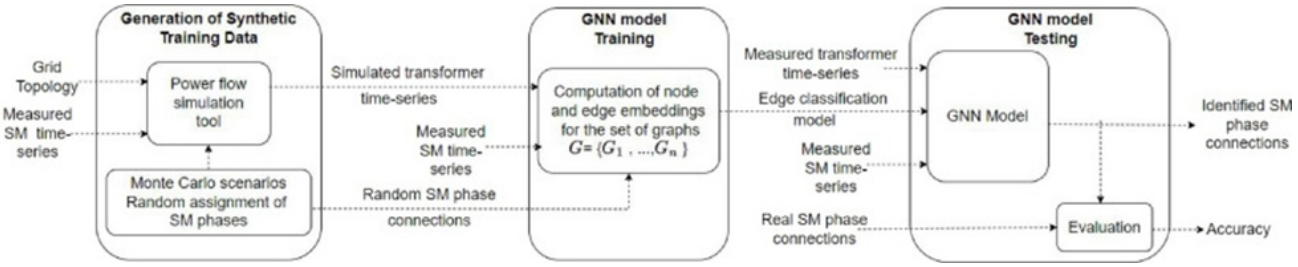


Figure 1. Proposed CPI methodology.

CPI Methodology

The CPI technique outlined in this paper comprises three core phases, as shown in **Figure 1**. The initial phase involves generating synthetic data to support model training. Following that, the GNN model is trained using an edge classification strategy. In the closing phase, the GNN model is evaluated using transformer power readings, and the resulting load-phase assignments are matched against known values.

Graph neural networks

GNNs are a category of deep learning techniques designed to understand graph relations by exchanging messages between graph nodes. They essentially expand standard neural networks to handle graph-structured data in a non-Euclidean setting. Lately, GNNs have gained popularity as a key approach for analyzing graphs, particularly in domains where information naturally organizes into graph structures [34]. Accordingly, the GNN structure aligns directly with the examined system's arrangement, allowing accurate insights into system components and their detailed mutual links. Power distribution grids mirror this arrangement in graphs with nodes (buses) and edges (lines) representing connections. Therefore, the use of GNNs in power systems is attracting increasing attention, spanning standard applications such as fault detection, time-series forecasting, power flow computation, and data generation [35].

A graph is typically expressed as G = {V, E}, where V denotes the collection of nodes and E stands for the collection of edges, or the links joining nodes. As noted in Hu *et al*. [36], three primary graph learning activities exist in research:

• Node-level tasks, centered around nodes, encompass node classification, node regression and node clustering.

• Edge-level tasks, aimed at labeling edge categories or forecasting links between two nodes, encompass edge classification and link prediction.

• Graph-level tasks, centered around acquiring graph representations, encompass graph classification, graph regression and matching.

For this work, the central aim is to establish how loads connect to grid phases, thereby determining the topology, which in turn enables applying the GNN edge-prediction approach [37]. Considering supervision aspects, training configurations for graph learning activities may be supervised using labeled data, unsupervised using unlabeled data, or semi-supervised with limited labeled nodes. In the present situation, all nodes and edges possess labels or values obtained from time-series data across both training and testing phases, leading to the adoption of supervised graph learning.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

Generation of synthetic training data

During the opening step, the necessary training information for the GNN model is created from (i) active power time-series readings collected from every SM throughout a specified timeframe and (ii) grid layout details, specifically electrical specifications for distribution transformers and lines essential for conducting power flow computations. In practice, most DSOs possess knowledge of these grid specifications as they perform associated evaluations, such as power flow and short-circuit computations. As a result, this paper presumes full awareness of the grid specifications. The pool of synthetic information needs to be robust enough to support strong outcomes, resilience, and flexibility in diverse grid environments for the GNN model. Creation of synthetic information follows a Monte Carlo procedure in which every case randomly allocates phases to all SMs connected to one particular distribution transformer. Within each Monte Carlo case, a three-phase power flow computation simulates the transformer time-series pattern based on the measured active power patterns from the SMs. Any suitable electric power system tool capable of handling three-phase power flow studies may be used. The number of Monte Carlo cases varies with grid scale and SM count. In conclusion, the synthetic training data production phase delivers (i) simulated power patterns of the transformers and (ii) preset random phases assigned to SMs. These outputs then proceed to the following phase of GNN model training.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

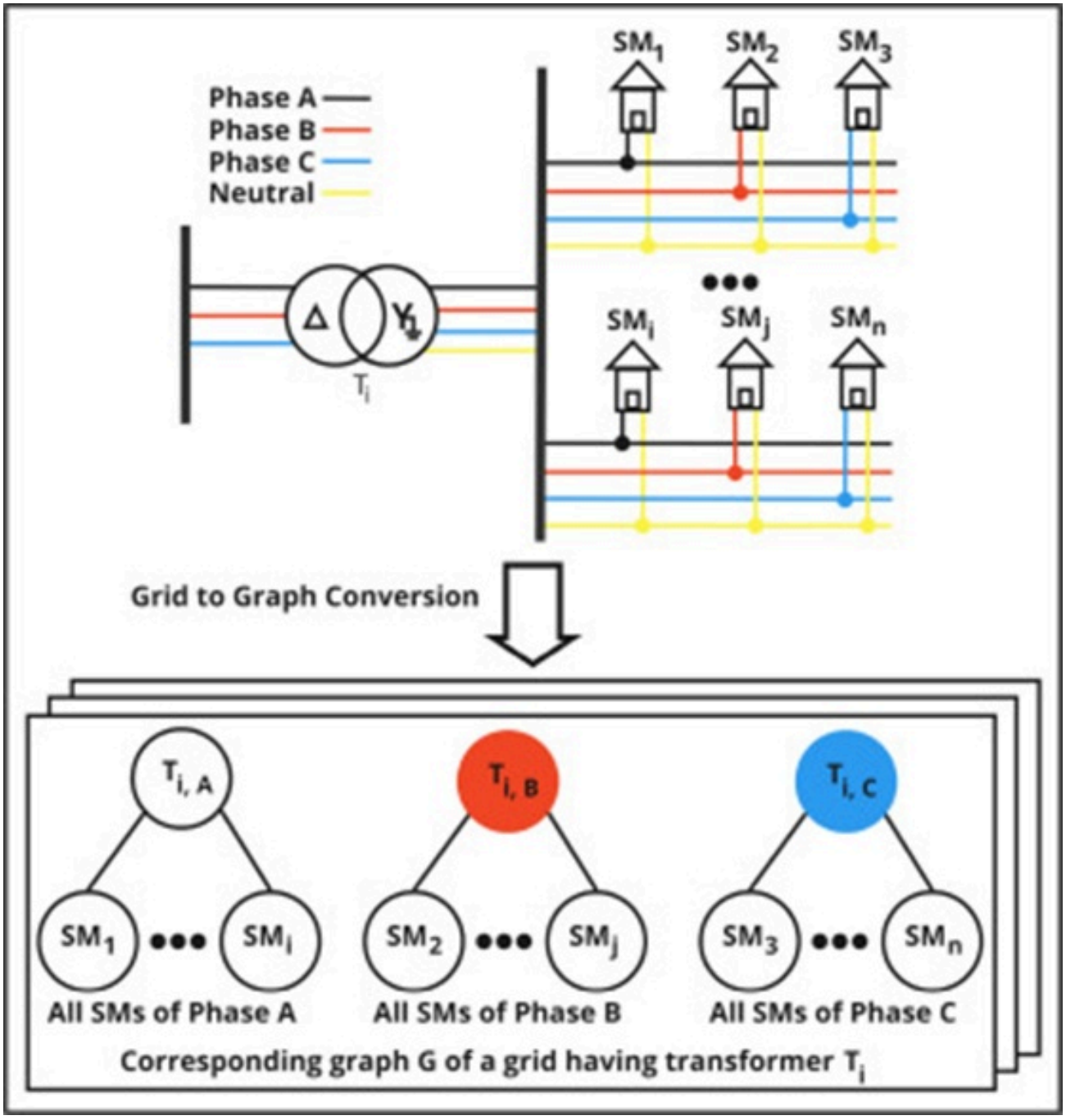


Figure 2. Conversion of distribution grid to a GNN graph.

GNN model training

The GNN model training phase draws on outputs from synthetic data generation (specifically the simulated transformer power pattern and random SM phase assignments) and on active power readings from all SMs. The objective is to structure the issue as an edge classification model, generating a set of N graphs from N Monte Carlo trials, represented as G = {G_1, G_2, …, G_N}. Each graph G_i= (, E) consists of nodes V and edges E, with every node carrying a collection of input features given by { h_v: $\forall\, v \in V$ }. As training advances, specific transformation operations process each node feature. After transformation, the resulting node and edge features are known as node embeddings and edge embeddings, respectively. Like standard neural networks, a GNN features an input layer, several hidden layers (with their

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

number termed "depth"), and an output layer where transformations occur to produce the final embeddings. The model handles graphs and outputs edge embeddings suitable for edge classification. Classification divides edges into two categories labeled "1" and "0"; specifically, an edge present between nodes receives an embedding of "1", while absence results in "0".

The representation of the distribution grid layout relies on graph-theoretic principles, matching grid buses to graph nodes and lines to graph edges, as depicted in **Figure 2**. Within the grid, each transformer is represented by three graph nodes, each corresponding to a unique phase. Additionally, every SM shown as a graph node is connected to the appropriate phase node on the transformer. An illustration in **Figure 2** demonstrates SMn linking to phase C, forming an edge between nodes Ti, C and SMn. Each graph node also includes the active power pattern belonging to the corresponding SM as its feature.

The transformation of the active power pattern for each SM starts with a linear projection that maps node features into a lower-dimensional space. Then, in line with Hamilton *et al*. [38], the GraphSAGE algorithm extracts node embeddings by aggregating information from a node and its neighbors. Embeddings for node v at hidden layer k, $\forall\ k \epsilon K$, shown through a collection of node embeddings { h_v^k: $\forall\ v \epsilon V$}, calculate as:

$$h_v^k = \sigma\left(W_{\text{self}}^k h_v^{k-1} + W_{\text{neigh}}^k h_{N(v)}^k + b^k\right) \quad (1)$$

$$h_{\mathcal{N}(v)}^k = \sum_k h_u^{k-1}, \quad \forall u \epsilon \mathcal{N}(v) \quad (2)$$

Where:

• σ is a non-linear activation function;

• W_{self}^k is a weight at layer k, and "self" indicates the weight applied to node v;

• W_{neigh}^k is a weight at layer k, and "neigh" indicates the weight applied to neighboring nodes;

• $h_{N(v)}^k$ are the features from the neighboring nodes;

• b^k is optional bias.

From (1)–(2), each node embedding within a neural network layer merges the node's own features from layer k - 1 together with features from neighboring nodes at layer k. The ability to extract knowledge from neighboring nodes, rather than relying solely on individual data, is a notable characteristic of GNNs. Upon attaining a defined depth K, once node embeddings reach certain neural network layers, edge embeddings emerge at the output layer through:

$$e_{ij} = \left(h_{v_i}^k\right) \cdot \left(h_{v_j}^k\right) \quad (3)$$

Where e_{ij} is an edge connected between v_i and v_j.. Generation of the edge embedding occurs via multiplication of the embeddings from the two nodes joined by that edge. This edge embedding then supports classification of edge presence between nodes. Through this process, the specific phase (Ti,A, Ti,B , or Ti,C) associated with each single-phase SM is identified. **Figure 3** presents the GNN architecture employed in this paper, including the input, hidden, and output layers.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

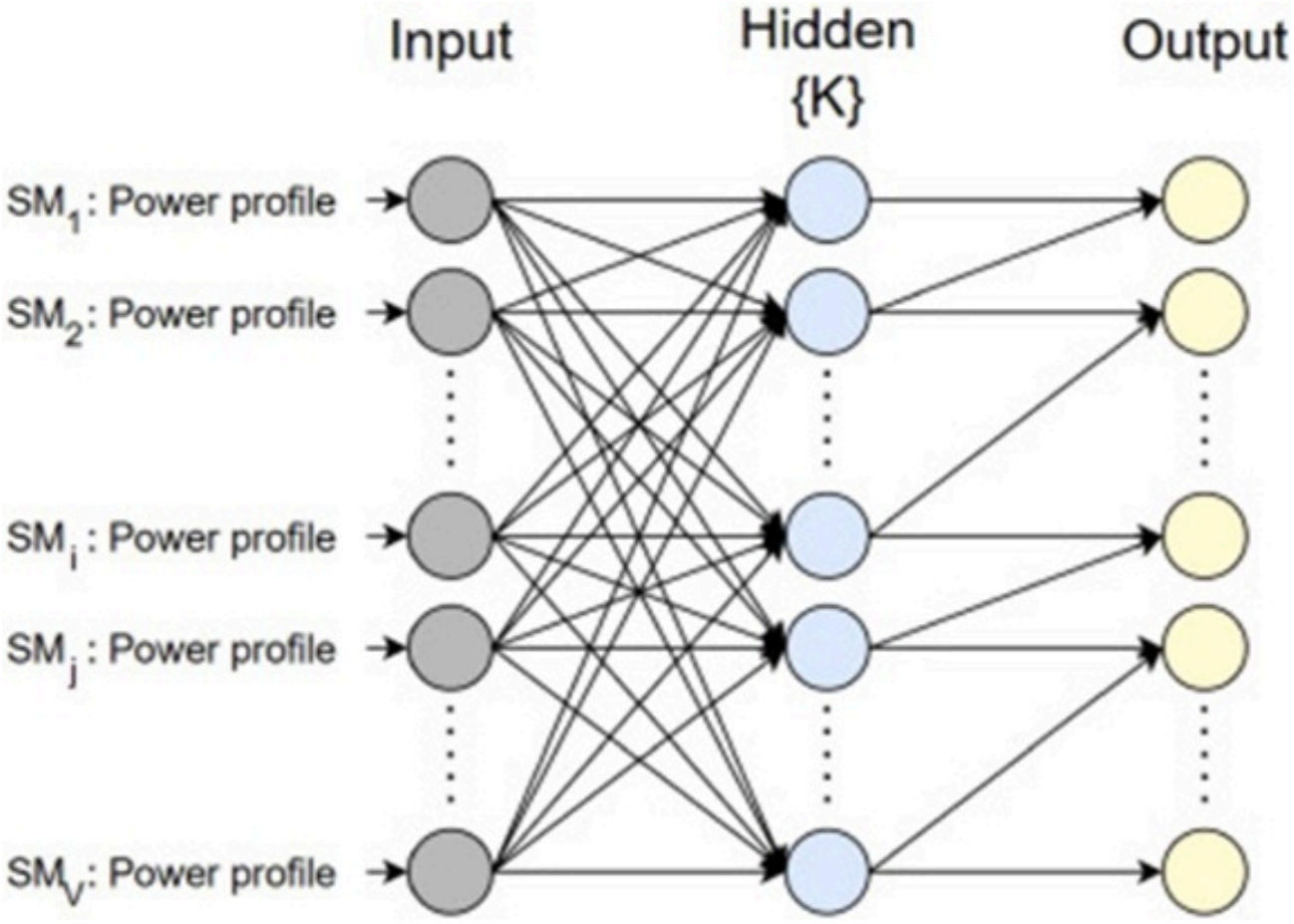
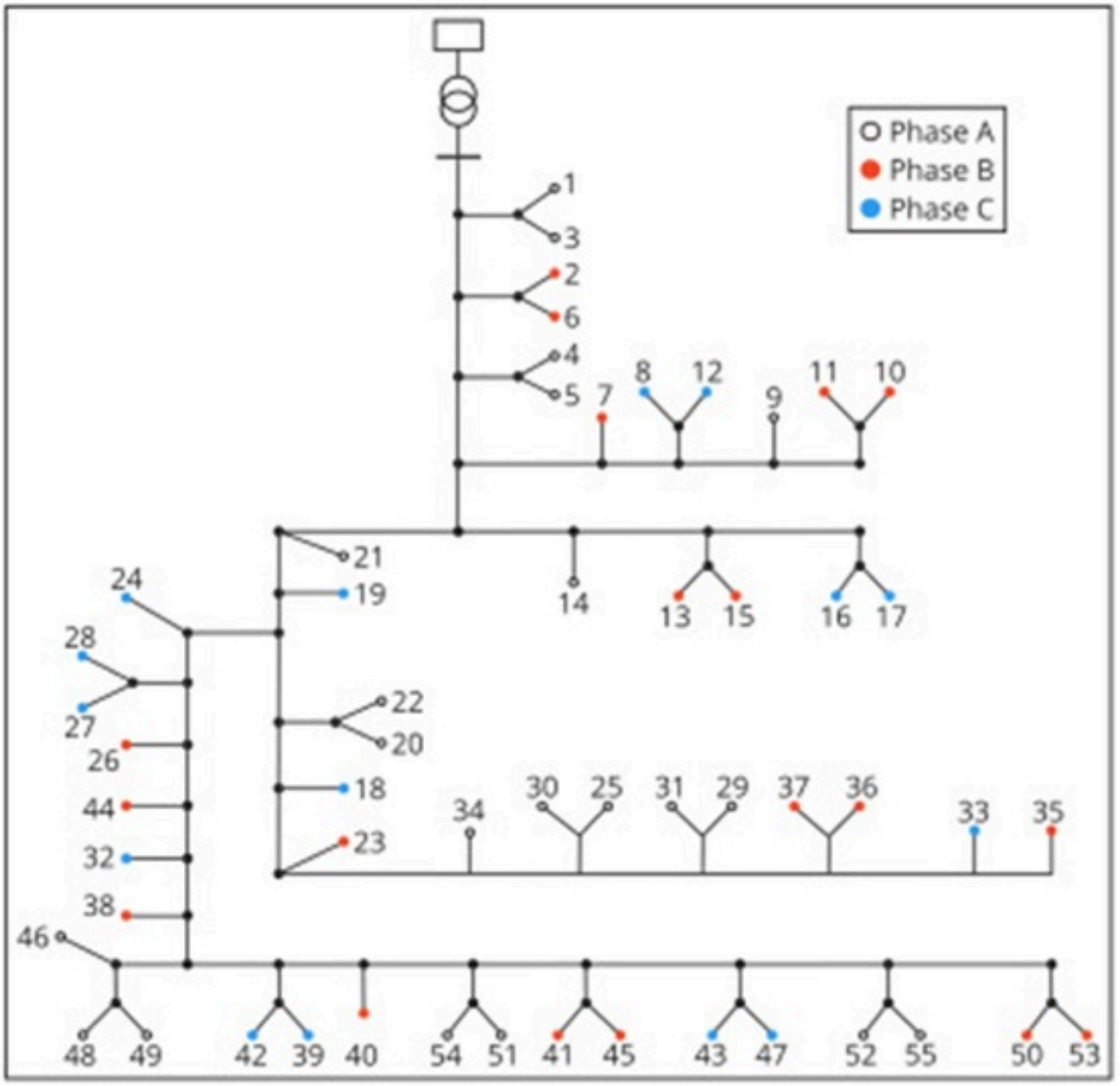


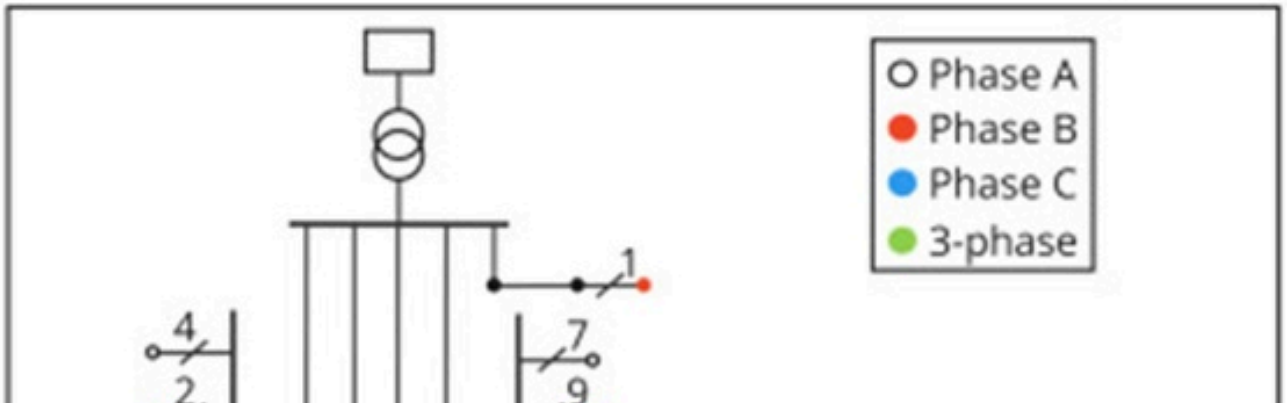
Figure 3. GNN model architecture adopted in this work.

Within each iteration, known as an epoch, calculations cover predictions for every possible edge embedding, followed by the computation of binary cross-entropy loss with respect to ground truth data. Ground truth assigns an edge embedding of "1" if an edge exists between two nodes, and "0" otherwise. The resulting loss undergoes backpropagation, adjusting the weights W to better align predictions with ground truth data in subsequent epochs. Predefined hyperparameters include the training/validation split, learning rate, maximum number of epochs, and batch size. Additional explanations regarding each hyper-parameter are available in Hamilton *et al*. [38].

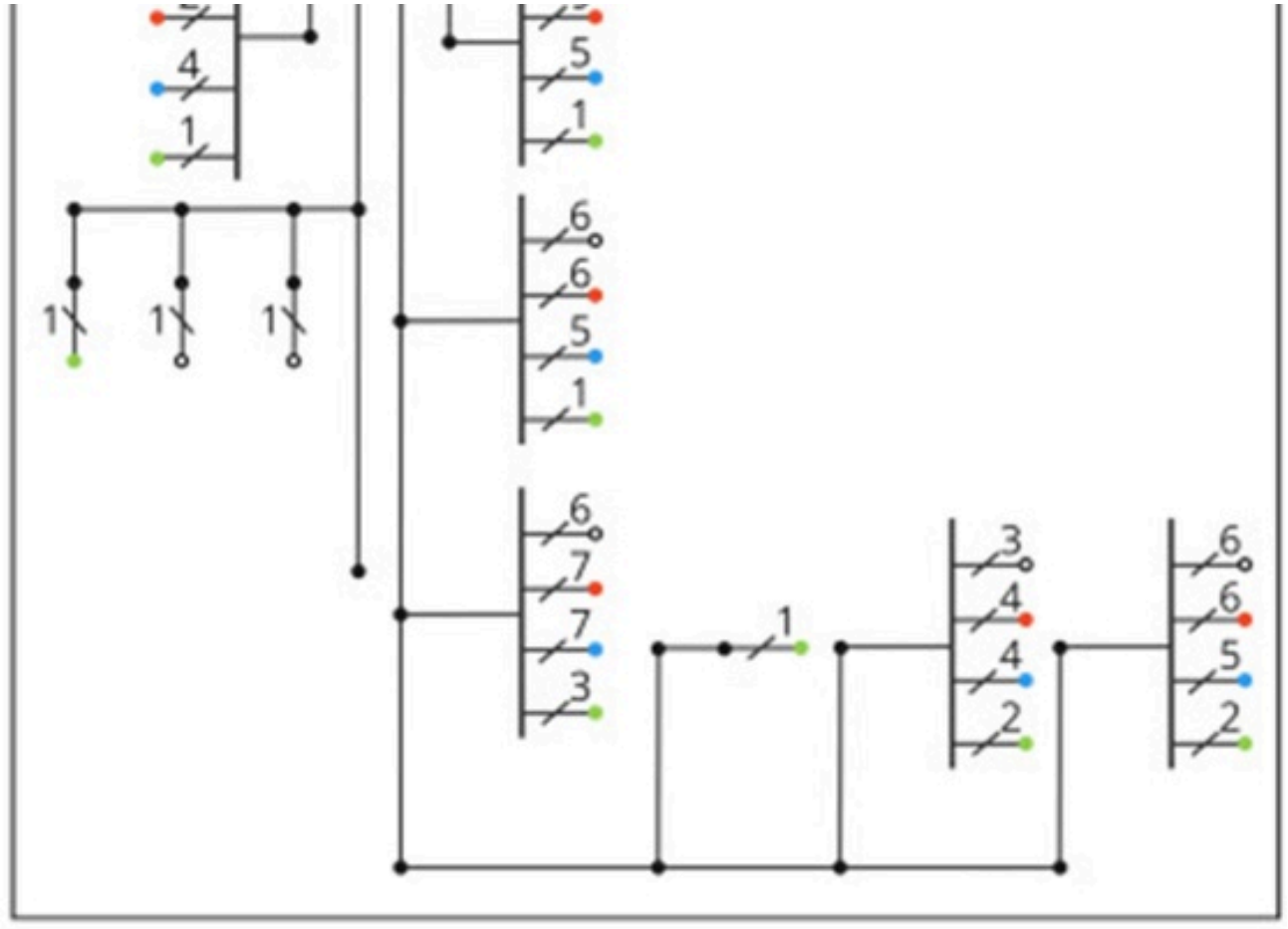
© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press



(a)

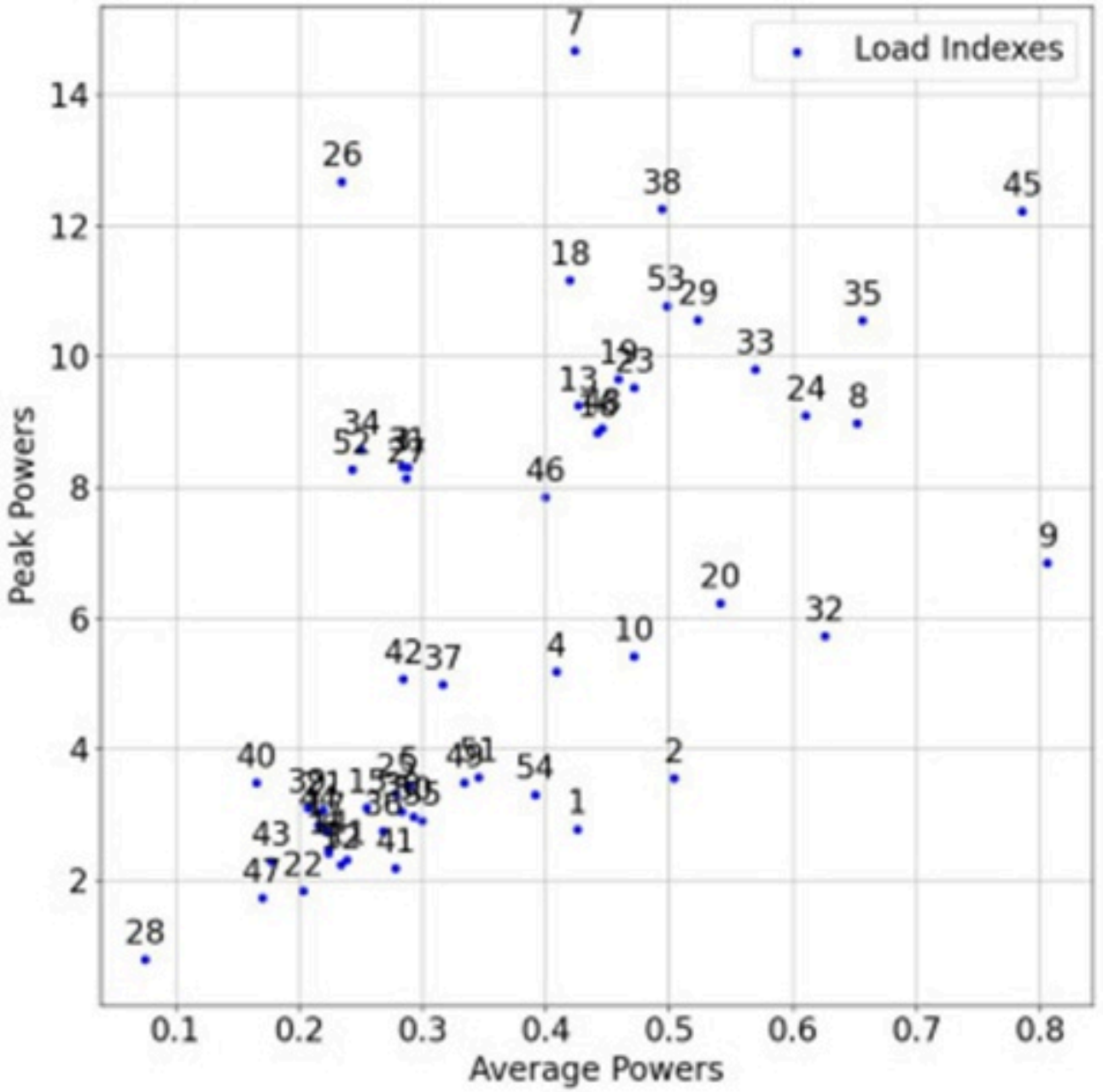


© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press



(b)

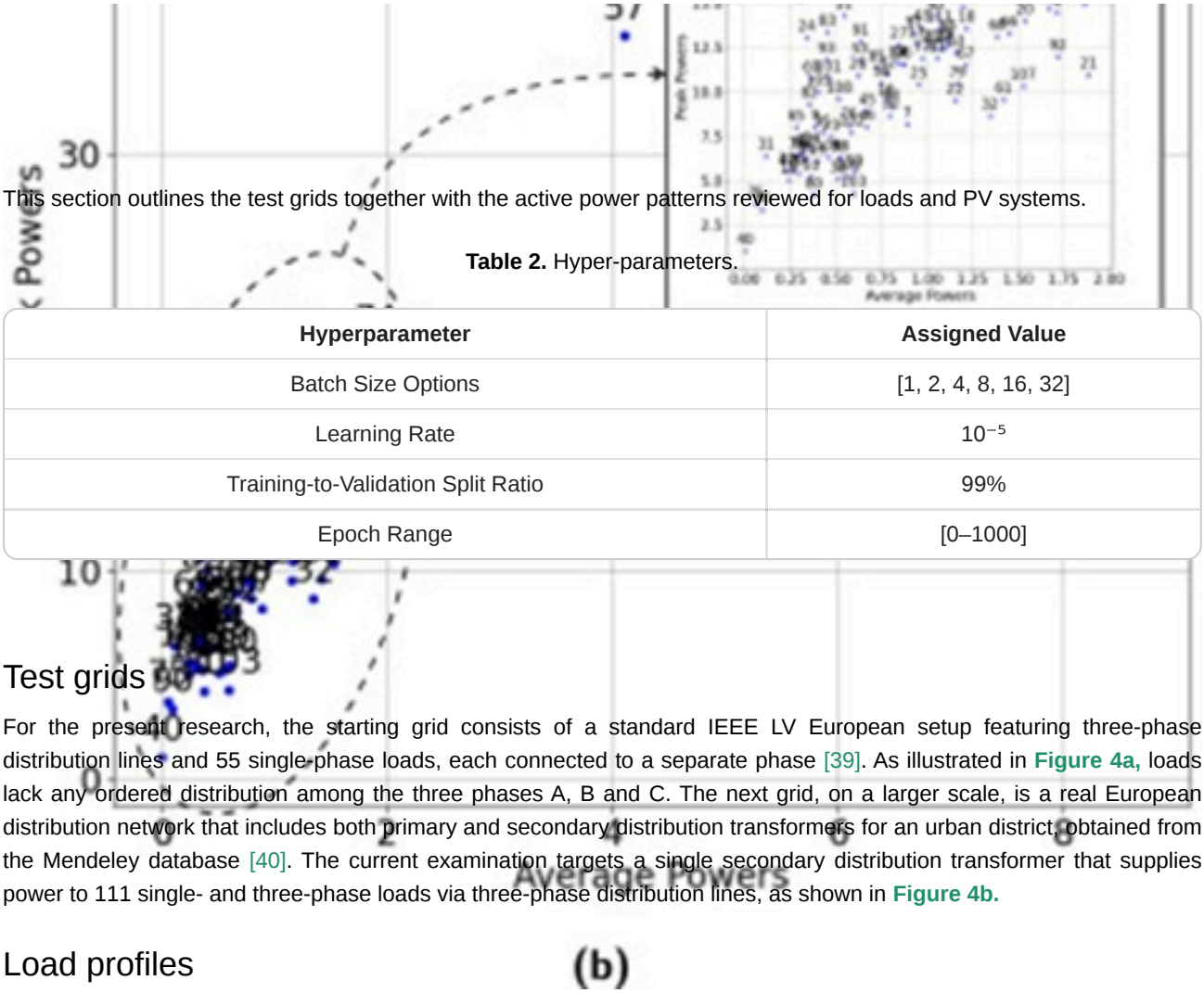
© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press



(a)



© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press



This section outlines the test grids together with the active power patterns reviewed for loads and PV systems.

Table 2. Hyper-parameters.

Hyperparameter	Assigned Value
Batch Size Options	[1, 2, 4, 8, 16, 32]
Learning Rate	10^{-5}
Training-to-Validation Split Ratio	99%
Epoch Range	[0–1000]

Test grids

For the present research, the starting grid consists of a standard IEEE LV European setup featuring three-phase distribution lines and 55 single-phase loads, each connected to a separate phase [39]. As illustrated in **Figure 4a,** loads lack any ordered distribution among the three phases A, B and C. The next grid, on a larger scale, is a real European distribution network that includes both primary and secondary distribution transformers for an urban district, obtained from the Mendeley database [40]. The current examination targets a single secondary distribution transformer that supplies power to 111 single- and three-phase loads via three-phase distribution lines, as shown in **Figure 4b.**

Load profiles

Two distinct datasets support the load patterns for the grids noted above.

The dataset associated with the IEEE LV European benchmark grid contains 55 daily active power patterns at one-minute resolution [39]. Consequently, every time series contains 1440 data points, and individual patterns link to each single-phase SM. To examine differences in power patterns, **Figure 5a** plots peak power for each pattern against average power. Obviously, average power ranges from 0.1 kW to 0.8 kW, while peak power ranges from 1 kW to 15 kW. The power patterns clearly display a range of features, demonstrating the dataset's diversity. This spread helps check the strength and capability of the CPI methodology over a wide variety of power patterns applied during load flow computations. Later studies might explore how heterogeneity influences GNN model performance by selecting patterns with varying levels of power fluctuation.

The dataset for the non-synthetic European LV Grid stems from the Irish CEER database and covers 6436 active power patterns collected from different customers at 30-minute resolution from 2009 to 2010 [41]. In this research, 111 randomly selected power patterns from the CEER database are selected before assignment to nodes carrying single- and three-phase loads. Power patterns for three-phase loads are split evenly over the three phases. Matching the first dataset from the IEEE LV grid, a time series containing 1440 data points from 30 consecutive days in 2009–2010 is assigned to loads within the second sample grid.

PV profiles

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

For PV patterns used on both test grids, arbitrary positions within the Swiss cities of Basel, Geneva, and Zurich are selected. Any end-customer in the grids might support a maximum of one single-phase PV unit set at 180 degrees azimuth and 35 degrees tilt. At the same time, total system losses equal 10% of the momentary output. Because the studied grids have no PVs, the active power patterns in summer and winter are supplied by Renewables. Ninja web application [42]. All patterns assume a 10 kW rated power and a time series of 1440 data points. Even though 10 kW PVs are normally connected in three-phase, the analysis assumes single-phase connection to examine the most challenging scenarios involving single-phase PV addition. Also, the PV meter captures only generated power, avoiding mixing with household usage. Summer and winter patterns are assigned to nodes equipped with PV systems to reflect changes in PV power generation.

Hardware and software attributes

Hardware consists of an Intel i7 processor with 16 GB RAM and an NVIDIA GeForce RTX 2070 to manage every phase of the CPI methodology, from synthetic data generation to GNN model training to testing. Data preprocessing is performed using MATLAB 2024a [43], and the OpenDSS 10.0.2 open-source simulator performs three-phase power flow simulations [44]. Synthetic training data is generated using Python 3.12.0 [45] for smooth operation with OpenDSS 10.0.2. Torch geometric Python library handles the GNN model. Moreover, Python 3.11 is required to run the GNN model, as the ML library (PyTorch) remains incompatible with Python 3.12 for now.

Regarding hyperparameters, several values for batch size and epoch count were tested to achieve the best fit, as shown in **Table 2**. Final decisions also fix the learning rate and training/validation ratio.

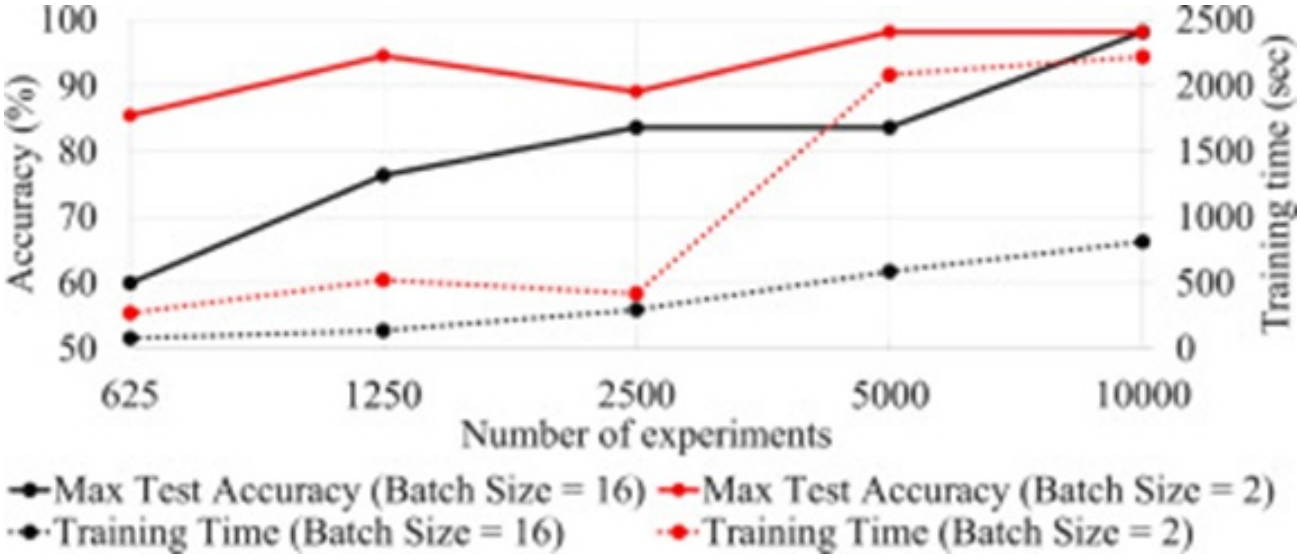


Figure 6. GNN model accuracy and training time with respect to the number of experiments and different batch sizes for the IEEE LV grid.

Results and Discussion

This section presents and explains findings from multiple examinations, then discusses the limitations of the GNN model and possible future directions.

Table 3. GNN model accuracy and training time with respect to the batch size for the IEEE LV grid.

Batch size	Training time (s)	Test accuracy (%)

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

1	4534	98.2
2	2219	98.2
4	2382	98.2
8	1380	96.4
16	813	98.2
32	690	92.7

Parametric analysis

To begin, review the varied batch sizes and the counts of Monte Carlo trials. **Figure 6**, combined with **Table 3,** shows that the top precision of 98% occurs at a batch size of 2, provided a minimum of 5000 trials are performed. Increased batch sizes reduce the number of graphs available for training and therefore require additional trials before reaching the mentioned precision level. The chief gain from larger batch sizes is reduced training time, balanced against the extra Monte Carlo runs needed to maintain equal precision. Raising the batch size does not guarantee superior test precision, even with reduced training duration, as shown in **Table 3.** Specifically, batch sizes 8 and 32 yield lower precision (96.4% and 92.7%, respectively) than sizes 1-4. Nonetheless, no obvious downward trend is observed, with 98% precision at a batch size of 16.

Regarding the smallest epoch count needed for peak precision, multiple figures are being checked within the IEEE LV grid. Shown in **Figure 7**, precision of 98% arrives at 80 or 90 epochs, although brief ups and downs appear at midway values. Because precision levels off after 50 or more epochs, this number is adopted for the IEEE LV grid. Given the larger size of the non-synthetic European LV grid, 300 epochs are selected using the same logic.

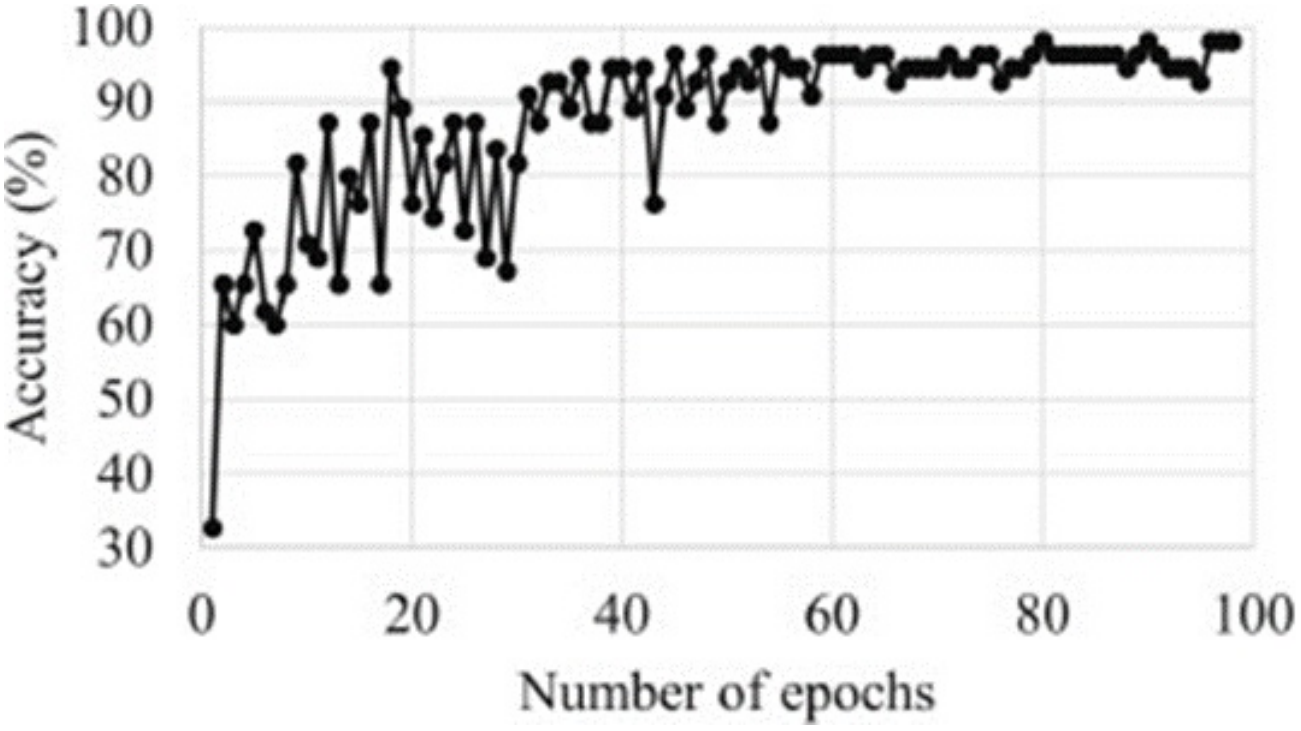


Figure 7. GNN model accuracy with respect to the number of epochs for the IEEE LV grid.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

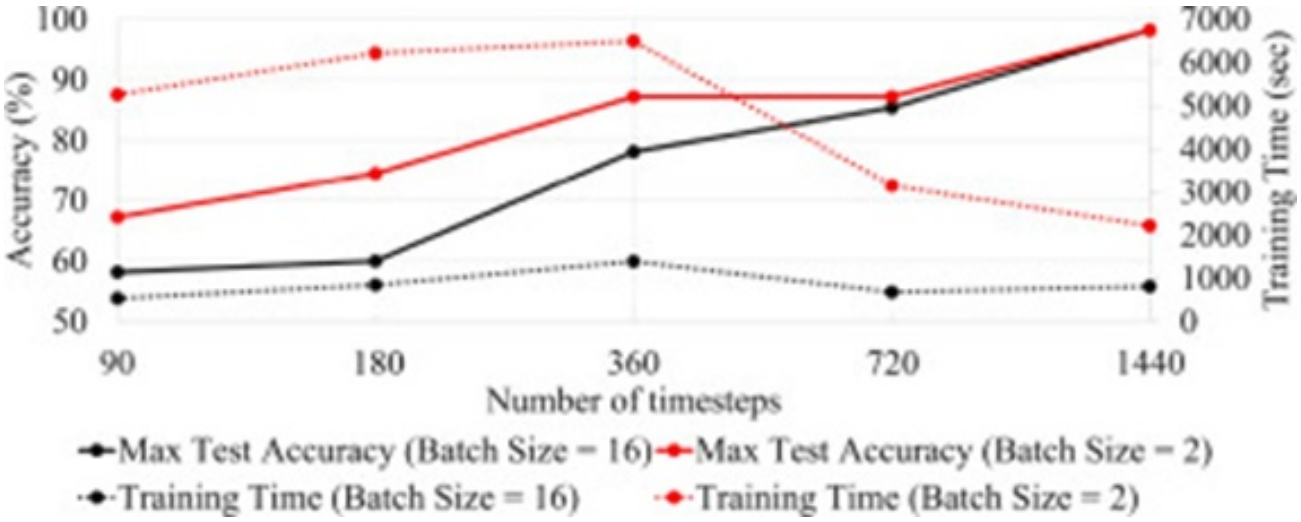


Figure 8. GNN model accuracy and training time with respect to the number of timesteps for the IEEE LV grid.

Temporal resolution within time-series patterns leaves the GNN model untouched, since each timestamp power value stands alone with no time-based connections. Therefore, timestamp length (e.g., 15 min or 30 min) does not affect the GNN model's performance. On the other hand, the timestep quantity can strongly shape GNN model precision, as evident in **Figure 8**. Within the IEEE LV grid, achieving 98% precision clearly requires 1440 timesteps regardless of the batch size. Besides, a batch size of 2 paired with 1440 timesteps results in a notable reduction in training duration compared to lower timestep counts. As mentioned before, larger batch sizes tend toward briefer training periods, yet such gains remain bounded as larger timesteps are adopted.

Adopting a batch size of 2, 1440 timesteps, and 300 epochs enables a precision review across different Monte Carlo trials on the non-synthetic European LV grid. The primary intention here is to find the required trial quantity, scaled to the grid dimensions, so that precision remains above 90%. As shown in **Table 4**, at least 12,000 trials are required to achieve 90% test precision on the non-synthetic grid with 111 loads. The IEEE LV grid with 55 loads attains roughly 98% precision after 5000 trials. Both situations suggest that Monte Carlo trial numbers remain proportional to the SM count multiplied by a factor of 120-150.

Table 4. GNN model accuracy and training time with respect to the number of experiments for the non-synthetic European LV grid.

Number of experiments	Test accuracy (%)
10 000	85
12 000	90

Evidently, Monte Carlo trials do not require matching every possible phase combination among the SMs. For the IEEE grid containing 55 end-customers, there are 355 total combinations. Still, 5000 trials prove enough for 98% precision. In the LV grid of 111 end-customers, despite 3^{110} possible combinations, 12,000 trials produce 90% precision. Trials can also include repeated combinations, since random phase choice occurs at each grid node independently of other trials. Hence, the GNN achieves strong performance with only a fraction of the total possible combinations. This strength enhances GNN's suitability for real-world applications where end-customer demand rises sharply.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

Another verification step in this review assesses how grid losses affect the GNN model's results for the IEEE LV grid. Thus, Monte Carlo trials follow two separate approaches for expressing power at transformer LV terminals:

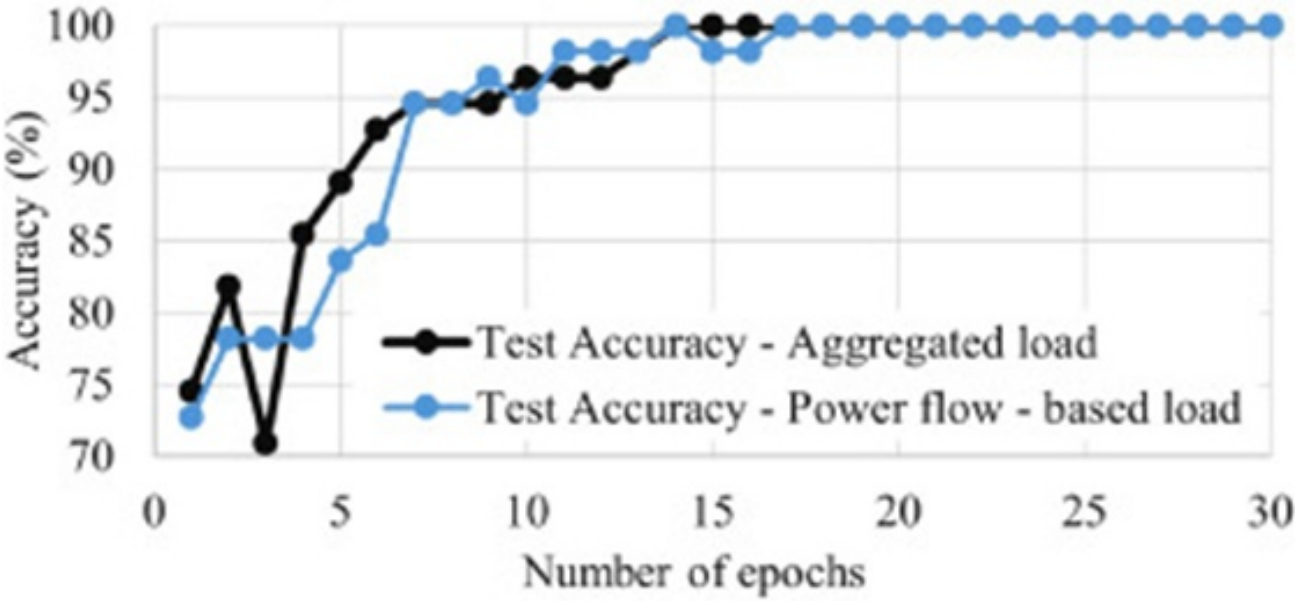


Figure 9. Comparison of GNN model accuracies with aggregated and power-based load for the IEEE LV grid.

• Aggregating the individual profiles measured by all SMs for the determination of the transformer profile, ignoring the grid losses;

• Running power flows for the simulation of transformer profile (which considers the grid losses).

Training of the GNN model is performed over multiple epochs under both conditions, while accuracy is compared in **Figure 9.** Results show that after 17 epochs, both methods achieve 100% accuracy. Grid losses therefore hold almost no influence over GNN model results and can safely be dropped out. This situation eliminates the need for load flow computations, easing the production of synthetic training data and sharply reducing preparation time. Parallel checks that add multi-phase loads await future efforts.

Table 5. GNN model comparison with state-of-the-art methods in case of noisy measurements for the IEEE LV grid.

Method	**Test accuracy (%)**	
Empty Cell	**With noise**	**No noise**
GNN	95	100
LASSO	80	100
k-NN	68	82

Measurement errors

In real-world settings, SMs, depending on their accuracy class, may introduce measurement inaccuracies reaching ±2% under disturbed environments [16]. Here, Gaussian noise at a 5% rate is added to the load and transformer power

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

patterns of the IEEE LV grid to examine the GNN model's performance under elevated noise. Afterwards, the GNN model trains on these noisy patterns, yet noiseless patterns are used for testing. The precision of the GNN model is compared with that of the LASSO and k-nearest neighbors (k-NN) algorithms, which are commonly used in phase detection studies. As shown in **Table 5**, the LASSO algorithm yields results similar to those of the GNN model under standard conditions without noise addition.

In contrast, the k-NN algorithm achieves the lowest precision. Still, the GNN model proves more resilient than LASSO and k-NN algorithms under intense noise. Specifically, the GNN model maintains roughly 95% precision, while LASSO declines to around 20% and k-NN reaches only 68%.

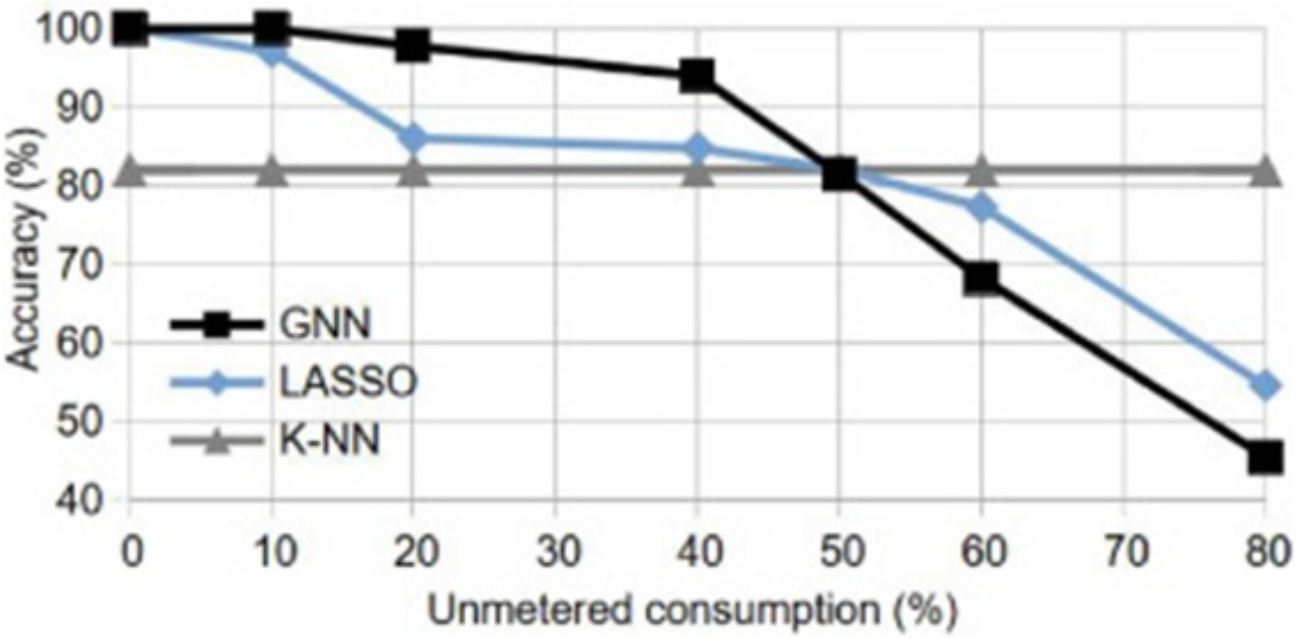


Figure 10. GNN model accuracy with respect to different levels of unmetered consumption.

Unmetered consumption

Regarding unmetered consumption, an initial 10% share applies to the IEEE LV grid, meaning that 5 SM patterns are considered unknown. Consequently, the remaining 50 measured power patterns participate in Monte Carlo trials during synthetic training data creation. During GNN model testing, the measured transformer pattern also accounts for the unmeasured SM patterns. Yet the GNN model pinpoints phases solely for the 50 measured single-phase SMs. When using the LASSO algorithm, the 10% unmetered consumption case is repeated 100 times across random selections of unmetered loads, after which the minimum precision across all runs is reported. As shown in **Figure 10**, the GNN model outperforms both LASSO and k-NN, even though LASSO achieves a solid 97% precision. In addition, the GNN model demonstrates consistent performance with no precision loss when shifting from fully metered to 10% unmetered conditions. Subsequently, the GNN model, LASSO, and k-NN algorithms are assessed across increasing levels of unmetered consumption. **Figure 10** shows that the GNN model's precision declines steadily from 98% to 45% at unmetered consumption levels of 20% and 80%. Nonetheless, the GNN model achieves steadier outcomes than LASSO up to 40% unmetered consumption, with only about a 5% reduction in precision. By contrast, LASSO shows a sharp drop of more than 15% in precision as unmetered consumption rises to 40%. Moreover, LASSO follows a similar pattern, exceeding the GNN model's performance beyond 50% unmetered consumption. The k-NN algorithm maintains 82% precision regardless of the amount of unmetered consumption. This steady behavior stems from k-NN determining unknown phase links by correlating individual SM power patterns with the transformer pattern, independent of other SM

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

readings. According to the assessment, the GNN model clearly exceeds leading techniques for unmetered consumption, achieving up to 50%.

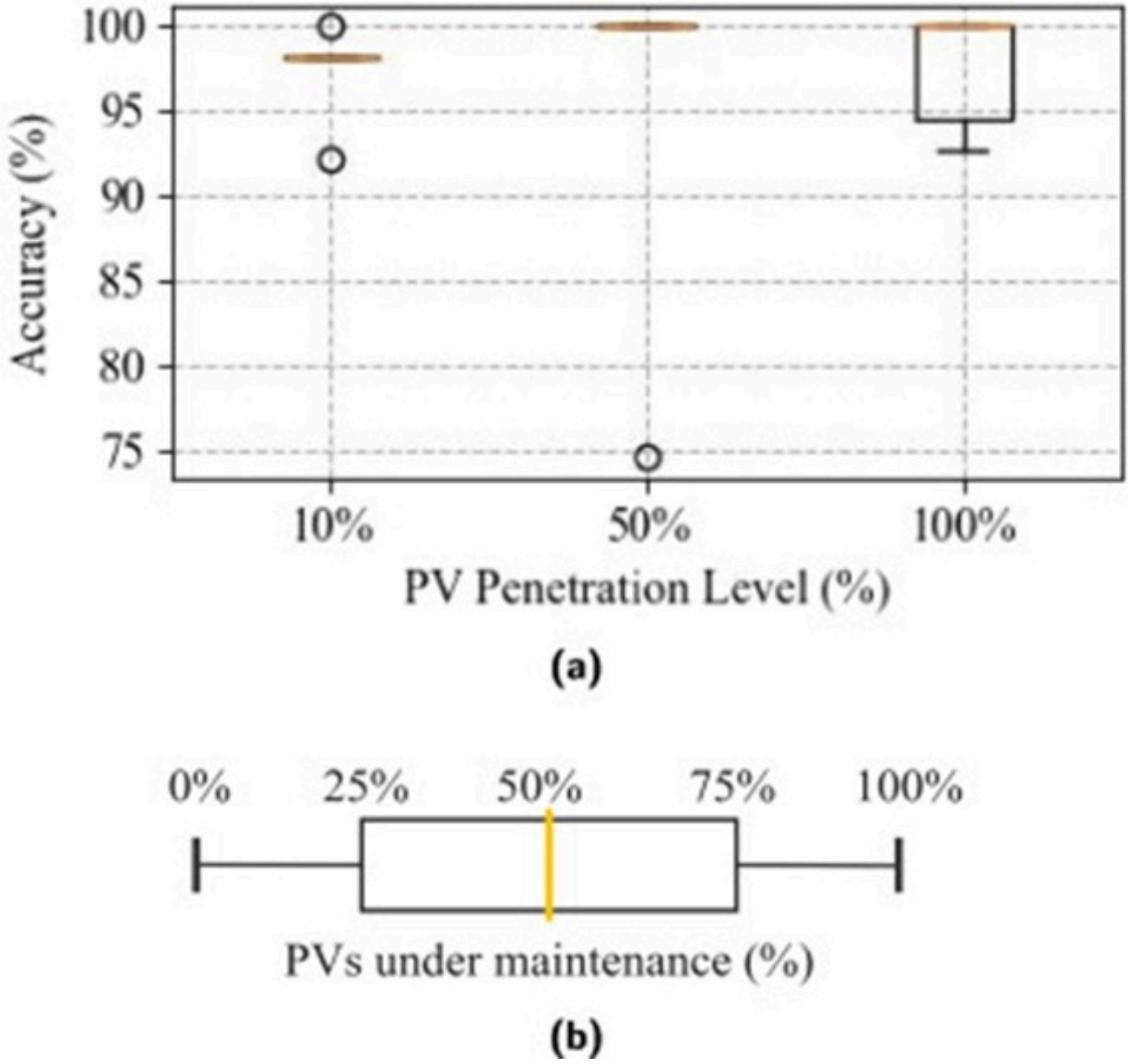


Figure 11. GNN model accuracy: (a) for different PV penetration levels (%) (b) based on a boxplot of quartiles for PVs under maintenance.

PV sensitivity analysis

Initially, this review considers three PV penetration rates: 10%, 50%, and 100%. For each PV penetration rate, varying proportions of PVs are considered under maintenance during testing of the GNN model. PVs stay disconnected during maintenance periods. Since the DSO may lack full awareness of PVs under maintenance, these are included in the GNN model training. Their power patterns are absent from the transformer pattern during GNN model testing, even though they appear in Monte Carlo trials during GNN model training. Notably, PVs under maintenance are not included when executing power flows to compute the simulated transformer pattern during the testing phase. Every PV scenario is simulated on the IEEE LV grid for both the training and testing phases, as no genuine measurements are available. Still,

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

the random phase mixing of PVs used for simulation during testing is omitted from the Monte Carlo trials used for training, hindering the GNN's proper performance.

Shares of 0%, 25%, 50%, 75%, and 100% apply to PVs under maintenance drawn from the primary five quartiles. In **Figure 11**, a box plot is shown for each PV penetration rate, based on the specified quartiles of PVs under maintenance. Outcomes indicate precision remains over 90% across all situations, except for the case of 50% PV integration with 100% maintained units. Here, greater precision could be achieved by extending GNN model training, but this is avoided to maintain comparable computation time across test cases. Consequently, testing precision falls to 75%. However, remaining cases at 50% penetration yield 100% precision. For PV integration rates of 10% and 100%, as more PVs enter maintenance, precision drops to roughly 92%. In total, the GNN model exhibits durability and strong performance amid increasing PV integration rates, including when substantial PV shares are accounted for in training but remain under maintenance in testing.

In the next step of this review, additional loads and PVs are added to the IEEE LV grid beyond the base loads. Specifically, the situation includes not only single-phase loads but also two- and three-phase loads and PVs. The purpose is to determine whether multi-phase loads and PVs affect the GNN model's performance. As in the initial review, random phase mixing for multi-phase loads and PVs is only for simulation. Thus, SM and transformer patterns for testing rely on load-flow simulations rather than actual measurements. Besides, random phase mixes for phase links examined for SMs and PVs in Monte Carlo trials are excluded in the testing phase to support suitable GNN model assessment. Under this framework, three arbitrary nodes are picked for installation:

• Single-phase units: Two loads and two PVs connected to the first node

• Two-phase units: One load and two PVs connected to the second node

• Three-phase units: One load and two PVs connected to the third node

To generate synthetic data for GNN model training, 10,000 Monte Carlo trials are performed. For testing, precision reaches 91.80%, revealing that multi-phase units may impact model outcomes. Nonetheless, precision remains high at elevated levels, indicating the GNN model's durability when facing mixtures of multi-phase loads and PVs.

The conclusion states that the GNN achieves high precision in detecting links for single-, two-, and three-phase units. The GNN trains according to the pattern of each phase presumed inside Monte Carlo trials, while the GNN outputs assign every pattern to one phase. Thus, for two- and three-phase units, the separate power pattern of each phase is assigned to the phase identified by GNN. This way, the GNN determines not only the phases of a multi-phase unit but also the exact power pattern for each phase. If two single-phase end-customers sharing matching power patterns are linked to two phases at a single grid node, the GNN can accurately identify those two phases. Yet pinpointing which end customer corresponds to each phase remains impossible because time-series match exactly, and the model does not use further customer details, such as identity numbers. In practice, linking two- or three-phase units with matching power patterns per phase, for instance PVs, electric vehicles and motors, occurs regularly, whereas connecting single-phase end-customers to separate phases at a single node with matching power patterns per phase happens rarely. Beyond this, the GNN is trained on power patterns spanning 1440 timesteps. As a result, the power patterns of single-phase end-customers connected to different phases at a single node should exhibit differences that enable the GNN model to operate correctly.

Uncertain parameters of grid assets

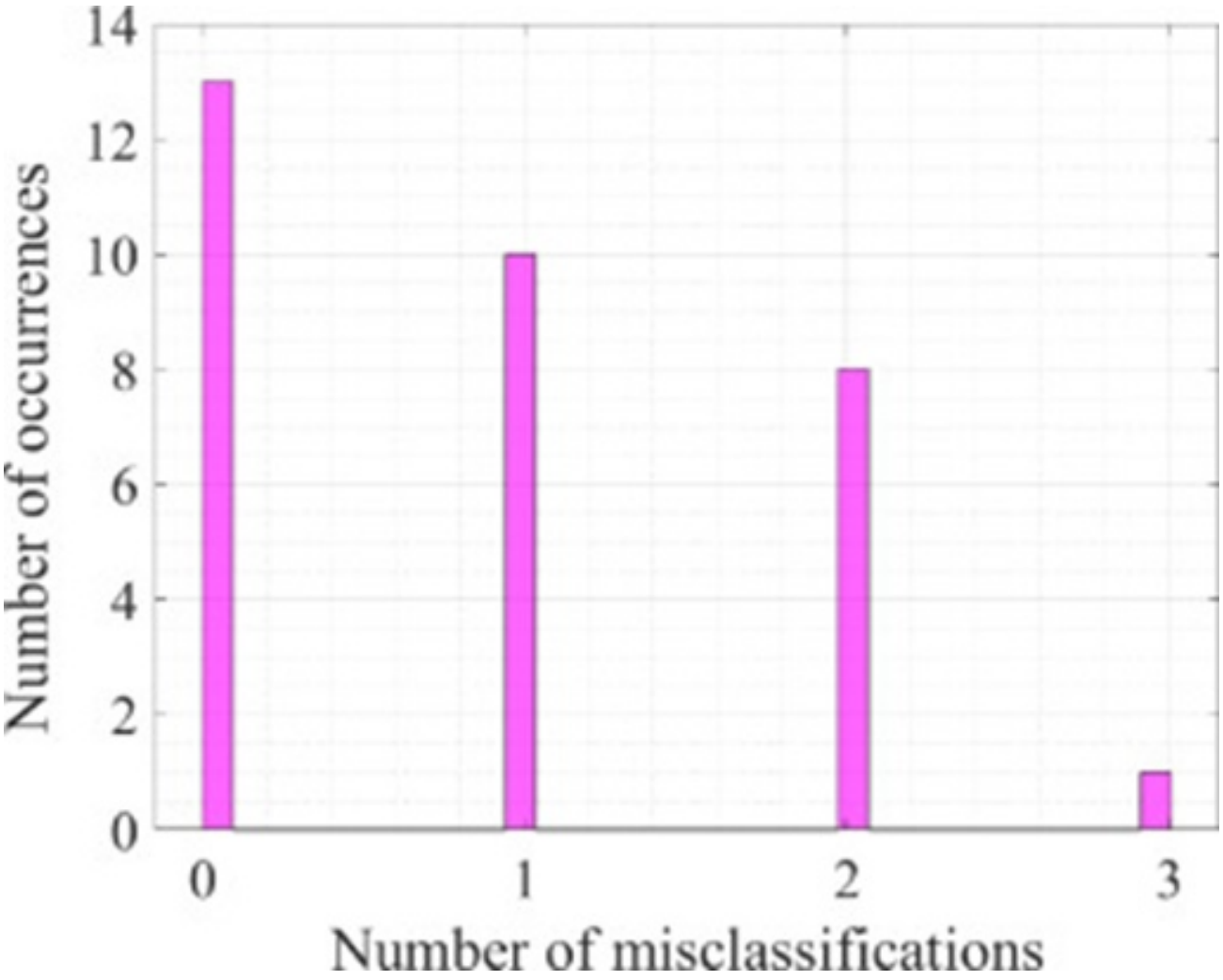
Complete knowledge of the grid specifications is assumed when generating synthetic training data via repeated power-flow simulations. Given that some tolerance may apply to the grid specifications, this section evaluates the influence of uncertainty in line resistance and inductance values on GNN outcomes for the IEEE LV grid.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

In this framework, resistance and inductance are treated as random variables with a uniform probability density function (PDF) spanning [–10%, +10 %] around their nominal values. Up to 32 separate mixes of deviations (in) for resistance and inductance emerge from the joint PDF through low-discrepancy quasi-random Sobol's sequences that deliver strong space coverage with limited samples [46, 47]. **Table 6** presents a sample portion of the generated matrix.

Table 6. Excerpt of the matrix containing the 32 combinations of % deviations of resistance and inductance.

Combination ID	Inductance deviation (%)	Resistance deviation (%)
1	6.21	−6.39
2	−9.66	8.39
3	−1.51	−2.88
...	...	...
32	−7.69	4.76



© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

Figure 12. Impact of the uncertain line parameters on the GNN model performance in terms of number of misclassified load phases.

Examination of GNN model outcomes under uncertainty in line specifications occurs on the IEEE LV grid. For each mix, the suggested process first applies the deviated resistance and inductance values across all distribution grid lines. Predicted phases are then calculated for each load before comparison with the actual phases. The count of misclassifications is finally logged and serves as a measure of the GNN model's performance.

As shown in **Figure 12**, roughly 40% of cases allow the GNN model to correctly forecast phases for all 55 loads. The largest misclassification count is 3, observed in only one of the 32 situations. Notably, negative shifts of −7.1% and −9.1% for line resistance and inductance, respectively, cause three misclassifications among the 55 load phases. Apart from that, mixes producing large shifts in resistance and inductance result in as many as three wrong phase links. Outcomes indicate that uncertain grid specifications have a limited effect on GNN model precision. From an engineering standpoint, resistances and inductances shape grid losses that build up inside the transformer power pattern. Yet, complete disregard of grid losses—in other words, resistances and inductances—leaves GNN model outcomes nearly unchanged, driven primarily by load patterns.

To explore the effects of extreme uncertainties on GNN outcomes, two additional situations undergo review for the IEEE LV grid: (a) ±10% shifts on resistance (R) and inductance (X) across all lines, and (b) ±20% shifts on resistance and inductance across all lines plus transformer impedance. **Table 7** summarizes outcomes from the noted situations. Evidently, the peak misclassification count under extreme uncertainties with the GNN stays between 1 and 2. In summary, the GNN model demonstrates robustness to realistic or even extreme uncertainties in grid asset specifications.

Table 7. Number of misclassifications in case of extreme deviations on resistances and inductances of lines and transformer.

Deviations (%)				
Transformer		**Line**		**Misclassifications**
X	**R**	**X**	**R**	
0	0	10	10	1
0	0	−10	−10	1
20	20	20	20	1
−20	−20	−20	20	2

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

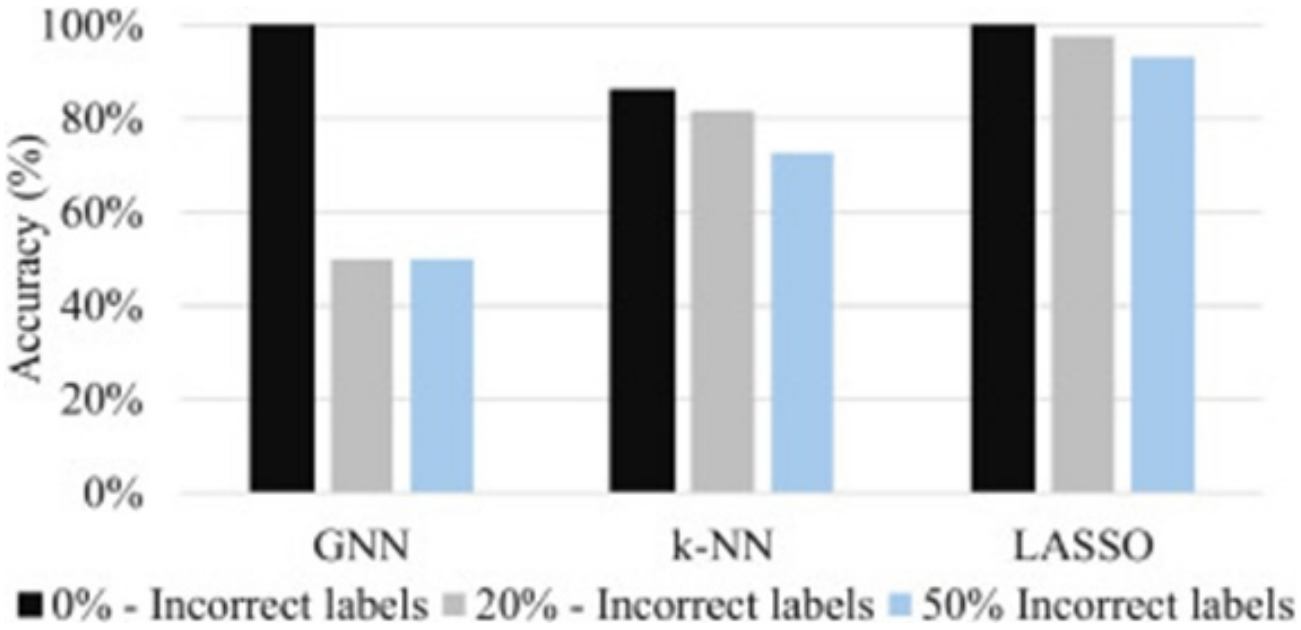


Figure 13. Accuracy of GNN model and state-of-the-art methods in case of incorrect known phase connections.

Inaccurate phase connections

In practice, certain phase links may be known and accepted when generating synthetic training data. Still, incorrectly labeled links can harm GNN model performance. Under this setting, an additional sensitivity review is conducted across varying levels of inaccurate phase links for the IEEE LV grid. With the assumption that 20% of phase labels count as known, three extents of wrong phase labels apply: (i) 0%, (ii) 20%, and (iii) 50%. Known phase labels stay fixed throughout Monte Carlo trials, so random phases are assigned only to the remaining unknown phase links.

As shown in **Figure 13**, the GNN model's precision falls to 50% when 20% or 50% of the phase labels are incorrect. A comparison of GNN model outcomes with k-NN and LASSO reveals that both algorithms outperform the GNN model. Specifically, precision for LASSO and k-NN algorithms ranges from 70% to 94%, indicating that incorrect labels can weaken GNN model performance. Nonetheless, this shortcoming carries no consequences for GNN model creation, as only the grid layout and power patterns are essential, rendering known phase links unnecessary. Owing to the danger of lowering GNN precision from inaccurate phase links, known phase links should stay excluded from Monte Carlo trials. Since GNN secures nearly 100% precision without known phase links, GNN outputs can verify the correctness of known labels. Thus, a GNN can spot erroneous labels, provided they are not used during training.

Wrong-phase links do not affect the k-NN algorithm's precision, as the model is trained by assessing correlation for each link separately without relying on known incorrect labels. For the LASSO algorithm, precision stays above the rest at values exceeding 90% across all extents of wrong-phase links. However, outcomes are tied to a specific mix of known phase labels, whereas LASSO outcomes might weaken across multiple random mixes of known links.

Discussion

The uncertainty of line resistance and inductance uses a standard uniform PDF that supports the creation of typical mixes. A parallel method could apply to the random mix of phase labels presumed for Monte Carlo trials. This approach would markedly reduce the required Monte Carlo trial count by selecting only typical mixes within the study range. Regarding GNN model outcomes, this could also raise precision when large trial counts become necessary due to extensive grid scale. The GNN model achieved up to 90% accuracy across 12,000 trials on the non-synthetic European

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

LV grid with 111 loads. A PDF-driven selection of typical mixtures for the 12,000 trials might improve GNN model precision.

Regarding the count of typical mixes used to estimate uncertainty in line resistance and inductance, the review is limited to 32 mixes sampled via quasi-random Sobol sequences. Still, more precise insights into the stochastic spread of misclassifications could be obtained by increasing the number of mixes to 1000. Another simplification involves applying uniform shifts to resistance and inductance across all lines within each mix. Yet a fuller evaluation could incorporate varied shifts per mix. Notably, lines occupy different zones with differing weather and geographic settings, so shifts per mix might differ. These aspects should be included when setting the shift collection for each mix.

The examination examines the influence of multi-phase loads and PVs on the GNN model's precision with only a small number of additional units. Clearly, attaching varied single-, two-, and three-phase loads and PVs to three nodes may reduce the GNN model's precision from 100% to roughly 92%. This approximately 8% reduction shows that GNN model outcomes respond to the inclusion of several multi-phase units. Therefore, a deeper sensitivity review is of major significance for the next part, which involves higher integration rates of multi-phase loads and PVs. Additionally, comparisons between GNN and leading methods should be conducted across both segments of the PV sensitivity review.

Regarding expandability, the GNN model is tested on two grids with 55 and 111 loads, but further checks are needed. Specifically, larger grids with more loads and nodes between 500 and 1000 are essential for a suitable evaluation of the expandability feature. This research shows that even on the grid with 111 loads, precision can reach 90% after 12,000 Monte Carlo trials. Because trial quantities would increase steadily for large grids, an efficient sampling technique such as PDF must first be applied to reduce the required Monte Carlo trial count.

This research focuses on assessing GNNs using the GraphSAGE algorithm. Its outcomes are compared with leading methods, such as LASSO and k-NN, that have been used in phase detection studies. Different GNN varieties, such as Graph Convolutional Networks, Graph Attention Networks, and Graph Recurrent Networks, differ in how they combine and propagate information across nodes, which shapes their outcomes. Given that GraphSAGE achieves better precision than leading methods for phase detection, future studies could explore whether the aforementioned GNN variants can deliver comparable results (e.g., the number of misclassifications). Moreover, a review comparing these GNN types in terms of computational time and required Monte Carlo trial count could also be conducted.

Conclusion

In this paper, a CPI methodology grounded in GNN that draws on both synthetic and measured power patterns is proposed. Analysis indicates that 5000 and 12,000 Monte Carlo trials can achieve precisions of 98% and 90%, respectively, for two low-voltage grids with 55 and 111 loads. In addition, grid losses do not appear to affect GNN model outcomes when the combined load serves as the transformer pattern, without performing power flow calculations.

Under disturbed conditions, the GNN model maintains a precision of 95% and outperforms the leading methods, LASSO and k-NN. For unmetered consumption, the GNN model surpasses both LASSO and k-NN algorithms, achieving up to 50% accuracy.

Across diverse Photovoltaic scenarios, the GNN model also delivers robust performance, with precision exceeding 90%. Notably, the GNN model is minimally affected by Photovoltaics omitted during testing because they remain disconnected for maintenance reasons. Moreover, the inclusion of mixed multi-phase loads and Photovoltaics does not appear to affect the GNN model's outcomes significantly.

Regarding uncertainty in line resistance and inductance, multiple shifts are tested under a uniform probability density function. Evidently, the GNN model maintains a precision of around 95% even with up to 3 incorrect phase labels. Thus,

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

shifts in resistance and inductance values seem to carry negligible influence on GNN model precision, much like grid losses.

One key restriction identified in this research is avoiding inaccurate phase labels when creating synthetic training data, as they risk weakening GNN performance. In future efforts, an appropriate sampling technique should be integrated to generate random phase mixes, thereby reducing the number of Monte Carlo trials required. This will greatly support GNN model evaluation on bigger grids. Lastly, additional situations involving the installation of numerous multi-phase Photovoltaics and loads await examination to confirm GNN model durability to a greater degree.

Nomenclature

Abbreviations

ADMS Advanced Distribution Management System(s)

CPI Consumer Phase Identification

DER(s) Distributed Energy Resource(s)

DFT Discrete Fourier Transformation

DSO(s) Distribution System Operator(s)

GIS Geographic Information System

GNN(s) Graph Neural Network(s)

GPRS General Packet Radio Service

GRA Grey Relational Analysis

HV High Voltage

LASSO Least Absolute Shrinkage and Selection

Operator

LV Low Voltage

MIP Mixed Integer Programming

MLR Multivariate Linear Regression

ML Machine Learning

MV Medium Voltage

PCA Principal Component Analysis

PDF Probability Density Function

PLC Power Line Communication

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

PV(s) Photovoltaic(s)

QP Quadratic Programming

SA Salient Analysis

SCADA Supervisory Control and Data Acquisition

SF Salient Frequency

SM(s) Smart Meter(s)

SMI Smart Meter Infrastructure

Sets

E Set of edges

G Set of graphs

K Set of hidden layers

N Set of neighboring nodes

O Set of output layers

V Set of nodes and input layers

Subscripts

neigh Neighboring

self Self

i Index for numbering transformers

Superscripts

bus Bus

k Hidden layer

o Output layer

Symbols

σ Non-linear activation function

b Optional bias

e Edge embedding

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

h	Node embedding
v	Node
W	Weight
x	Input node feature
T	Transformer

Acknowledgements

This work was supported by the Common European-scale Energy Artificial Intelligence Federated Testing and Experimentation Facility (EnerTEF), which is a project funded by the European Union's Horizon Europe Research and Innovation Program (Grant 101172887).

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 05 Oct 2021 Revised: 11 Dec 2021 Accepted: 28 Dec 2021
Published online: 15 June 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Hosseini ZS, Khodaei A, Paaso A. Machine learning-enabled distribution network phase identification. IEEE Trans Power Syst. 2021;36(2):842-50.

2. Pappu SJ, Bhatt N, Pasumarthy R, Rajeswaran A. Identifying topology of low voltage distribution networks based on smart meter data. IEEE Trans Smart Grid. 2018;9(5):5113-22.

3. Pezeshki H, Wolfs PJ. Consumer phase identification in a three phase unbalanced LV distribution network. In: 2012 3rd IEEE PES Innov Smart Grid Tech Eur (ISGT Europe). 2012:1-7.

4. Mitra R, Kota R, Bandyopadhyay S, Arya V, Sullivan B, Mueller R, et al. Voltage Correlations in Smart Meter Data. In: Proc 21th ACM SIGKDD Int Conf Know Disc Data Min. 2015:1999-2008.

5. Hashmi MU, Brummund D, Lundholm R, Koirala A, Hertem DV. Consensus based phase connectivity identification for distribution network with limited observability. Sustain Energy Grids Netw. 2023;34:101070.

6. Wen MHF, Arghandeh R, Meier AV, Poolla K, Li VOK. Phase identification in distribution networks with micro-synchrophasors. In: 2015 IEEE Power Energy Soc Gen Meet. 2015:1-5.

7. Arya V, Seetharam D, Kalyanaraman S, Dontas K, Pavlovski C, Hoy S, et al. Phase identification in smart grids. In: 2011 IEEE Int Conf Smart Grid Commun (SmartGridComm). 2011:25-30.

8. García S, Merchan JMM, Larios DF, Personal E, Parejo A, Leon C. Phase topology identification in low-voltage distribution networks: A Bayesian approach. Int J Elect Power Energy Syst. 2023;144:108525.

9. Jimenez VA, Will A, Rodriguez S. Phase identification and substation detection using data analysis on limited electricity consumption measurements. Elect Power Syst Res. 2020;187:106450.

10. Zhou L, Zhang Y, Liu S, Li K, Li C, Yi Y, et al. Consumer phase identification in low-voltage distribution network considering vacant users. Int J Elect Power Energy Syst. 2020;121:106079.

11. Hoogsteyn A, Vanin M, Koirala A, Hertem DV. Low voltage customer phase identification methods based on smart meter data. Elect Power Syst Res. 2022;212:108524.

12. Xu M, Li R, Li F. Phase identification with incomplete data. IEEE Trans Smart Grid. 2018;9(4):2777-85.

13. Therrien F, Blakely L, Reno MJ. Assessment of measurement-based phase identification methods. IEEE Open Access J Power Energy. 2021;8:128-37.

14. Wang W, Yu N. Maximum marginal likelihood estimation of phase connections in power distribution systems. IEEE Trans Power Syst. 2020;35(5):3906-17.

15. Olivier F, Ernst D, Fonteneau R. Automatic phase identification of smart meter measurement data. In: 24th Int Conf Elect Distrib (CIRED). 2017:1-5.

16. Efkarpidis N, Geidl M, Wache H, Peter M, Adam M. Smart metering applications, main concepts and business models. Lect Notes Energy – Springer Cham. 2023:1-135.

17. Brint A, Poursharif G, Black M, Marshall M. Using grouped smart meter data in phase identification. Comp Oper Res. 2018;96:213-22.

18. Kumar P, Arya V, Bowden DA, Kohrmann L. Leveraging DERs to improve the inference of distribution network topology. In: 2017 IEEE Int Conf Smart Grid Commun (SmartGridComm). 2017:52-7.

19. Dilek M, Broadwater RP, Sequin R. Phase prediction in distribution systems. In: 2002 IEEE Power Eng Soc Winter Meet. 2002:985-90.

20. Jayadev SP, Rajeswaran A, Bhatt NP, Pasumarthy R. A novel approach for phase identification in smart grids using Graph Theory and Principal Component Analysis. In: 2016 Am Control Conf (ACC). 2016:5026-31.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

21. Arya V, Jayram TS, Pal S, Kalyanaraman S. Inferring connectivity model from meter measurements in distribution networks. In: 4th Int Conf Future Energy Syst. 2013:173-82.

22. Tang X, Milanovic JV. Phase Identification of LV Distribution Network with Smart Meter Data. In: 2018 IEEE Power Energy Soc Gen Meet (PESGM). 2018:1-5.

23. Santos JLT. Phase identification in low voltage smart grids using smart meter data. 1-9.

24. Zhao G, Zhao R, Gu Y, Chang B, Hu C, Xue Y. Phase identification of low-voltage users with novel grey relational analysis method. IOP Conf Ser: Earth Env Sci. 2020;546(042007):1-5.

25. Qingning P, Xutao L, Feihu H, Jin M, Kaimin S. Consumers' Phase Identification in Low Voltage Station Area Based on Wavelet Analysis of Consumption Data. In: 2021 IEEE Int Conf Power Intell Comput Syst (ICPICS). 2021:346-50.

26. Yi Y, Liu S, Zhang Y, Xue Y, Deng W, Li Q. Phase identification of low-voltage distribution network based on stepwise regression method. J Mod Power Syst Clean Energy. 2023;11(4):1224-34.

27. Yu H, Wu Y, Guan W, Zhang D, Yu T, Liu Q. Practical method for data-driven user phase identification in low-voltage distribution networks. Front Energy Res. 2021;9:1-7.

28. Zhang M, Luan W, Guo S, Wang P. Topology Identification Method of Distribution Network Based on Smart Meter Measurements. In: 2018 China Int Conf Elect Distrib (CICED). 2018:372-6.

29. Liu J, Zang W, Lu Y, Zhang Y, Cong R. Transformer-customer relationship and phase identification method of low-voltage distribution networks based on critical time segments. J Phys: Conf Ser. 2023;2666(012010):1-6.

30. Sun L, Chen Y, Du Q, Xu H, Wang W. Identification of low-voltage phase lines using IEC 61850 and K-means clustering. Elect Power Syst Res. 2024;234:110597.

31. Dahale S, Pahwa A, Natarajan B. Phase identification in unobservable distribution systems. IEEE Trans Power Del. 2023;38(5):3067-75.

32. Mugerwa G, Megahed TF, Elsabrouty M, Abdelkader SM. Data-driven consumer-phase identification in low-voltage distribution networks considering prosumers. Front Energy. 2024;18:827-40.

33. Vanin M, Acker TV, D'hulst R, Hertem DV. Phase identification of distribution system users through a MILP extension of state estimation. Elect Power Syst Res. 2024;229:110107.

34. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: A review of methods and applications. AI Open. 2020;1:57-81.

35. Liao W, Jensen BB, Pillai JR, Wang Y, Wang Y. A review of graph neural networks and their applications in power systems. J Mod Power Syst Clean Energy. 2022;10(2):345-60.

36. Hu W, Fey M, Zitni M, Dong Y, Ren H, Liu B, et al. Open graph benchmark: Datasets for machine learning on graphs. In: 34th Conf Neural Inf Process Syst (NeurIPS 2020). 2020:1-16.

37. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: 22nd ACM SIGKDD Int Conf Know Disc Data Min. 2016:855-64.

38. Hamilton WL, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs. In: 31st Conf Neural Inf Process Syst (NIPS 2017). 2017:1-19.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

39. IEEE. IEEE PES test feeder: 2015 test feeder cases - European low voltage test feeder. Available: https://cmte.ieee.org/pes-testfeeders/resources/.

40. Koirala A, Ramón LS, Mohamed B, Arboleya P. Non-synthetic European low voltage test system. Int J Elect Power Energy Syst. 2020;118:105712.

41. Commission for Energy Regulation (CER). Smart metering project – Electricity customer behaviour trial, 2009 – 2010 [dataset]. 1st ed. Irish Social Science Data Archive. SN: 0012-00; 2012. Available: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/. Accessed November 26, 2024.

42. Pfenninger S, Staffell I. Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. Energy. 2016;114:1251-65.

43. MATLAB. 2024. Available: https://www.mathworks.com/products/new_products/release2024a.html. Accessed November 27, 2024.

44. EPRI. OpenDSS. 2024. Available: https://www.epri.com/pages/sa/opendss. Accessed November 27, 2024.

45. Python. 2024. Available: https://www.python.org/downloads/release/python-3120/. Accessed November 27, 2024.

46. Ginocchi M, Ponci F, Monti A. Sensitivity analysis and power systems: Can we bridge the gap? A review and a guide to getting started. Energies. 2021;14(24):8274.

47. Niederreiter H. Random number generation and Quasi-Monte Carlo methods. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics; 1992.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press