energies

MDPI

Article

Machine-Learning-Driven Identification of Electrical Phases in Low-Sampling-Rate Consumer Data

Dilan C. Hangawatta, Ameen Gargoom * and Abbas Z. Kouzani

School of Engineering, Deakin University, Geelong, VIC 3216, Australia; dnaranapitihang@deakin.edu.au (D.C.H.); abbas.kouzani@deakin.edu.au (A.Z.K.)
* Correspondence: a.gargoom@deakin.edu.au

Abstract: Accurate electrical phase identification (PI) is essential for efficient grid management, yet existing research predominantly focuses on high-frequency smart meter data, not adequately addressing phase identification with low sampling rates using energy consumption data. This study addresses this gap by proposing a novel method that employs a fully connected neural network (FCNN) to predict household phases from energy consumption data. The research utilizes the IEEE European Low Voltage Testing Feeder dataset, which includes one-minute energy consumption readings for 55 households over a full day. The methodology involves data cleaning, preprocessing, and feature extraction through recursive feature elimination (RFE), along with splitting the data into training and testing sets. To enhance performance, training data are augmented using a generative adversarial network (GAN), achieving an accuracy of 91.81% via 10-fold cross-validation. Additional experiments assess the model's performance across extended sampling intervals of 5, 10, 15, and 30 min. The proposed model demonstrates superior performance compared to existing classification, clustering, and AI methods, highlighting its robustness and adaptability to varying sampling durations and providing valuable insights for improving grid management strategies.

Keywords: phase identification; low sampling rate; energy consumption data; fully connected neural network; recursive feature elimination; generative adversarial network

check for updates

Academic Editor: Marcin Kaminski

Received: 5 November 2024
Revised: 18 December 2024
Accepted: 20 December 2024
Published: 31 December 2024

Citation: Hangawatta, D.C.; Gargoom, A.; Kouzani, A.Z. Machine-Learning-Driven Identification of Electrical Phases in Low-Sampling-Rate Consumer Data. *Energies* **2025**, *18*, 128. https://doi.org/10.3390/en18010128

1. Introduction

1.1. Overview

The rapid deployment of distributed energy resources (DERs) has significantly increased the complexity of operating and maintaining low-voltage distribution systems. These systems, being the most dynamic and locally controllable components of the electrical grid, require precise and up-to-date information to ensure their reliable functioning. One critical piece of information is the phase connectivity of consumers. In electrical power systems, consumer loads are typically divided into different phases to balance the load across the electrical grid as shown in Figure 1.

Proper identification of these phases is critical for several reasons. Accurate phase identification helps in balancing the electrical load across the phases, which is essential for maintaining the efficiency and stability of the power distribution system. Understanding which loads are on which phases allows for better detection of faults and more targeted maintenance, which can reduce downtime and improve reliability. For utilities and energy providers, knowing the phase distribution helps in optimizing energy distribution and minimizing losses, leading to more efficient energy use and cost savings.

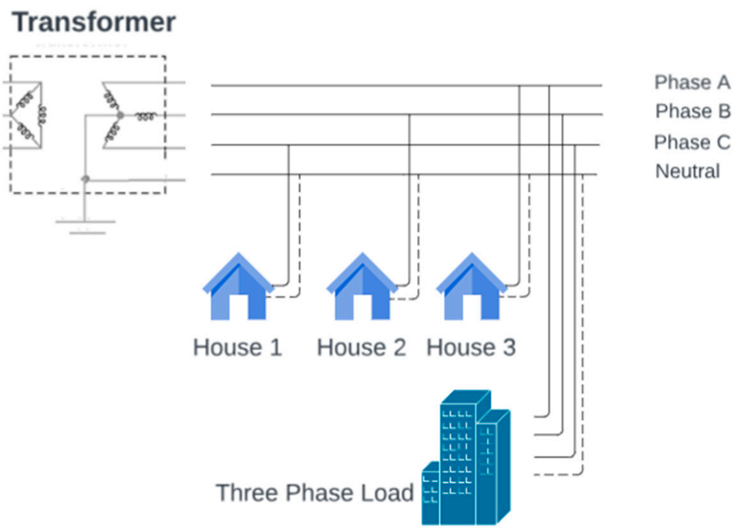

Figure 1. Phase connection of a distribution network.

However, distribution network operators frequently face a significant challenge due to the lack of detailed phase information, which impairs their ability to maintain system stability and performance. This gap in phase knowledge can lead to inefficient load balancing, increased risk of overloads, and reduced overall grid reliability. To mitigate these issues, it is imperative to develop and implement reliable methods for generating and acquiring accurate phase information.

The advent of smart meter data has significantly transformed the process of phase identification, enabling more accurate and data-driven methods to determine the phases of electricity consumption. The utilization of smart meter data for electrical phase identification signifies a shift towards data-driven methodologies in energy management. By leveraging advanced analytics, clustering algorithms, and machine learning models, utilities can extract valuable insights from smart meter data to enhance network operations, improve efficiency, and ensure the reliability of electrical systems. The comprehensive analysis of smart meter data not only facilitates phase identification but also enables a wide array of applications that drive innovation and optimization in the energy sector.

In this article, a fully connected neural network model is proposed for phase identification of consumers, based on their energy consumption data with longer sampling period data (i.e., smart meter readings greater than 1 min). The results are compared with other existing machine learning algorithms.

Existing methods for phase connectivity identification can be categorized into two main types: traditional methods and advanced methods. Each category presents distinct advantages and limitations that are crucial for selecting the appropriate approach based on specific scenarios. Traditional methods primarily include manual inspection and testing techniques. In visual inspection, electricians or technicians physically examine the distribution network, tracing connections from transformers to service drops to identify which phases are linked to specific loads. The signal injection method involves sending a reference signal from one of the three phases at the substation while another crew member receives it at the destination point. Despite their effectiveness, these traditional methods face practical challenges, such as implementation complexities, time consumption, and high labor demands, making them less feasible for widespread application. Advanced methods can be further divided into data-driven approaches and those utilizing high-precision meter devices. The literature [1,2] highlights the use of time-synchronized measurements from high-precision phasor measurement units, particularly micro-synchro phasors (μPMUs).

With the rapid deployment of smart meters, data-driven methods for phase identification have gained popularity, and they can be categorized into three subtypes: optimization-based, machine-learning-based, and neural-network-based approaches. Optimization-based methods [3–7] leverage power measurements from both the customer end and the distribution transformer to establish phase connections. These methods utilize the principle of conservation of power, asserting that the sum of power measurements from all consumers connected to a specific phase should equal the power recorded at the transformer end for that phase. Machine-learning-based methods can be divided into two categories: those using voltage measurements and those utilizing energy consumption data. Studies referenced in [8–24] explore various machine learning techniques that employ voltage measurements from smart meters, grounded in the hypothesis that time-series voltage data from end nodes connected to the same phase exhibit strong similarities. This similarity allows machine learning algorithms to effectively identify and classify consumer phases based on voltage patterns. Focusing on energy consumption data, several methods in the literature rely solely on kWh readings [25–35]. Compared to voltage-measurement-based methods, power consumption data offer notable advantages in terms of availability, practicality, and less complexity, and voltage measurements often show small variations.

In [25], the author presents a hybrid approach combining graph theory, energy conservation, and principal component analysis (PCA) to determine phase connectivity. While this method simplifies connectivity analysis by modeling transformers as parent nodes, the specific role of PCA is unclear, and noise modeling needs refinement for real-world applications. Methods [26–28] rely on the premise that consumer demand closely aligns with the supply phase. In [26], significant variations in household power consumption are extracted for correlation analysis, though feature extraction criteria remain undefined. Method [27] introduces a "modified k-means clustering" technique that uses aggregated phase data as centroids to enhance accuracy. Method [28] employs discrete wavelet transform (DWT) to analyze load data variations, using Daubechies 4 wavelets for comparison. Additional studies explore diverse phase identification methods, including those of Xiong et al. [29], who categorize power consumption data, and Gao et al. [32], who apply K-Means++ and GCN for precise appliance monitoring, underscoring advancements in utilizing smart meter data for phase identification

Neural-network-based methods for phase identification are relatively scarce and have not been extensively applied to this specific problem. However, insights from studies [36–39] can inform the development of effective neural network models for phase identification.

A key challenge in phase identification using power consumption data is its modeling, as proposed by Arya et al. [3,5,16] through a linear model with binary variables. Accurate data from multiple customers are essential, but issues like incomplete measurements and communication errors can lead to significant identification errors, affecting load balancing and grid stability. The reliability of phase identification methods relies on the availability and quality of power consumption data. Incomplete or noisy data can lead to inaccurate phase identification, resulting in inefficient load balancing and higher operational costs for utilities. To improve accuracy, researchers suggest using wavelet analysis to extract features from consumption data, though effectiveness remains contingent on data quality.

Privacy and security concerns arise from collecting and analyzing power consumption data for phase identification. Li [40] emphasizes the risks of data leakage and privacy breaches in the power Internet of Things, advocating for trusted decision fusion methods to safeguard sensitive information. Ethical issues related to data collection and consumer privacy rights may also provoke debate.

To address the challenges of electrical phase identification using low sampling rates, it is essential to evaluate how sampling duration affects the accuracy of identification techniques. Increasing the sampling frequency can enhance accuracy, particularly when analyzing complex waveforms under periodic stress. However, lengthening the observation period introduces significant challenges in achieving high accuracy. Among the referenced studies, the work in [41] is particularly relevant, exploring the variability in smart meter sampling frequencies, from fast rates to longer intervals, which aligns with the context of phase identification. The study discusses using deep learning techniques to analyze power consumption data over extended periods, which is crucial for this task. Additionally, the reference in [42] provides insights into smart meter data analytics, highlighting methodologies and challenges pertinent to electrical phase identification with longer sampling periods. Furthermore, Hoogsteyn et al. [31] offer specific methodologies tailored to smart meter data for phase identification, addressing the challenges posed by longer sampling periods. Improving phase identification accuracy with long sampling data is complicated by factors like measurement errors and data distortion, as discussed by Foggo and Yu [43], who achieved a significant accuracy increase from 51.7% to 97.3% using supervised methods. Challenges remain, particularly in ensuring consistent results, as noted by Yu Wang and Yu [44], and comprehensive data collection is critical for precision [45]. Factors like missing data and synchronization issues further complicate accurate phase identification, as highlighted by Cleenwerck [46].

1.2. Related Work

Several studies have shown that using energy consumption data, specifically kWh readings, can be more practical for phase identification than voltage measurements. This section reviews relevant works that leverage power consumption data and outlines their contributions and limitations, clarifying how our work contributes to the field. Graph theory and PCA [25] propose modeling transformers as parent nodes and utilizing PCA to simplify phase connectivity analysis. While novel and effective for modeling, the use of PCA for noise handling and its specific implementation require refinement for real-world use. Correlation analysis methods, such as those in [26], focus on analyzing significant demand variations and correlating them with phase data but suffer from undefined feature extraction criteria, impacting consistency and reproducibility. Clustering techniques, like the modified k-means clustering in [27], use aggregated data as centroids to enhance phase identification accuracy. However, these methods are sensitive to data variations and may underperform with noisy real-world data. Wavelet analysis, as applied in [28] with discrete wavelet transform (DWT) and Daubechies 4 wavelets, is effective for feature extraction and phase identification in well-conditioned datasets but can be computationally intensive and vulnerable to data quality issues. Advanced categorization techniques, such as those by Xiong et al. [29] and Gao et al. [32], employ K-Means++ and graph convolutional networks (GCNs) for phase identification and appliance monitoring, contributing to more refined data processing techniques. Nevertheless, these approaches do not adequately address the challenges associated with extended sampling periods or handling limited data. Our work advances the field by proposing a robust fully connected neural network model that maintains high accuracy across various sampling intervals and data limitations, offering a more practical solution for real-world smart grid applications.

While the aforementioned studies have made valuable contributions, they also have limitations that our work addresses. We introduce an innovative fully connected neural network (FCNN) model specifically designed for phase identification using energy consumption data, surpassing traditional and other advanced methods in terms of accuracy and robustness. Our model demonstrates strong performance even with longer sampling periods (e.g., from 1 min to 30 min intervals), maintaining accuracy with less than 10% degradation, which contrasts with other methods that often experience significant accuracy losses as sampling periods increase. Additionally, our approach is robust in scenarios involving limited data, a critical consideration given the scarcity of comprehensive datasets in phase identification research. Through rigorous 10-fold cross-validation, our model achieved an impressive accuracy of 91.81% and an F1 score of 0.9591, outperforming existing machine learning models (e.g., SVM, decision trees, LSTM, GRU) and other data-driven methods.

The main contributions of this article are clearly defined as follows:

- Novel Methodology: The primary contribution of this study is the development of a novel method that employs a fully connected neural network (FCNN), enhanced by data augmentation through a generative adversarial network (GAN) and recursive feature elimination (RFE) for feature selection. This innovative approach facilitates accurate phase identification from energy consumption data sampled at low rates, addressing a significant limitation in previous research that has predominantly focused on higher-frequency data. The method has been tested across various sampling intervals, demonstrating effective performance under different data conditions, which makes it suitable for practical applications in energy management.
- Comprehensive Performance Evaluation: We conducted a thorough benchmarking of our proposed model against various established methods from the literature, including traditional classification algorithms (e.g., SVM, decision trees) and advanced neural network architectures (e.g., LSTM, GRU). This extensive comparison emphasizes the strengths of our approach, particularly in its ability to effectively handle data with low sampling rates.
- Handling Limited Data: Our model's performance is resilient in scenarios with limited data, addressing the challenge of data scarcity in phase identification tasks. This is especially valuable as there is a gap in research that effectively handles long sampling periods with limited datasets.

2. Materials and Methods

2.1. Proposed Model Architecture

The proposed model is a six-layered fully connected neural network (FCNN) model that consists of an input layer, 4 hidden layers, and an output layer combining active functions and regularization techniques to properly train the model. The dataset is split into training and testing sets, ensuring that the model can be evaluated on unseen data. Firstly, preprocessing is applied, which includes the normalization of input data and encoding of categorical variables to ensure optimal model performance. Then, feature selection is carried out to identify the most relevant time-based features using the recursive feature elimination (RFE) method. Further, the generative adversarial network (GAN) model is used for data augmentation as the original dataset is limited. The augmentation is only applied on training data. The basic architecture of the proposed model is depicted in Figure 2.

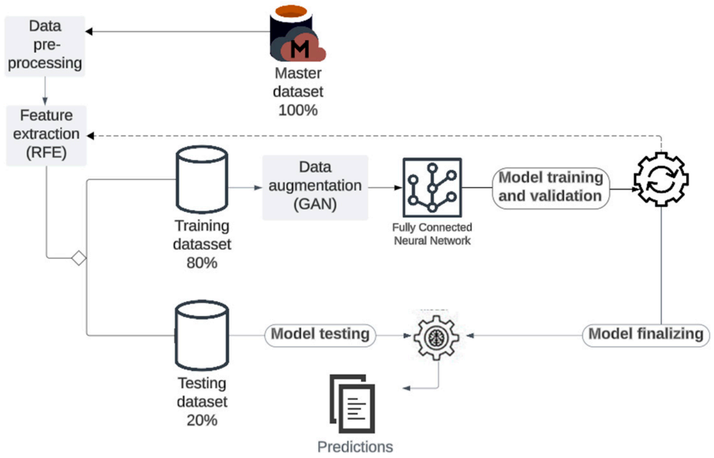

Figure 2. Proposed model architecture.

2.2. Dataset

This paper utilizes data from the IEEE European Low Voltage (LV) Test Feeder [47], a representative three-phase low-voltage distribution system in Europe, operating at 50 Hz. The dataset, detailed in Table 1, includes energy consumption data measured in kilowatt-hours (kWh) from 55 houses over a one-minute sampling period, totaling 1440 samples per house for one day. The load distribution comprises 21 loads on Phase A, 19 on Phase B, and 15 on Phase C, providing a comprehensive view of the energy usage patterns within the network. We utilize this dataset as the master dataset for our proposed model architecture, which involves three main steps: preprocessing, feature extraction, and data augmentation. During preprocessing, we clean and standardize the data to ensure its quality and consistency. Next, feature extraction identifies key characteristics that enhance the model's performance. Finally, data augmentation expands the dataset through techniques that create synthetic variations, helping to improve the robustness and generalizability of the model. These steps are crucial for effectively training and validating our architecture using the LV Test Feeder data.

Table 1. Characteristics of the dataset.

Characteristics	Description
Dataset	IEEE European LV test feeder data
Type	energy consumption (kWh)
Number of houses	55
Sampling period	1 min
Samples per house	1440 (i.e., recordings for 1 day)
Load distribution	21 loads—Phase A, 19 loads—Phase B, 15 loads—Phase C

2.3. Preprocessing

In our preprocessing phase, we implement data normalization to adjust the scale of the data, using the formula shown in Equation (1). This process standardizes the data by subtracting the mean and dividing it by the standard deviation, which helps to bring all features onto a similar scale.

$$x_{norm} = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the mean of the data and σ is the standard deviation of the data.

We also encode categorical variables, transforming non-numeric data into a numerical format that the model can effectively interpret. Additionally, data cleaning is performed to eliminate noise and irrelevant information, such as duplicates and missing values. This thorough approach to preprocessing not only enhances the quality of the dataset but also ensures that the neural network can leverage the data effectively for robust training and testing. By addressing these aspects, we improve the model's accuracy and reliability, leading to better performance in real-world applications.

2.4. Feature Extraction

Feature extraction plays a crucial role in neural network modeling, focusing on retaining relevant features while discarding redundant ones. This process enhances model performance by improving accuracy, reducing the risk of overfitting, and lowering computational costs. In our proposed model, we employ recursive feature elimination (RFE), a systematic technique that iteratively removes less significant features to refine the dataset [48,49]. RFE constructs models using various subsets of features to evaluate their performance, enabling the identification of the most influential attributes that contribute to predictive power.

The RFE process involves several key steps: It starts with training a model on the entire feature set and then ranks features based on their importance. Subsequently, the least significant features are removed, and the model is retrained [50]. This cycle continues until the optimal subset of features is identified, maximizing the model's predictive capabilities while maintaining interpretability. By systematically eliminating features that do not significantly enhance the model, RFE ensures a more efficient training process. The detailed steps of the RFE procedure are outlined in Algorithm 1, providing a structured approach to feature selection that supports robust neural network training.

Algorithm 1. Recursive Feature Elimination (RFE)

```
Step 1: Train the model using all the features (i.e., 1440).
Step 2: Determine model's testing accuracy.
Step 3: Determining feature ranking using feature importance as follows:
Consider a dataset $\{X\}$ with $(n)$ features and a target variable $\{y\}$. We denote the feature matrix as
$X = [x_1,\ x_2, \ldots\ldots, x_n]$
where $x_i$ represents the $i$-th feature. A model $f$ is on the dataset.
$y = f(X)$
The importance of features is calculated using logistic regression.
The importance of a feature $x_i$ is given by the absolute value of its coefficient $|\beta_i|$.
Importance $x_i$ = $|\beta_i|$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$
where $\beta_i$ is the coefficient of $x_i$ in the model and $\varepsilon$ is the error term.
Step 4: Feature elimination
Features with the lowest importance score are eliminated. Let us denote the feature set at step $k$ as $X_k$.
After removing the least important feature:
$X_{k+1} = X_k$ {feature with lowest importance}.
Step 5: For each subset $S_i$, $i = 1, \ldots\ldots,\ N$ with important features
and train the model.
Step 6: Evaluate the accuracy of the model.
Step 7: Use the model corresponding to the appropriate number
of optimal features.
```

2.5. Data Augmentation

Data augmentation is a fundamental technique in machine learning [51,52] which involves expanding the original dataset by applying transformations. This process aims to enhance model performance, improve generalization, and mitigate overfitting by providing the model with a more diverse and extensive set of training examples.

Time-series data, known for their sequential nature and temporal dependencies, pose unique challenges due to their high dimensionality and time-correlated features. Augmenting time-series data is essential for expanding the size and quality of training datasets, which is important for the successful application of models and has been shown to be instrumental in improving model performance by expanding the dataset and reducing overfitting.

Various techniques are available to effectively augment time-series data. The proposed model uses generative adversarial networks (GANs) to create synthetic data [53,54] that closely mimic the original time-series data's distribution. By generating new samples that capture the temporal patterns and dependencies present in the data, GAN-based augmentation methods have shown promise in improving model accuracy and robustness. This technique is particularly valuable in scenarios where the availability of labeled data is limited and it is impractical to expand the real data like limited time-series data where augmenting the dataset can significantly enhance the performance of deep learning models. A generator and a discriminator are the two main parts of the GAN. They are trained simultaneously during adversarial training [55]. The generator's goal is to synthesize data samples from random noise that is indistinguishable from real data, and the discriminator correctly separates samples as real or fake. Figure 3 shows the process of augmentation on the training dataset using GANs.

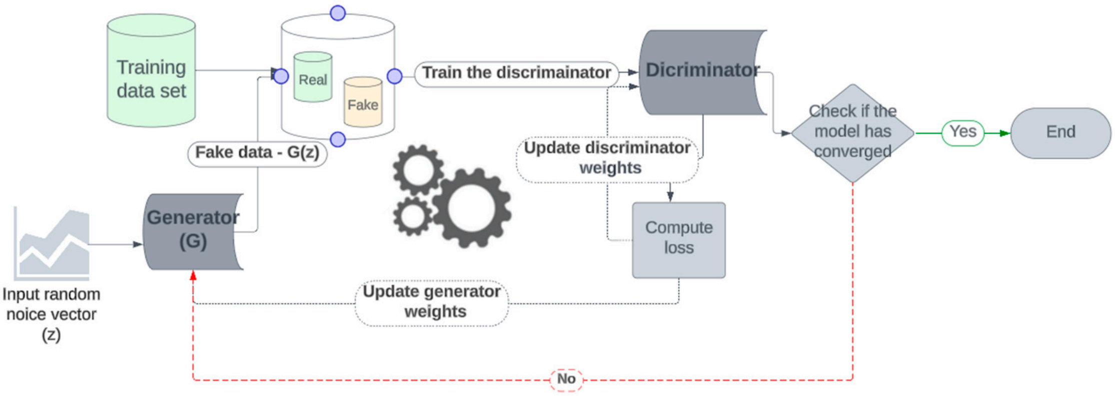

Figure 3. Data augmentation using GANs.

2.6. *Proposed Fully Connected Neural Network*

Figure 4 shows the detailed structure of the proposed model which is designed to accurately predict the electrical phases of the houses with a limited amount of data. This detailed structure consists of three main parts, which are data preparation, proposed multi-layer model, and model evaluation.

In Figure 4, the green color represents the data preprocessing part of the proposed model which is discussed in Section 2.3. The orange color area is the second part of the proposed model, which is the multi-layered fully connected neural network. And the blue color section represents the model validation and performance analysis. This proposed fully connected neural network consists of an input layer, four hidden layers, and an output layer with neurons that help to perform complex computations and make accurate predictions for electrical phases.

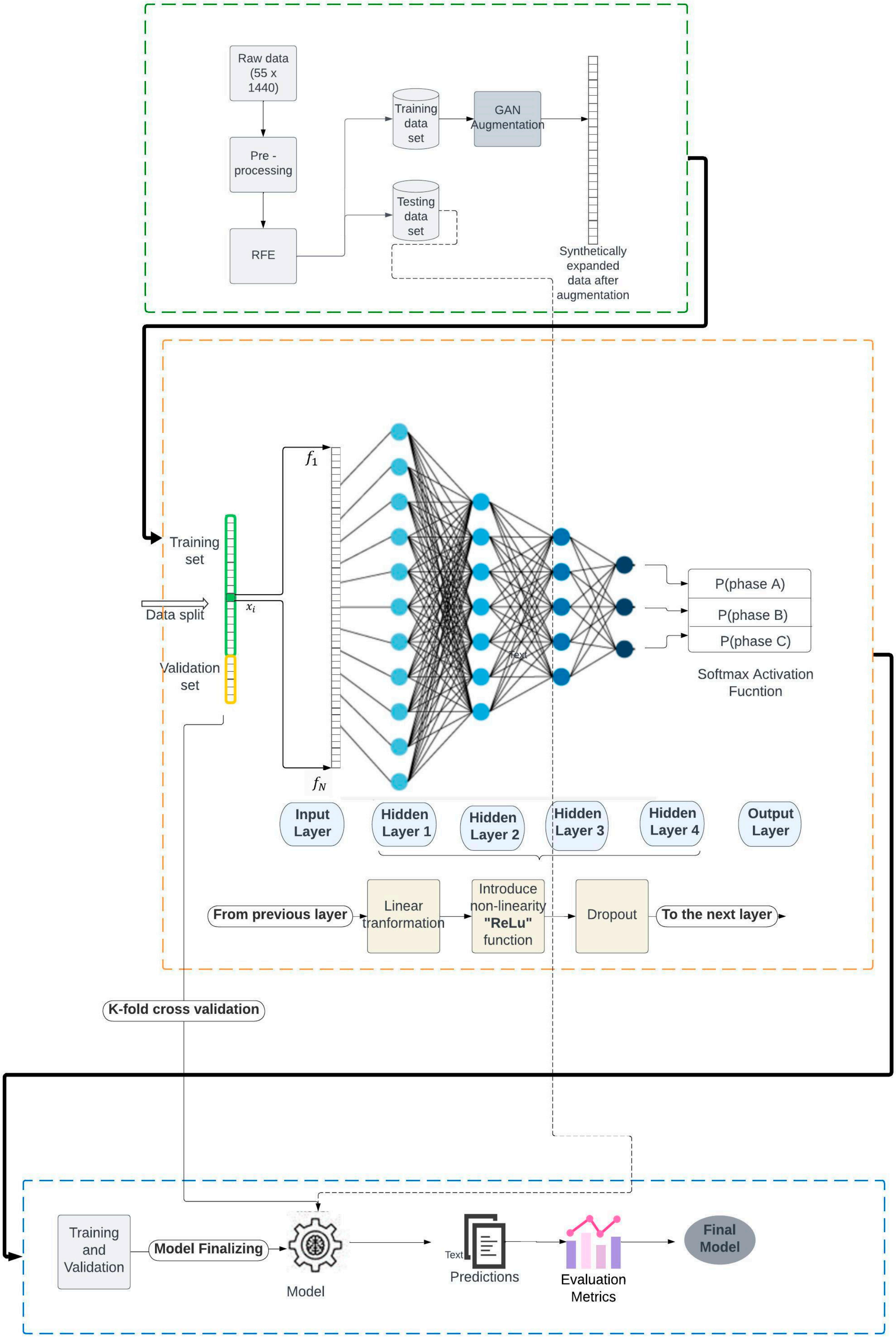

Figure 4. Proposed fully connected neural network.

The input layer of the model receives all features from a single house at a time and passes these data to the subsequent hidden layers for further processing. Since the input layer does not transform or process the data, it does not apply an activation function; instead, it simply transmits the preprocessed input data to the first hidden layer. In the initial iteration, before applying recursive feature elimination (RFE) for feature reduction, all available features are utilized in the model. Each house is equipped with smart meters that provide readings at one-minute intervals, resulting in 1440 readings per day. Consequently, the input layer is designed with 1440 neurons, each corresponding to one minute of reading. These neurons are responsible for capturing and transmitting the relevant information from the smart meter data, facilitating the learning process in the hidden layers. By effectively utilizing all features at this stage, the model aims to fully understand the data before optimizing them through feature selection techniques like RFE. This structured approach ensures comprehensive data representation for improved model performance.

Each layer in the network transforms its input through a combination of linear transformation (matrix multiplication and bias addition) and non-linear activation functions, with dropout applied during training to prevent overfitting. During the training process, the network adjusts the weights and biases of its interconnected neurons.

The proposed neural network model is characterized by several key parameters that significantly influence its architecture and performance. It consists of four dense (fully connected) layers, enabling the network to learn complex patterns through comprehensive connections between neurons. A batch size of 32 is employed, which balances computational efficiency with gradient descent stability, ensuring effective updates to the model's parameters. The hidden layers feature neurons in a decreasing sequence of [256, 128, 64, 32], facilitating the capture of high-level features while managing model complexity. To assess the model's performance and prevent overfitting, 20% of the training data are set aside as a validation split. Training occurs over 30 epochs, providing sufficient exposure to the dataset while avoiding excessive training that could lead to overfitting. A dropout rate of 50% is applied to enhance robustness by randomly deactivating half of the neurons during training. The Adam optimizer, known for its efficiency, adapts learning rates for each parameter, promoting rapid convergence. The categorical cross-entropy loss function is utilized for multi-class classification, guiding the model to make accurate predictions by measuring the discrepancy between predicted probabilities and actual class labels. The ReLU (Rectified Linear Unit) activation function introduces non-linearity, defined as follows:

$$f(x) = max\ (0, x) \tag{2}$$

enabling the model to learn intricate mappings. Finally, the softmax function is applied in the output layer, transforming the output into probability distributions across multiple classes, which is essential for effective classification tasks.

The parameter setting for the proposed fully connected layer is represented in Table 2.

Table 2. Parameter setting for the proposed model.

Parameter	**Value**
Number of dense layers	4
Batch size	32
Hidden neurons	[256, 128, 64, 32]
Validation split	20%
Epochs	30
Drop rate	0.5
Optimizer	Adam
Loss function	Categorical cross entropy
Activation function	ReLU
Activation function output	Softmax

2.7. Model Training

Training the proposed neural network involves presenting it with a dataset to learn from and adjusting its weights and biases through backpropagation to minimize errors and improve performance. During training, the model iteratively refines its parameters to optimize its ability to make predictions or classifications based on the input data. The training process aims to enable the neural network to generalize well to unseen data, thereby enhancing its predictive capabilities.

2.8. Model Validation

Validation is another critical step in the neural network development process, serving to fine-tune the model and prevent overfitting. Validation involves assessing the model's performance on a separate dataset, distinct from both the training and testing sets, to ensure that the model generalizes well. Further, validation helps to fine-tune the model's architecture, regularization techniques, and other parameters to achieve optimal performance.

Monitoring the training and validation loss curves is essential for evaluating the proposed model's performance, optimizing training processes, and ensuring robust generalization to unseen data. Figure 5 shows the training and validation loss curves of the proposed model.

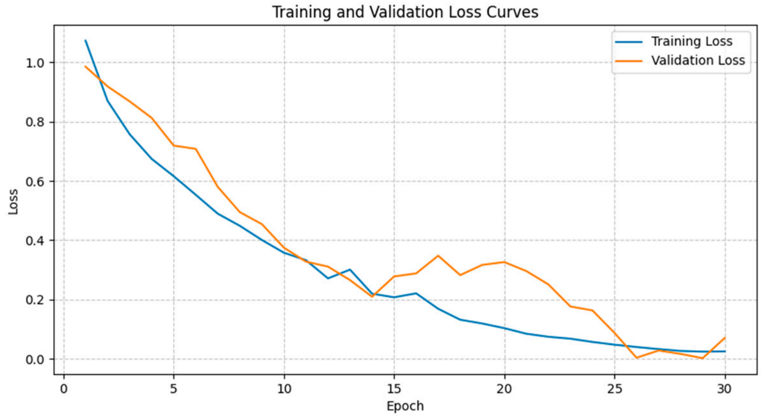

Figure 5. Training and validation loss curves of the proposed model.

The training loss curve (blue graph in Figure 5) illustrates changes in the loss function over epochs, indicating the model's fit to the training data. Initially, the training loss decreases rapidly from 1.072 at epoch 0 to 0.449 at epoch 8, reflecting effective learning. However, the curve flattens out from epoch 21 to 30, stabilizing around 0.05, indicating incremental improvements. Conversely, the validation loss curve (orange graph) shows the model's performance on unseen data. It decreases initially from 0.985 at epoch 1 to 0.311 at epoch 12 but then increases, suggesting potential overfitting. After epoch 20, the validation loss decreases again, reaching 0.071 at epoch 30, indicating good generalization to unseen data.

3. Results

3.1. Model Testing

Once the neural network is trained on a dataset, it is crucial to evaluate its performance on unseen data to assess its effectiveness and generalization ability. Testing of the proposed model involves feeding it with a separate dataset (i.e., 20% of the original data) that it has not encountered during training to evaluate how well it can make predictions. The testing accuracy for the proposed model is 91%. This means that out of 11 testing houses, the model is able to predict the electric phases of 10 households correctly. The actual phase group of "node 20", which is group A, is wrongly identified as group B. Table 3 shows the predictions of the proposed model on the testing dataset.

Table 3. Predictions of proposed FCNN on testing dataset.

Node	Ground Truth	Prediction
Node 5	A	A
Node 10	B	B
Node 15	B	B
Node 19	C	C
Node 20	A	B
Node 25	A	A
Node 37	B	B
Node 42	C	C
Node 46	A	A
Node 49	A	A
Node 50	B	B

In order to assess the model performance comprehensively the test result is further analyzed using evaluation metrics like confusion matrix, accuracy, macro avg, weighted avg, precision, recall, and F1 score.

The result from the confusion matrix is described in Table 4.

Table 4. Confusion matrix on testing dataset.

Class A	Class B	Class C
True Positives (TP): 4	True Positives (TP): 4	True Positives (TP): 2
False Positives (FP): 1 (classified as A but actually B)	False Positives (FP): 0 (classified as B but actually A)	False Positives (FP): 0
False Negatives (FN): 0	False Negatives (FN): 0	False Negatives (FN): 0

For overall performance evaluation, we use the evaluation metrics represented in Table 5.

Table 5. Performance evaluation of the proposed model.

Metric	Formula	Value
Sensitivity	Rate of positives correctly classified =TP/(TP + FN)	0.9191
Specificity	Rate of negatives correctly classified =TN/(TN + FP)	0.9524
F1 score	Harmonic mean between precision and recall =2 × (precision × recall/(precision + recall)	0.9091
AUC	Area Under the Curve–Measure of the trade-off between the TP rate and FP rate	0.9286
AUPRC	Area Under Precision-Recall Curve	0.9242

3.2. Experiments

In this section, further experiments are carried out to gain a comprehensive understanding of the proposed model's effectiveness.

3.2.1. Ensure the Model's Robustness from K-Fold Cross-Validation

The proposed model employs cross-validation to enhance its robustness and consistency across varying data splits. This process involves dividing the dataset into multiple subsets, or folds, which are then used alternately for training and validation. By using different segments of the data for these purposes, cross-validation helps ensure that the model is not overly fitted to any specific subset, promoting generalization to unseen data.

During the cross-validation process, the model is trained multiple times, each time using a different fold for validation while the remaining folds are utilized for training. This iterative approach allows for a comprehensive evaluation of the model's performance. The results from this cross-validation revealed an impressive average accuracy of 91.81%, along with a standard deviation of 0.053. This indicates that the model consistently performs well across the various folds, exhibiting relatively low variability. Such performance metrics reflect the model's reliability and effectiveness in making predictions, suggesting that it can generalize well to new data while minimizing the risk of overfitting. Overall, the use of cross-validation significantly strengthens the credibility of the model's results.

3.2.2. Accuracy Variation with Feature Extraction and Data Augmentation

Table 6 shows how the accuracy is varied with the incorporation of feature extraction methods and data augmentation methods, highlighting their impact on performance.

Table 6. Performance variation from RFE and data augmentation.

	Method	Accuracy
1.	Using all the features and no data augmentation	45.45%
2.	Using RFE feature extraction and no data augmentation	63.63%
3.	Using RFE feature extraction and data augmentation with Range Shift (RS), Time Shift (TS), Magnitude Warping (MW), and Time Warping (TW)	72.72%
4.	Proposed model (i.e., RFE feature extraction + GAN data augmentation + 10-fold CV	91.81%

3.2.3. Model Performance Comparison with Other Existing Classification and AI Methods on Same Data

The same dataset is utilized to evaluate other existing classification and AI methods, allowing for a comparative analysis of the proposed model's performance. Table 7 presents the accuracies obtained from various approaches. From this analysis, it is clear that the proposed model significantly surpasses the performance of other methods, achieving higher accuracy levels. Specifically, all traditional classification methods and neural network models yield accuracy rates below 30% when applied to the same preprocessed data. In contrast, cross-correlation-based methods achieve an accuracy of 40%, suggesting some level of correlation in power consumption patterns among the houses. However, this correlation alone proves inadequate for effective modeling with statistical methods. The substantial difference in accuracy underscores the advantages of the proposed model, demonstrating its superior capability to capture complex relationships within the data, leading to more reliable predictions. This highlights the model's effectiveness in addressing the challenges of the dataset compared to other methodologies.

Clustering-based methods that incorporate feature engineering techniques like PCA, t-SNE, and RFE show improved performance compared to traditional methods. However, the proposed model outperforms these methods due to its advanced feature engineering, sophisticated architecture—including specialized layers, activation functions, and regularization techniques—along with better-tuned hyperparameters and effective data augmentation. Traditional methods and neural networks often face challenges such as overfitting or underfitting due to inadequate training settings and limited data. In contrast, the proposed model is designed with assumptions that align more closely with the characteristics of the data. Its customized approach, which integrates novel algorithms and techniques, further enhances its performance. This comprehensive strategy ensures that the model not only captures the complexities of the data effectively but also provides robust predictions, solidifying its position as the superior method in this comparative analysis.

Table 7. Performance comparison with other methods.

Method Type	Method Name	Accuracy %	Sensitivity	Specificity	F1 Score	AUC
Statistical methods	Cross-correlation method	40%	0.3553	0.4389	0.3437	0.343
Clustering methods	PCA + Kmeans	71%	0.8123	0.7944	0.8092	0.8091
	TSNE + DBSCAN	68%	0.6743	0.6642	0.6834	0.6830
Classification methods	Decision Tree (DT)	21.43%	0.1965	0.2054	0.2105	0.2035
	Random Forest (RF)	14.29%	0.1378	0.1411	0.1392	0.1333
	K Nearest Neighbor (NN)	21.43%	0.2120	0.2140	0.2092	0.2209
	Gaussian Naïve Bayes (GNB)	28.57%	0.2789	0.2901	0.2788	0.2820
	Logistic Regression (LR)	14.29%	0.1372	0.1433	0.1392	0.1333
	Support Vector Machine (SVM)	35.71%	0.3491	0.3677	0.3503	0.3489
Neural network models	Simple Recurrent NN	18.18%	0.1719	0.1934	0.1835	0.1797
	GRU	36%	0.3593	0.3608	0.3589	0.3536
	LSTM	18.18%	0.1720	0.1936	0.1835	0.1799
	Bidirectional LSTM	27.27%	0.2699	0.2800	0.2702	0.2689
	Proposed fully connected model	91.81%	0.9191	0.9524	0.9091	0.9286

3.2.4. Performance Evaluation of the Proposed Model with Low-Sampling-Rate Data

As mentioned before, the original dataset is one-minute sampling period data for 55 households. In order to evaluate the performance of the proposed model over the longer sampling period data, the same dataset is downsized for 5 min, 10 min, 15 min, and 30 min samples. Then, the accuracy is observed, and the avg accuracy is determined for each downsized dataset with 10-fold cross-validation. Figure 6 shows the accuracy variation of the proposed method over the longer sample periods.

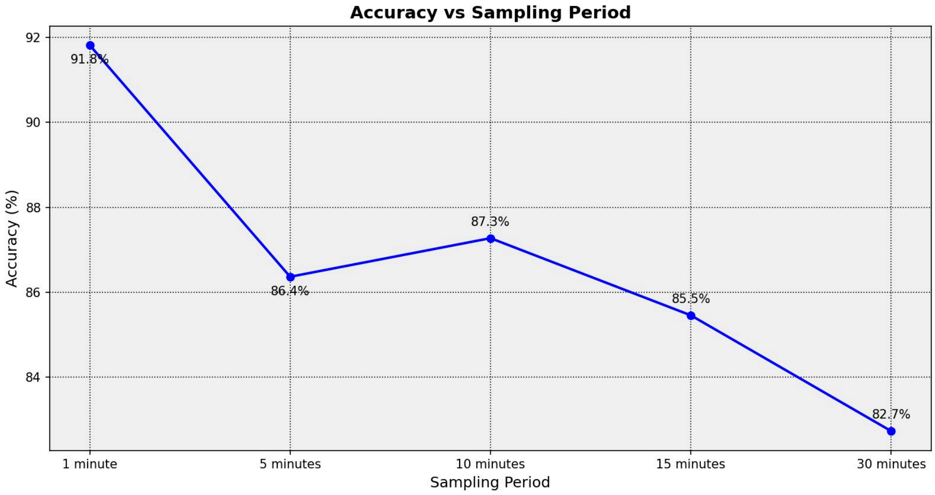

Figure 6. Proposed model accuracy variation over sampling periods.

According to Figure 6, the results from our proposed neural network model for phase identification demonstrate impressive performance at low sampling rates, with accuracies of 91.81% for 1 min and around 86% for 5 to 10 min. These results highlight the model's effectiveness in processing data from smart meters, which typically use low sampling frequencies in household applications. Achieving such high accuracy with limited data points is particularly significant, as it aligns well with the practical constraints of standard smart meters in residential settings. This capability enhances the feasibility of implementing advanced phase identification techniques in real-world scenarios, improving grid management and energy efficiency.

4. Discussion

In the context of maintaining the balance of the electrical distribution network and enhancing the reliability of the network, identifying the correct electrical phase connection of households should be the central focus, as absent or misleading information leads to a reduction in the longevity of the network resources. There are various methods available for that purpose. Our focus is on machine learning methods based on the smart meter reading of households.

4.1. Model Performance

We use IEEE European test feeder data and develop a fully connected neural network model to accurately identify the phase group of each house. We obtained 91.81% average accuracy with 10-fold cross-validation on the test data. This accuracy is for a 1 min sampling period. This accuracy level is comparatively significant compared to the other statistical, clustering, classification methods, and neural network models performed with the same data as shown in a table. This indicates that the model correctly classifies 92% of the samples on average, which is typically considered excellent performance, especially in many practical scenarios. Furthermore, the standard deviation of 0.053 (or 5.3%) is relatively low, which means that the accuracy is consistent across the different folds. A low standard deviation implies that the model's performance is stable and does not fluctuate.

Moreover, we examine the variation in the accuracy with each step before finalizing the model as shown in Table 6. The results obtained from that table validate the effectiveness of the proposed fully connected neural network using feature extraction with RFE and its capacity to recursively learn from the model itself and identify the most important features during the training process as it completely outperforms the results achieved using the total number of features. When the conventional data augmentation methods are replaced with the GAN model to expand the training data without losing its originality, the accuracy exhibits a significant difference. This reflects that by collaboratively using recursive feature extraction, data augmentation, and cross-validation, the performance can be enhanced significantly. According to the classification report, the model performs well with an accuracy of 91%. It has high precision and recall across all classes. Based on the class-specific performance, class A and class C have very high precision and recall, while class B has slightly lower precision but perfect recall. A high macro average reflects the average performance across all classes, treating each class equally. A high weighted average indicates that the performance is weighted by the number of instances in each class, accounting for class imbalance.

We also use five metrics (i.e., sensitivity, specificity, F1 score, AUC, and AUPRC) to analyze the overall performance (Table 5) of the model. The proposed method resulted in high values in all the metrics, indicating the model completely outperforms the results achieved by other methods. An F1 score of 0.9091 indicates that the model has a good balance between precision and recall. Specifically, it means that the model performs well in both identifying positive instances and minimizing false positives and false negatives. Moreover, a sensitivity (recall) of 0.9091 means that the model correctly identified 90.91% of all actual positive instances. This high value indicates that the model is very effective at detecting positive cases. The value of 0.9524 for specificity indicates that the model is effective at avoiding false positives and is reliable in identifying negative cases. An AUC close to 1.0 suggests that the model has a high degree of separability between the classes, and a high AUPRC shows that the model has good performance in scenarios where the positive class is rare or when dealing with class imbalance. This reinforces the fact that using a fully connected neural network model results in better performance overall.

From Figure 6, it can be seen that the proposed model is not only effective in shorter sampling periods but also in longer sampling periods without significant degradation of its performance. Compared to the other existing methods used on the same dataset, it is clear that the proposed model outperforms the conventional phase identification methods using the power consumption data. When the sampling period increases from 1 min to 30 min the accuracy is decreased only 9.9%. The available literature on phase identification using combined power consumption data with longer sampling periods is rare. This model shows excellent performance and opens new paths to use power consumption data which can be easily retrieved from smart meters for phase identification with practical longer sampling periods.

4.2. FCNN over Other Models

Architectures like CNNs and RNNs, including their variants, have the potential to capture temporal dependencies in data. However, our decision to use the FCNN was based on comparative results, as detailed below:

- Comparative Performance Analysis: We conducted an extensive analysis comparing various machine learning and neural network models, including both traditional algorithms and deep learning architectures, to assess their suitability for phase identification using the IEEE European Low Voltage Testing Feeder dataset. The results are presented in Table 7. These results indicate that while traditional methods and RNN variants demonstrated limited performance for this task, our FCNN model significantly outperformed all other approaches.
- Effectiveness for the Task: The FCNN showed substantial performance improvement, achieving 91.81% accuracy with 10-fold cross-validation. The high accuracy, combined with the simplicity of implementation, made the FCNN the most effective choice for our study.
- Model Complexity: RNNs, including GRUs and LSTMs, are designed for learning temporal dependencies and are more complex to train. Despite this, they did not yield satisfactory results (e.g., GRU achieving only 36%). The FCNN, on the other hand, was able to leverage the engineered features effectively without the complexity of sequence modeling, proving that temporal dependencies may not be as critical for this phase identification task as initially considered.
- Feature Engineering Impact: The use of data augmentation (GANs) and feature selection (RFE) with the FCNN was pivotal in achieving high accuracy. These enhancements allowed the FCNN to learn from the most relevant features and generalize well to the task at hand.

While our results indicate that the FCNN is highly effective for phase identification with low-sampling-rate data, we recognize the potential of exploring more complex architectures like CNNs and RNNs in future studies. This would help determine whether these architectures can further improve performance, particularly when applied with advanced data augmentation and feature engineering techniques.

4.3. Computational Cost of Scaling the Proposed Model

The computational demands of our models are crucial for assessing their practical applicability. In this study, we used the IEEE European Low Voltage Testing Feeder dataset, which is relatively small compared to larger, real-world datasets. The training and data augmentation were performed on standard hardware with a modest GPU, yielding efficient computation times. However, scaling this approach to larger datasets with more households would require more advanced hardware, such as GPUs or TPUs with increased memory and processing power. Additionally, GAN training is computationally intensive due to its adversarial nature. Efficient management of memory and resources will be critical to avoid bottlenecks as the dataset expands. To address these challenges, we plan to optimize the FCNN architecture by using lighter versions, applying model pruning, and exploring parallelized training across multiple GPUs or distributed computing frameworks. For GAN training, we are investigating more efficient architectures, such as conditional GANs and Wasserstein GANs, to improve convergence and computational efficiency.

4.4. Privacy, Security Aspects, and Potential Attacks of the Proposed Model

Given the importance of security in energy systems and machine learning applications, it is essential to address potential vulnerabilities that may arise from the use of machine learning models in grid management:

- Potential Security Attacks: Machine learning models, including our proposed approach, could be susceptible to various forms of cyberattacks that aim to compromise their performance. One concern is poisoning attacks, where an attacker manipulates training data to subvert the model's accuracy and behavior. This could lead to incorrect phase identification, undermining grid stability and reliability. Such attacks are detailed in studies like "Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare" [56], which explore the potential for adversarial manipulation of training data.
- Data Integrity and Privacy: Ensuring data integrity is crucial in the context of grid management. The security of data transmission, especially when augmented data are generated and shared, must be protected to prevent unauthorized access or tampering.
- Data Anonymization: We ensure that the dataset used, the IEEE European Low Voltage Testing Feeder, is anonymized, preventing the identification of individual households. This protects the privacy of users by ensuring that the model cannot infer personal data or habits from the data provided.
- Secure Data Handling: Our methodology relies on data preprocessing and feature extraction that are performed on aggregated data rather than individual-level data, reducing the risk of revealing personal information. In practice, the application of our model would involve secure data transmission and storage practices that align with data protection regulations (e.g., GDPR).
- Ethical Model Usage: We emphasize that the use of our model should be governed by clear ethical guidelines. The deployment of such a model in real-world scenarios must involve informed consent from participants and be limited to use cases where privacy risks are minimized. Our proposed approach could be adapted to ensure that sensitive information is not disclosed without proper safeguards and permissions.
- Security Measures: To mitigate these risks, it is essential to implement robust data verification processes, anomaly detection mechanisms, and regular model audits to identify and respond to potential attacks. Incorporating energy-efficient, long-term continuous monitoring systems, as discussed in publications like "Energy-Efficient Long-term Continuous Personal Health Monitoring" [57], could provide an additional layer of security by continuously monitoring for unusual data patterns and potential intrusions.

4.5. Limitations

We acknowledge several limitations in our study. First, the dataset used, the IEEE European Low Voltage Testing Feeder dataset, is comprehensive but may not fully represent the variations found in real-world grid configurations and household energy profiles. As a result, the model's applicability to other regions with different grid setups or consumption patterns remains uncertain. The dataset's limited availability of phase information also restricted the scope of our analysis to this particular dataset, which may not reflect broader conditions.

Second, although the use of GANs for data augmentation significantly improved model performance, we did not conduct a thorough analysis of the quality or potential biases in the GAN-generated data. This oversight may affect the generalizability of the model, particularly if the generated data contains inaccuracies or inherent biases.

Third, our study focused on data from a single day, which limits our ability to evaluate the model's performance over extended periods or across different seasons. The energy consumption patterns could vary significantly based on the time of day, week, or year, and examining this variability is crucial for assessing the model's robustness in real-world scenarios. Lastly, scaling the model to handle larger datasets was not fully explored. The computational cost associated with processing and training on larger datasets using FCNN and GANs could present challenges, and efficient techniques for resource management and optimization will be needed.

5. Conclusions and Future Work

In this article, we developed a fully connected neural network model designed to enhance the accuracy of phase identification using power consumption data from smart meters. The performance of our proposed model was rigorously compared with other established statistical and machine learning methods. The results demonstrate that our model consistently delivers highly accurate predictions, regardless of whether the sampling periods are short or extended. A significant advantage of our neural network model is its robustness in handling limited data. Despite the constraints of available datasets, particularly for phase identification with longer sampling periods, the model exhibits strong performance. This represents a valuable advancement, given the scarcity of research addressing phase identification using power consumption data with extended sampling intervals.

The architecture of our proposed model comprises an input layer, four hidden layers, and an output layer, all intricately connected to perform complex computations for precise phase prediction. Through 10-fold cross-validation, our model achieved an impressive accuracy rate of 91.81% and a high F1 score of 0.9591. These metrics surpass those of alternative methods evaluated on the same dataset.

A key achievement of this model is its ability to maintain performance even when data samples are reduced from 1 min intervals to 30 min intervals. The accuracy degradation remains under 10%, a notable improvement compared to other methods that experience significant accuracy loss with longer sampling periods.

Future work will focus on addressing the limitations identified in this study. We plan to extend the evaluation to more diverse datasets, especially those that cover different geographic regions and household types, to assess the generalizability of the model in real-world scenarios. Additionally, while the current model is effective, exploring more complex architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) could provide further improvements, particularly in capturing temporal dependencies in energy consumption patterns. Another important area of future research will involve investigating the quality of the data generated by GANs, ensuring that any biases or inaccuracies are addressed to maintain model performance and generalizability.

We also recognize the need to extend the evaluation to longer periods of data, across multiple days and seasons, to better understand how the model adapts to varying energy consumption patterns over time. Finally, the computational scalability of our approach remains an area of concern. As the model is scaled to handle larger datasets, exploring optimization techniques and resource-efficient strategies will be essential for ensuring its feasibility in real-world applications.

In summary, the proposed neural network model demonstrates exceptional capability in phase identification using limited power consumption data over extended periods.

Author Contributions: D.C.H. contributed to conceptualization, methodology, validation, writing—original draft preparation, writing—review and editing, and visualization; A.G. contributed to supervision, writing—review and editing; and A.Z.K. contributed to validation, supervision, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study are based on the IEEE European Low Voltage Test Feeder, which can be accessed at https://ewh.ieee.org/soc/pes/dsacom/testfeeders.html accessed on 9 March 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, C.S.; Ku, T.T.; Lin, C.H. Design of Phase Identification System to Support Three-Phase Loading Balance of Distribution Feeders. *IEEE Trans. Ind. Appl.* **2011**, *48*, 191–198. [CrossRef]
2. Wen, M.H.F.; Arghandeh, R.; von Meier, A.; Poolla, K.; Li, V.O.K. Phase Identification in Distribution Networks with Micro-Synchrophasors. In Proceedings of the 2015 IEEE Power & Energy Society General Meeting, Denver, CO, USA, 26–30 July 2015; pp. 1–5. [CrossRef]
3. Arya, V.; Seetharam, D.; Kalyanaraman, S.; Dontas, K.; Pavlovski, C.; Hoy, S.; Kalagnanam, J. Phase Identification in Smart Grids. In Proceedings of the 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm), Brussels, Belgium, 17–20 October 2011; pp. 25–30.
4. Heidari-Akhijahani, A.; Safdarian, A.; Aminifar, F. Phase Identification of Single-Phase Customers and PV Panels via Smart Meter Data. *IEEE Trans. Smart Grid* **2021**, *12*, 4543–4552. [CrossRef]
5. Arya, V.; Jayram, T.S.; Pal, S.; Kalyanaraman, S. Inferring Connectivity Model from Meter Measurements in Distribution Networks. In Proceedings of the e-Energy '13, Berkeley, CA, USA, 21–24 May 2013.
6. Pappu, S.J.; Bhatt, N.; Pasumarthy, R.; Rajeswaran, A. Identifying Topology of Low Voltage Distribution Networks Based on Smart Meter Data. *IEEE Trans. Smart Grid* **2018**, *9*, 5113–5122. [CrossRef]
7. Zhou, L.; Zhang, Y.; Liu, S.; Li, K.; Li, C.; Yi, Y.; Tang, J. Consumer Phase Identification in Low-Voltage Distribution Network Considering Vacant Users. *Int. J. Electr. Power Energy Syst.* **2020**, *121*, 106079. [CrossRef]
8. Seal, B.K.; McGranaghan, M.F. Automatic Identification of Service Phase for Electric Utility Customers. In Proceedings of the 2011 IEEE Power and Energy Society General Meeting, Detroit, MI, USA, 24–28 July 2011; pp. 1–3.
9. Pezeshki, H.; Wolfs, P. Correlation Based Method for Phase Identification in a Three Phase LV Distribution Network. In Proceedings of the 2012 22nd Australasian Universities Power Engineering Conference (AUPEC), Bali, Indonesia, 26–29 September 2012; pp. 1–7.
10. Pezeshki, H.; Wolfs, P.J. Consumer Phase Identification in a Three Phase Unbalanced LV Distribution Network. In Proceedings of the 2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), Berlin, Germany, 14–17 October 2012; pp. 1–7.
11. 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*, 1964–1971. [CrossRef]
12. Watson, J.D.; Welch, J.; Watson, N.R. Use of Smart-Meter Data to Determine Distribution System Topology. *J. Eng.* **2016**, *2016*, 94–101. [CrossRef]
13. Olivier, F.; Ernst, D.; Fonteneau, R. Automatic Phase Identification of Smart Meter Measurement Data. *CIRED-Open Access Proc. J.* **2017**, *2017*, 1579–1583. [CrossRef]
14. Mitra, R.; Kota, R.; Bandyopadhyay, S.; Arya, V.; Sullivan, B.; Mueller, R.; Labut, G. Voltage Correlations in Smart Meter Data. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 1999–2008.
15. Arya, V.; Mitra, R.; Mueller, R.; Storey, H.; Labut, G.; Esser, J.; Sullivan, B. Voltage Analytics to Infer Customer Phase. In Proceedings of the IEEE PES Innovative Smart Grid Technologies, Europe, Istanbul, Turkey, 12–15 October 2014; pp. 1–6.
16. Arya, V.; Mitra, R. Voltage-Based Clustering to Identify Connectivity Relationships in Distribution Networks. In Proceedings of the 2013 IEEE International Conference on Smart Grid Communications (SmartGridComm), Vancouver, BC, Canada, 21–24 October 2013; pp. 7–12.
17. Wang, W.; Yu, N.; Foggo, B.; Davis, J.; Li, J. Phase Identification in Electric Power Distribution Systems by Clustering of Smart Meter Data. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 259–265.
18. Simonovska, A.; Ochoa, L.F. Phase Grouping in PV-Rich LV Feeders: Smart Meter Data and Unconstrained K-Means. In Proceedings of the 2021 IEEE Madrid PowerTech, Madrid, Spain, 28 June–2 July 2021; pp. 1–6.

19. Overington, S.; Edwards, D.; Trinkl, P.; Buckley, A. Application of Constrained K-Means Algorithm for Phase Identification. In Proceedings of the 2021 31st Australasian Universities Power Engineering Conference (AUPEC), Perth, Australia, 26–30 September 2021; pp. 1–6. [CrossRef]
20. Wang, W.; Yu, N. Advanced Metering Infrastructure Data Driven Phase Identification in Smart Grid. In Proceedings of the GREEN 2017, Taipei, Taiwan, 21–24 December 2017.
21. Ma, Y.; Fan, X.; Tang, R.; Duan, P.; Sun, Y.; Du, J.; Duan, Q. Phase Identification of Smart Meters by Spectral Clustering. In Proceedings of the 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, China, 20–22 October 2018; pp. 1–5.
22. Liu, S.; Cui, X.; Lin, Z.; Lian, Z.; Lin, Z.; Wen, F.; Qiu, H. Practical Method for Mitigating Three-Phase Unbalance Based on Data-Driven User Phase Identification. *IEEE Trans. Power Syst.* **2020**, *35*, 1653–1656. [CrossRef]
23. Blakely, L.; Reno, M.J. Phase Identification Using Co-Association Matrix Ensemble Clustering. *IET Smart Grid* **2020**, *3*, 490–499. [CrossRef]
24. Blakely, L.; Reno, M.J.; Feng, W.C. Spectral Clustering for Customer Phase Identification Using AMI Voltage Timeseries. In Proceedings of the 2019 IEEE Power and Energy Conference at Illinois (PECI), Champaign, IL, USA, 28 February–1 March 2019; pp. 1–7.
25. Jayadev, P.S.; Rajeswaran, A.; Bhatt, N.P.; Pasumarthy, R. A Novel Approach for Phase Identification in Smart Grids Using Graph Theory and Principal Component Analysis. In Proceedings of the 2016 American Control Conference (ACC), Boston, MA, USA, 6–8 July 2016.
26. Xu, M.; Li, R.; Li, F. Phase Identification With Incomplete Data. *IEEE Trans. Smart Grid* **2018**, *9*, 2777–2785. [CrossRef]
27. Hosseini, Z.S.; Khodaei, A.; Paaso, A. Machine Learning-Enabled Distribution Network Phase Identification. *IEEE Trans. Power Syst.* **2020**, *36*, 842–850. [CrossRef]
28. 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 Proceedings of the 2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 29–31 July 2021; pp. 346–350. [CrossRef]
29. Xiong, Y.; Zhang, L.; Liu, J. Categorization of Power Consumption Patterns for User Screening: High, Normal, and Low Regions. *J. Energy Manag.* **2022**, *15*, 123–135.
30. Blakely, J.; Reno, M. Advanced Data Processing Techniques for Effective Phase Information Extraction in Electrical Systems. *IEEE Trans. Power Syst.* **2020**, *35*, 3050–3061.
31. Hoogsteyn, A.; Vanin, M.; Koirala, A.; Hertem, D. Low Voltage Customer Phase Identification Methods Based on Smart Meter Data. *arXiv* **2022**, arXiv:2204.06372. [CrossRef]
32. Gao, X.; Wang, Y.; Zhang, L. Advanced Data Analytics Techniques for Phase Identification in Power Distribution Systems. *IEEE Trans. Smart Grid* **2021**, *12*, 3456–3468.
33. Wang, Y.; Zhang, L.; Liu, J. Data-Driven Phase Identification in Power Distribution Systems Using Machine Learning Techniques. *IEEE Trans. Power Syst.* **2019**, *34*, 3456–3468.
34. Liao, Z.; Liu, Y.; Wang, B.; Tao, W. Topology Identification of Active Low-Voltage Distribution Network Based on Regression Analysis and Knowledge Reasoning. *Energies* **2024**, *17*, 1762. [CrossRef]
35. Yan, X.; Zhang, L.; Liu, J. Adaptability of Distribution Network Electrical Topology Identification Algorithms Based on Edge Computing. *IEEE Trans. Smart Grid* **2023**, *14*, 1234–1245. [CrossRef]
36. Doseděl, M.; Kopecny, L.; Kozovský, M.; Hnidka, J.; Havránek, Z. Detection of the Interturn Shorts of a Three-Phase Motor Using Artificial Intelligence Processing Vibration Data. In Proceedings of the 2022 20th International Conference on Mechatronics—Mechatronika (ME), Pilsen, Czech Republic, 7–9 December 2022. [CrossRef]
37. Yao, M.; Xie, W.; Mo, L. Short-Term Electricity Price Forecasting Based on BP Neural Network Optimized by SAPSO. *Energies* **2021**, *14*, 6514. [CrossRef]
38. Morais, L.; Castro, A. Competitive Autoassociative Neural Networks for Electrical Appliance Identification for Non-Intrusive Load Monitoring. *IEEE Access* **2019**, *7*, 111746–111755. [CrossRef]
39. Liu, Z.; Ye, B. Adaptive Inverse Control of Hydro Electric Unit Based on Wavelet Neural Networks. *Adv. Mater. Res.* **2012**, *591–593*, 1200–1203. [CrossRef]
40. Li, X.; Zhang, Y.; Wang, L. Impact of Electric Vehicle Charging on Power Distribution Networks: Challenges and Solutions. *IEEE Trans. Power Syst.* **2023**, *38*, 1234–1245.
41. Li, X.; Wang, Y.; Zhang, Z. Analyzing Long-Term and High Instantaneous Power Consumption of Buildings from Smart Meter Big Data with Deep Learning and Knowledge Graph Techniques. *IEEE Trans. Smart Grid* **2023**, *14*, 123–135.
42. Wang, Y.; Zhang, Z.; Liu, J. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. *IEEE Trans. Smart Grid* **2019**, *10*, 3125–3148. [CrossRef]
43. Foggo, B.; Yu, N. Improving Supervised Phase Identification Through the Theory of Information Losses. *IEEE Trans. Smart Grid* **2020**, *11*, 2337–2346. [CrossRef]

44. Wang, W.; Yu, N. Maximum Marginal Likelihood Estimation of Phase Connections in Power Distribution Systems. *IEEE Trans. Power Syst.* **2020**, *35*, 3906–3917. [CrossRef]
45. Wh, X.; Wang, Y.; Liu, J. Comprehensive Data Collection for Precise Phase Identification in Electrical Distribution Networks. *IEEE Trans. Smart Grid* **2023**, *14*, 123–135.
46. Cleenwerck, L. Factors Affecting Phase Identification in Electrical Distribution Networks. *IEEE Trans. Smart Grid* **2023**, *14*, 234–245.
47. IEEE PES Distribution Systems Analysis Subcommittee Radial Test Feeders. The IEEE European Low Voltage Test Feeder. Available online: http://ewh.ieee.org/soc/pes/dsacom/testfeeders.html (accessed on 5 June 2022).
48. Raja, A.; Shashidhar, K. Feature Selection Techniques for Machine Learning: A Survey. *Int. J. Inf. Technol.* **2023**, *15*, 1175–1185.
49. Zhang, L.; Liu, C.; Sun, Y. Recursive Feature Elimination and Support Vector Machine-Based Feature Selection Method for Fault Diagnosis in Mechanical Systems. *Mech. Syst. Signal Process.* **2023**, *174*, 109058.
50. Sharma, S.; Ghosh, S. A Hybrid Feature Selection Method for Intrusion Detection System Using Recursive Feature Elimination and Machine Learning. *J. Ambient. Intell. Humaniz. Comput.* **2018**, *9*, 1459–1469.
51. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
52. Tavares, M.A.; Nascimento, L.P. Data Augmentation for Machine Learning: A Review. *Comput. Mater. Contin.* **2022**, *70*, 3455–3480.
53. Brophy, E.; Wang, Z.; She, Q.; Ward, T. Generative Adversarial Networks in Time Series: A Systematic Literature Review. *ACM Comput. Surv.* **2023**, *55*, 199. [CrossRef]
54. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119. [CrossRef]
55. Zheng, N.; Li, Z.; Chen, Y.; Yang, X.-H.; Wei, Z.; Heidari, A.A.; Chen, H.; Hu, H.; Zhou, Q.; Guan, Q.; et al. Generative Adversarial Networks in Medical Image Augmentation: A Review. *Comput. Biol. Med.* **2022**, *144*, 105382. Available online: https://www.sciencedirect.com/science/article/pii/S0010482522001743 (accessed on 9 March 2024).
56. Mozaffari-Kermani, M.; Sur-Kolay, S.; Raghunathan, A.; Jha, N.K. Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1893–1905. [CrossRef]
57. Nia, A.; Mozaffari Kermani, M.; Sur Kolay, S.; Raghunathan, A.; Jha, N.K. Energy-Efficient Long-term Continuous Personal Health Monitoring. *IEEE Trans. Multi-Scale Comput. Syst.* **2015**, *1*, 85–98. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.