Electric Vehicle Identification Model Based on Net Load Decomposition and Two-Stage Decision

Shuxian Yi , Guowu Li , Saining Yin , Zezhong Wang , MA Xinsheng , Zhao Zhen *

Posted Date: 23 March 2026

doi: 10.20944/preprints202603.1670.v1

Keywords: distributed photovoltaics; electric vehicle identification; ensemble learning; net load decomposition; contextually supervised source separation

 doi:10.20944/preprints202603.1670.v1

Article

Electric Vehicle Identification Model Based on Net Load Decomposition and Two-stage Decision

Shuxian Yi [1], Guowu Li [2], Saining Yin [1], Zezhong Wang [2], MA Xinsheng [1] and Zhao Zhen [3,*]

[1] Electric Power Research Institute of State Grid Jibei Electric Power Co., Ltd, Beijing 100032, China

[2] State Grid Jibei Electric Power Co., Ltd, Beijing 100052, China

[3] Department of Power Engineering, North China Electric Power University, Baoding 071003, China

* Correspondence: georgiazhz@foxmail.com

Abstract

To address the challenges of identifying electric vehicles (EVs) in user-side scenarios where multi-source load data is coupled by high-penetration distributed photovoltaics (PV), this paper proposes a robust EV identification framework based on net load decomposition and a two-stage decision-making process. Initially, a context-aware source-supervised separation (CSSS) algorithm is employed to decouple PV output from the net load, effectively eliminating PV-induced interference by constructing targeted feature vectors. Subsequently, four key features characterizing EV charging behavior are extracted to feed into a hierarchical identification model. The first stage utilizes a Composite Charging Characteristic Index (CCCI) for rapid preliminary screening, while the second stage implements sample-adaptive weighted Stacking ensemble learning for high-precision detection. Experimental results demonstrate that the proposed method achieves an identification accuracy of 96.33%, with the load decomposition stage contributing a 1.2% improvement. This framework provides a reliable technical foundation for load analysis and demand-side management in distribution networks with high PV integration.

Keywords: distributed photovoltaics; electric vehicle identification; ensemble learning; net load decomposition; contextually supervised source separation

1. Introduction

With the profound transformation of the global energy structure and the continuous advancement of "carbon peak and carbon neutrality" strategies, user-side distributed resources, represented by distributed photovoltaics (PV) and electric vehicles (EVs), are being integrated into the power grid at an unprecedented scale and speed [1–4]. While these resources optimize the energy structure and enhance system flexibility, they also present new challenges for the accurate perception, operational control, and resource coordination of the power grid [5, 6]. In the context of "New Power Systems," the traditional passive distribution network is evolving into an active, bidirectional energy ecosystem where edge intelligence plays a crucial role [7].

In the refined management and intelligent operation of distribution networks, achieving observability, measurability, and controllability of various distributed resources is a fundamental prerequisite [3, 8]. For EVs specifically, as a significant flexible load resource, their accurate identification and perception directly influence the precision of charging load forecasting, guided orderly charging, demand response strategy formulation, and the exploration of vehicle-to-grid (V2G) interaction potential [9–11]. Accurately identifying whether a user owns an EV and evaluating their charging behavior patterns helps grid operators assess the impact of charging facility integration, optimize charging guidance strategies, and provide basic data support for EV participation in ancillary service markets [5, 12].

 doi:10.20944/preprints202603.1670.v1

However, most distributed energy resources operate in "Behind-the-Meter" (BTM) configurations, meaning grid operators can only access the aggregated "net load" power. This creates a significant "observability gap," as the temporal coupling and morphological masking between PV generation and EV charging sessions severely obstruct accurate feature extraction. Such interference limits the efficacy of traditional management and optimization strategies for demand-side resources [13, 14].

Currently, research on EV identification has made some progress. Existing methods can be categorized into three types based on their core technical paths:

Load Disaggregation-based Methods: These methods focus on separating the EV charging component from the total user load signal. For instance, some studies utilize the low-frequency characteristics of EV charging loads for non-intrusive extraction [15], or use combinations of event detection, state transition removal, and sequence matching for online identification [16]. While these focus on signal-level stripping, they often assume load components are relatively independent; when processing net loads with high PV penetration, the inability to effectively separate the PV component leads to the masking of EV features.

Feature Extraction-based Methods: These methods extract key features reflecting EV charging behavior and analyze them through rules or statistical methods. Examples include knowledge-driven feature extraction via pre-defined rules, which offer high interpretability but suffer from limited adaptability [17, 18]. Other techniques utilize two-stage decomposition to extract features from low-frequency data or apply independent component analysis (ICA) with multi-level pattern matching to extract three-stage EV charging features [19]. These methods require high data quality and feature stability, and their performance degrades when PV fluctuations mask the characteristics.

Data-Driven Methods: These techniques leverage machine learning and deep learning to learn charging behavior patterns from data. Research includes fusion-based seasonal analysis to detect EV households and the use of Generative Adversarial Networks (GANs) or Convolutional Neural Networks (CNNs) for regional load identification [20, 21]. Furthermore, the Viterbi algorithm and hidden Markov models have been used to improve steady- and transient-state mixed feature extraction [22]. Despite their strong pattern learning capabilities, these methods are prone to overfitting or insufficient generalization in multi-source coupled scenarios.

In summary, existing methods perform well in single-load or known PV output scenarios but face significant challenges in complex scenarios with high-penetration distributed PV. During sunrise and sunset, PV output and EV charging loads overlap temporally, causing confusion in load curve morphology and interfering with feature extraction. Furthermore, most methods rely on prior information like meteorological data or PV capacity, which are costly or difficult to obtain in practice. Most importantly, existing research often focuses on a single stage of decomposition or identification, lacking a complete framework from net load decomposition to EV identification.

To address this, this paper proposes an EV identification model based on net load decomposition and two-stage determination. First, by constructing PV and base load feature vectors and utilizing a context-aware source-supervised separation algorithm, the PV output is effectively separated from the net load. Subsequently, feature engineering is performed on EV charging behavior, which is then combined with a data-driven ensemble learning model to construct a two-stage screening and identification model. This method establishes a complete "decomposition-identification" analysis framework, providing technical support for the reliable identification of EVs in distribution networks with high proportions of distributed PV.

2. Problem Formulation and Research Framework

For a specific user cluster, the time window $[0,1,2,...,T]$ is defined as the set of periods within a day. Smart meters measure the aggregated net load power $\boldsymbol{p}_{net}^{d} = [p_{net}^{d,1}, p_{net}^{d,2}, p_{net}^{d,3} ..., p_{net}^{d,T}]$ of the user cluster at time intervals Δt (in this study, Δt =1 hour) on day d, which represents the power purchased from the grid. The relationship between the net load, distributed PV output, and the base load of the user cluster is expressed as follows:

$$\boldsymbol{p}_{net}^{d} = \boldsymbol{p}_{load}^{d} - \boldsymbol{p}_{pv}^{d} \tag{1}$$

 doi:10.20944/preprints202603.1670.v1

Where $\boldsymbol{p}_{load}^{d} = [p_{load}^{d,1}, p_{load}^{d,2}, p_{load}^{d,3} \ldots, p_{load}^{d,T}]$ denotes the base load of the user cluster on day d, which includes both conventional appliances and the electric vehicle (EV) charging loads to be identified. $\boldsymbol{p}_{pv}^{d} = [p_{load}^{d,1}, p_{load}^{d,2}, p_{load}^{d,3} \ldots, p_{load}^{d,T}]$ represents the output of the distributed PV systems installed on the user side on day d; if no distributed PV is installed in the cluster, $\boldsymbol{p}_{pv}^{d} = 0$. Since distributed energy storage for residential users is currently not widely deployed, this paper focuses on three primary components: distributed PV power, conventional power load, and EV charging load.

The overall research workflow is illustrated in Figure 1. The EV identification process comprises three main phases: the load decomposition stage, the feature extraction stage, and the two-stage identification process. First, a modified load curve $\boldsymbol{p}_{load}^{d}$, which contains only the base load and EV charging components, is obtained through load decomposition. Subsequently, a four-dimensional feature vector characterizing EV charging behavior is constructed. Finally, the target user cluster is analyzed through rapid initial screening and refined identification to determine the presence of EVs.

The remainder of this paper is organized as follows: Section 2 describes the load decomposition method based on the context-aware source-supervised separation (CSSS) algorithm, and Section 3 details the two-stage identification model based on EV feature vector extraction.

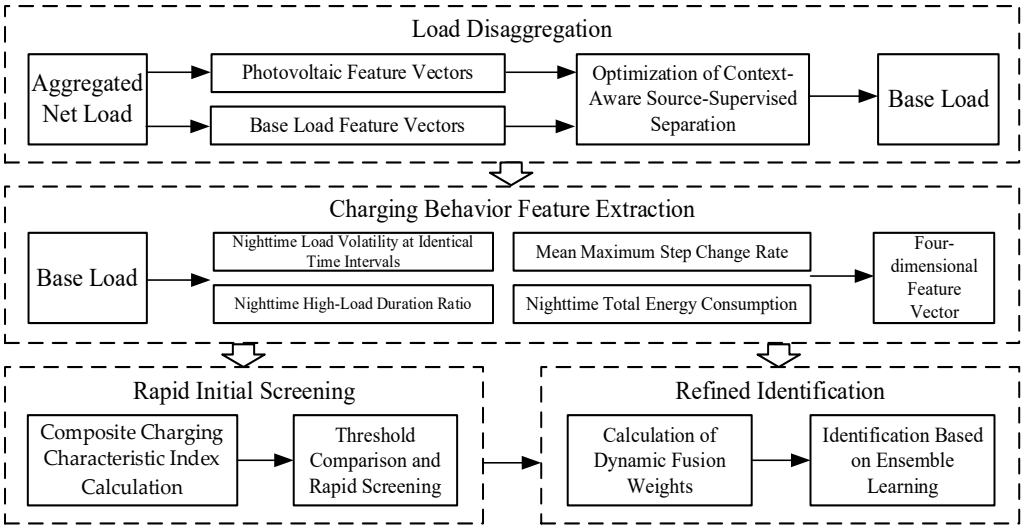

Figure 1. The proposed EV identification framework integrating CSSS-based decomposition and two-stage decision logic.

3. Load Disaggregation Based on the Context-Aware Source-Supervised Separation Algorithm

The primary objective of load disaggregation using the Context-aware Source-supervised Separation (CSSS) algorithm is to decouple the PV generation profile from the composite net load signal. By effectively stripping away the solar component under known net load conditions, this method generates a "refined" modified load curve—one that exclusively encapsulates the base load and electric vehicle (EV) charging components. This process serves as a critical preprocessing step, ensuring a robust and reliable data foundation for the subsequent stages of EV identification and behavior analysis.

3.1. Context-Aware Source-Supervised Separation Algorithm

For a target user cluster (equipped with distributed PV), the aggregated net load on day *d* can be modeled as a linear combination of its constituent components. Specifically, the PV generation output and the base load (excluding PV) are represented as the product of their respective feature vectors—characterizing the fundamental daily fluctuation patterns (profiles)—and their corresponding aggregation coefficients (amplitudes):

$$\boldsymbol{p}_{pv}^{d} \approx \theta_{pv}^{d} \cdot \boldsymbol{x}_{pv}^{d} \tag{2}$$

 doi:10.20944/preprints202603.1670.v1

$$\boldsymbol{p}_{load}^{d} \approx \theta_{load}^{d} \cdot \boldsymbol{x}_{load}^{d} \tag{3}$$

Where x_{pv}^{d} and x_{load}^{d} denote the feature vectors for distributed PV and base load, respectively; θ_{pv}^{d} and θ_{load}^{d} are the corresponding coefficients to be solved, representing the aggregation scaling ratios of the target cluster's PV output and base load relative to their respective feature vectors[23].

Since load disaggregation is inherently an ex-post analysis and processing of consumer load data, feature vectors for both distributed PV and the base load can be constructed based on known observation data. By optimizing the scaling coefficients θ_{pv}^{d} and θ_{load}^{d}, the linear representation is tuned to approximate the actual values as closely as possible while simultaneously satisfying the energy balance constraint.

The formulations allow the disaggregation process to be treated as an optimization problem that minimizes the sum of squared errors while adhering to physical energy balance constraints:

$$\min_{\theta_{pv},\theta_{\text{load}}} \frac{1}{2}\left(\| \boldsymbol{p}_{pv}^{d} - \theta_{pv}^{d}\boldsymbol{x}_{pv}^{d} \|_2^2 + \| \boldsymbol{p}_{load}^{d} - \theta_{load}^{d}\boldsymbol{x}_{load}^{d} \|_2^2\right) \tag{4}$$

$$s.t.\ \boldsymbol{p}_{net}^{d} = \boldsymbol{p}_{load}^{d} - \boldsymbol{p}_{pv}^{d} \tag{5}$$

$$\boldsymbol{p}_{load}^{d} \geq 0 \tag{6}$$

$$\boldsymbol{p}_{pv}^{d} \geq 0 \tag{7}$$

Specifically, the objective function aims to minimize the sum of squared errors (SSE) of the PV and base load estimations. Equality constraints are employed to ensure that the sum of the disaggregated components remains equal to the net load measured by the smart meter, while non-negativity constraints are applied to both PV generation and base load power.

The optimization model described above can be solved efficiently using the Gurobi solver. It is worth noting that disaggregating the signals becomes significantly more challenging when the two original sources are highly correlated. Fortunately, in most practical scenarios, there is minimal correlation between the base load and solar generation, enabling accurate PV and load disaggregation via the Context-aware Source-supervised Separation (CSSS) algorithm. Theoretically, the accuracy of the estimation improves as the correlation between the feature vectors and the actual physical values increases. Consequently, the construction of feature vectors is a pivotal phase that dictates the precision of the entire disaggregation process. The methodology for constructing these feature vectors is detailed in the following section.

3.2. Photovoltaic Feature Vectors

The output power of distributed photovoltaic (PV) systems is highly correlated with meteorological conditions and specific system parameters. Among these, solar irradiance, installed capacity, tilt angle, and azimuth are the most critical factors influencing power generation. Given that residential rooftop PV systems are typically small-scale, they often lack dedicated monitoring equipment to record technical parameters such as tilt and azimuth angles.

Although individual parameters—including capacity and installation tilt—may vary between systems, meteorological factors such as temperature and humidity remain approximately uniform within the same geographical region. These factors exert a consistent influence on the PV output across different users. Furthermore, to maximize energy yields and economic returns, most users adopt installation configurations that optimize irradiance based on the local latitude.

Consequently, assuming a consistent latitude within a given power supply area, these distributed residential PV systems can be expected to exhibit similar installation orientations and, by extension, analogous power generation patterns. Given that the total capacity of a user cluster is often unknown in practical scenarios, the PV feature vector can be derived using the following formulation:

$$\boldsymbol{x}_{pv}^{d} = \boldsymbol{I}_{irr}^{d}[\cos(90-\Theta)\sin(\beta)\cos(\phi-\alpha) + \sin(90-\Theta)\cos(\beta)] \tag{8}$$

Where I_{irr}^{d} denotes the irradiance vector received by the PV panel throughout day d; Θ represents the solar zenith angle, which varies temporally; β is the installation tilt angle of the PV panel; α signifies the solar azimuth angle; ϕ denotes the azimuth angle of the PV panel.

 doi:10.20944/preprints202603.1670.v1

3.3. Base Load Feature Vectors

The load behavior of an individual user is characterized by strong randomness and significant volatility. However, at the cluster level, the superposition of loads from a large number of users causes the aggregated load curve to become smoother and more regular, manifesting a typical daily load pattern. Consequently, by selecting a sufficient number of non-PV users within a specific region—whose net load is identical to their base load—their aggregated net load curve can serve as the feature vector for the base load. Assuming there are N_l non-PV users, the base load feature vector can be expressed as:

$$\boldsymbol{x}_{load}^{d} = \sum_{i=1}^{N_l} \boldsymbol{p}_{net,i}^{d} \tag{9}$$

Where $p_{net,i}^{d}$ denotes the net load vector of the i-th user cluster on day d.

4. Electric Vehicle Identification Considering Distributed Photovoltaics

After completing the photovoltaic (PV) output separation, the proposed identification method first extracts four key features of electric vehicle (EV) charging behavior from the base load curve to construct the identification feature vector. Subsequently, a two-stage identification framework is proposed:

Stage 1: Rapid Initial Screening. This stage is based on the Composite Charging Characteristic Index (CCCI) to achieve quick preliminary filtering, effectively removing typical non-EV user clusters from the dataset.

Stage 2: Refined Identification. This stage introduces a sample-adaptive weighted Stacking ensemble learning method. By integrating the complementary strengths of multiple machine learning models, it achieves high-precision identification of EV presence within the remaining clusters.

This "decomposition-identification" analysis framework ensures reliable EV detection even in distribution networks characterized by high proportions of distributed PV.

4.1. Charging Behavior Feature Extraction

Following the application of the Context-aware Source-supervised Separation (CSSS) algorithm, the impact of distributed PV generation is minimized, providing a reliable "modified" base load curve as input for precise electric vehicle (EV) identification. Based on this, a feature extraction scheme specifically designed for EV charging behavior is developed.

Guided by residential EV charging patterns and existing research, the feature variables in this study are designed based on the following assumptions[24]:

Concentrated Nighttime Charging: Most residents prefer charging their EVs during nighttime off-peak tariff periods or upon returning home from work. Consequently, the EV charging time window is defined as $\delta_{ev} = [t_s, 24] \bigcup [1, t_e]$, typically set with $t_s = 18:00$ and $t_e = 06:00$. Notably, during the start and end of this window, EV charging loads may partially couple with PV generation, creating significant interference due to the high volatility of both power signals.

Dominance of G2V Mode: Under current technical conditions, residential EV interaction is dominated by the Grid-to-Vehicle (G2V) charging mode. As Vehicle-to-Grid (V2G) discharge modes are not yet widely commercialized at the household level, this feature extraction scheme focuses exclusively on charging load characteristics.

Intermittency of Charging Behavior: EV charging is typically not a daily occurrence. The charging start time, duration, and energy consumption exhibit inherent randomness and variance between different user sessions.

By synthesizing these behavioral patterns, four categories of key features are extracted to effectively capture the presence of EV charging loads within the disaggregated base load signal.

4.1.1. Nighttime Load Volatility at Identical Time Intervals

Electric vehicle charging events do not occur on a daily basis, and even when they do occur, the start times and durations vary significantly. This inherent stochasticity causes the load values at the same time point within the charging window to exhibit pronounced fluctuations across different

 doi:10.20944/preprints202603.1670.v1

days. In contrast, the nighttime base load of user clusters without EVs typically remains relatively stable.

To capture this characteristic, an empirical time point characterized by typically high charging power within the defined charging window is selected. The standard deviation of the load at this specific time is then calculated over a one-month period (n days):

$$F_1 = \sqrt{\frac{\sum_{d=1}^{n} (p_{load}^{d}(t_c) - \overline{p_{load}^{d}}(t_c))^2}{n}} \tag{10}$$

Where $\overline{p_{load}^{d}}(t_c)$ represents the monthly average load value at time t_c. A higher F_1 value indicates more intense load fluctuations at that specific time point, suggesting a higher probability of the presence of intermittent, high-power EV charging loads.

4.1.2. Mean Maximum Step Change Rate

When an electric vehicle (EV) initiates a charging session, its power demand typically surges from zero to several kilowatts within a very short interval. This magnitude of power transition is significantly greater than the load steps caused by the switching of conventional household appliances. Consequently, this feature is designed to capture the most intense instantaneous variations in the load curve.

First, the first-order difference sequence of adjacent time points in the daily load curve is calculated:

$$\Delta p_{load}^{d}(t) = p_{load}^{d}(t+1) - p_{load}^{d}(t) \tag{11}$$

Next, the maximum value within the daily difference sequence—representing the maximum step change for that day—is identified. The mean value of these maximums over a period of n days is then calculated:

$$F_2 = \frac{1}{n} \sum_{d=1}^{n} \max_{1 \le t \le T-1} (\Delta p_{load}^{d}(t)) \tag{12}$$

A higher F_2 value indicates the frequent occurrence of large-amplitude power jumps within the load curve, serving as a strong indicator signal for the presence of EV charging behavior.

4.1.3. Nighttime High-Load Duration Ratio

Electric vehicle (EV) charging is characterized by high power consumption. Once a charging session begins, it typically maintains the household nighttime load at an elevated level for a significant duration, thereby extending the "peak" duration of the load curve. Consequently, the ratio of high-load sampling points to the total number of sampling points within the charging window is selected as an identification feature, calculated as follows:

$$S^{high} = \{ p_{load}^{d}(t) \mid t \in \delta_{ev}, p_{load}^{d}(t) > p_{threshold} \} \tag{13}$$

$$F_3 = card(S^{high}) / card(S) \tag{14}$$

Where $p_{threshold}$ is a power threshold used to distinguish the conventional base load from the high-load state caused by EV charging. It serves as one of the hyperparameters to be optimized during the model training phase. $card(S^{high})$ represents the number of sampling points that satisfy the condition $p_{load}^{d}(t) > p_{threshold}$ within the defined window. $card(S)$ represents the total number of sampling points within a day (or the designated charging period).

A higher F_3 value indicates that the nighttime load remains at an elevated level for a longer period, which is more likely to be the result of sustained EV charging.

4.1.4. Nighttime Total Energy Consumption

EV charging sessions typically involve a substantial amount of energy transfer. Even when accounting for the intermittency of charging behavior, user clusters with EVs exhibit significantly higher total energy consumption during nighttime hours compared to clusters without EVs.

 doi:10.20944/preprints202603.1670.v1

Consequently, the total energy consumption can be determined by integrating the base load over the designated charging window:

$$F_4 = \sum_{t=t_s}^{t_e} p_{load}^{d}(t)\Delta t \tag{15}$$

Where Δt represents the sampling interval. Utilizing F_4 as a feature provides a direct reflection of the absolute level of nighttime energy consumption, serving as an effective indicator for distinguishing between user groups with or without high-capacity power-consuming devices such as EVs.

4.2. Two-Stage Electric Vehicle Identification

4.2.1. Stage 1: Rapid Initial Screening Based on the Composite Charging Characteristic Index (CCCI)

To achieve efficient and reliable electric vehicle (EV) identification, this study first develops a rapid initial screening mechanism based on the Composite Charging Characteristic Index (CCCI). The core of this mechanism involves fusing the four previously extracted EV charging features into a single quantitative metric. This index is designed to comprehensively reflect the overall similarity between a user's load profile and typical EV charging patterns, thereby enabling efficient preliminary filtering of large-scale user datasets without requiring complex model computations. The calculation of the CCCI is formulated as follows:

$$CCCI = \frac{\alpha \cdot \frac{F_1}{F_1^{\text{base}}} + \beta \cdot \frac{F_2}{F_2^{\text{base}}} + \gamma \cdot \frac{F_3}{F_3^{\text{base}}} + \delta \cdot \frac{F_4}{F_4^{\text{base}}}}{\alpha + \beta + \gamma + \delta} \tag{16}$$

Where F_i^{base} denotes the benchmark value of the *i*-th feature, derived from the non-EV user clusters in the training set. It represents the typical baseline level achievable by non-EV users for that specific feature. α , β , γ , δ are the fusion weights for each feature, satisfying $\alpha + \beta + \gamma + \delta = 1$. The determination of these weights is based on the discriminative capability of each feature within the training set. Specifically, the Information Gain Ratio (IGR)—a classic core metric for feature selection in decision tree algorithms—is utilized for quantification and then normalized to obtain the final weights[25]. The CCCI characterizes the overall magnitude by which a user cluster exceeds the baseline levels of ordinary user clusters across all charging-related features. A high CCCI value indicates that the user cluster's load pattern deviates significantly from typical household profiles across multiple dimensions, aligning more closely with the "high volatility, large step changes, long high-load duration, and high energy consumption" characteristics introduced by EV charging.

Based on the CCCI, the following rapid screening rule is implemented: a low threshold θ_L is established. If $CCCI < \theta_L$, it indicates that the composite charging characteristics of the cluster are indistinguishable from those of ordinary users. Such clusters are directly classified as "Non-EV User Clusters" with high confidence, eliminating the need for subsequent refined identification.

4.2.2. Stage 2: Refined Identification Based on Ensemble Learning

The ensemble learning methodology adopted in this study is Stacking. Within this architecture, the first-stage models are referred to as base learners (or primary learners), while the second-stage model is designated as the meta-learner. The fundamental approach begins by partitioning the original dataset into a training set and a test set according to an appropriate ratio. Subsequently, on the balanced training set D, multiple base learners are trained using the K-fold cross-validation method. The classification outputs from these base learners, paired with the original labels of D, serve as the input features and target labels for the meta-learner, respectively, forming a new training set to optimize the meta-model. The final classification result is generated by the meta-learner, as illustrated in Figure 2.

Traditional Stacking algorithms typically utilize fixed weights for all test samples during the fusion of base learners, thereby neglecting the distributional differences of various samples within the feature space. To address this limitation, this paper introduces a sample-adaptive weighting mechanism to dynamically adjust the fusion weights of individual base learners. Furthermore, three

 doi:10.20944/preprints202603.1670.v1

machine learning algorithms with complementary characteristics are selected as base learners: Gradient Boosting Decision Tree (GBDT), Support Vector Machine (SVM), and Random Forest (RF). All base learners utilize the four-dimensional feature vector $[F_1, F_2, F_3, F_4]$ extracted in the previous stage as their input.

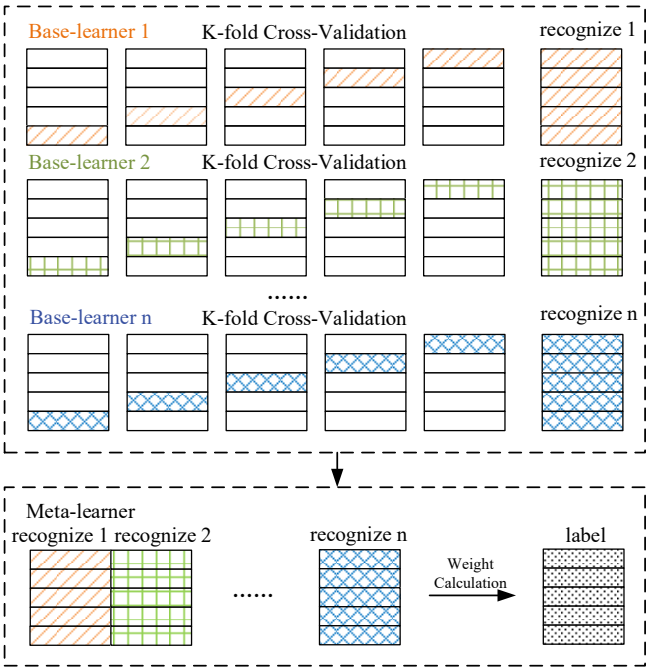

Figure 2. The approach of the Stacking algorithm.

For each test sample $\boldsymbol{x}_i$ to be identified, the K-nearest neighbors, denoted as $\mathcal{N}_K(\boldsymbol{x}_i)$, are identified within the training set. The normalized Euclidean distance is employed as the similarity metric to account for differences in feature scales. Subsequently, the identification accuracy of each base learner within this local neighborhood is calculated as follows:

$$Acc_m\left(\mathcal{N}_K\left(\boldsymbol{x}_i\right)\right)=\frac{1}{K}\sum_{\boldsymbol{x}_j\in\mathcal{N}_K(\boldsymbol{x}_i)}\mathbb{I}\left(y_j=\hat{y}_j^{(m)}\right) \tag{17}$$

Where y_j represents the ground-truth label of sample $\boldsymbol{x}_i$; $\hat{y}_j^{(m)}$ denotes the estimated label assigned to sample $\boldsymbol{x}_i$ by the *m*-th base learner.

Based on the local identification accuracy of each base learner, the dynamic fusion weights are calculated as follows:

$$w_m\left(\boldsymbol{x}_i\right)=\frac{\exp\left(\alpha\cdot Acc_m\left(\mathcal{N}_K\left(\boldsymbol{x}_i\right)\right)\right)}{\sum_{j=1}^{3}\exp\left(\alpha\cdot Acc_j\left(\mathcal{N}_K\left(\boldsymbol{x}_i\right)\right)\right)} \tag{18}$$

Where α is the sharpening parameter, used to regulate the concentration of the dynamic weight distribution across the base learners. When α is assigned a larger value, models with higher local accuracy receive significantly higher weights; conversely, when α is small, the weight distribution tends toward uniformity. In this study, α is optimized on the training set through cross-validation to strike a balance between the model's local sensitivity and its generalization ability.

The final output of the model is determined using a weighted voting strategy:

$$Vote\left(\boldsymbol{x}_i\right)=sign\left(\sum_{m=1}^{3}w_m\left(\boldsymbol{x}_i\right)\cdot\left(2\hat{y}_i^{(m)}-1\right)\right) \tag{19}$$

Where $Vote\left(\boldsymbol{x}_i\right)$ represents the weighted decision score (or aggregated result). The classification logic is defined as follows: if $Vote\left(\boldsymbol{x}_i\right)>0$, the target user cluster is identified as containing electric vehicles; conversely, if $Vote\left(\boldsymbol{x}_i\right)<0$, the cluster is identified as not containing electric vehicles.

 doi:10.20944/preprints202603.1670.v1

5. Case Study

5.1. Data Sources and Experimental Configuration

The dataset utilized in this study is derived from open-source load and distributed PV datasets provided by a public utility company in the U.S. Midwest, integrated with the Residential Energy Consumption Survey (RECS) dataset [26, 27]. The data spans from 1 January 2017 to 31 December 2017, with a temporal resolution of 1 hour. A total of 1120 residential users are included, each featuring comprehensive metered data—including net load, PV generation power, EV charging power, and base load. This high-fidelity dataset allows for a rigorous performance evaluation of the proposed identification model based on ground-truth observations.

To validate the effectiveness of the proposed model, an experimental sample set was constructed using random sampling with replacement. Specifically, 300 independent sampling trials were conducted; in each trial, 50 to 150 users were randomly selected to form a user cluster, resulting in 300 cluster samples of varying scales. This experimental design is intended to simulate the inherent diversity of user aggregation in real-world distribution networks and to assess the model's stability and generalization capability across different cluster sizes.

Hyperparameters for all models (including the benchmark models) were uniformly determined using Bayesian optimization during the training phase. All simulations were performed on a computing platform equipped with a 2.9 GHz Intel Core i5 processor and 8 GB of RAM. The optimization problems within the CSSS algorithm were solved using the Gurobi 9.1.2 solver.

5.2. Evaluation Metrics

A confusion matrix, also referred to as an error matrix, is a standard format for accuracy assessment in classification tasks. It organizes actual versus predicted classification results into an n×n grid, where n corresponds to the total number of categories. In the context of binary classification, outcomes are categorized as Positive and Negative. The structure and terminology of the confusion matrix are presented in Table 1. Based on these fundamental parameters, this study utilizes Equations (20)–(22) as the primary metrics for performance evaluation.

Table 1. Binary classification confusion matrix.

	Actual: Positive	Actual: Negative
Predicted: Positive	True Positive (TP)	False Positive (FP)
Predicted: Negative	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{20}$$

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

Furthermore, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) are selected to evaluate the performance of the load disaggregation process, as formulated in Equations (23)–(25):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{23}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{24}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{25}$$

Where y_i and $\hat{y}_i$ represent the ground-truth values and the estimated values of the disaggregated components, respectively, and N denotes the total number of samples.

5.3. Simulation Results and Discussion

 doi:10.20944/preprints202603.1670.v1

5.3.1. Load Disaggregation Results

In this study, the base load profile is derived through load disaggregation, serving as the foundational input for subsequent electric vehicle (EV) feature extraction and identification. Figure 3 illustrates the net load curve, the ground-truth base load curve, and the estimated base load curve obtained via the CSSS algorithm for a representative user cluster over a one-week duration.

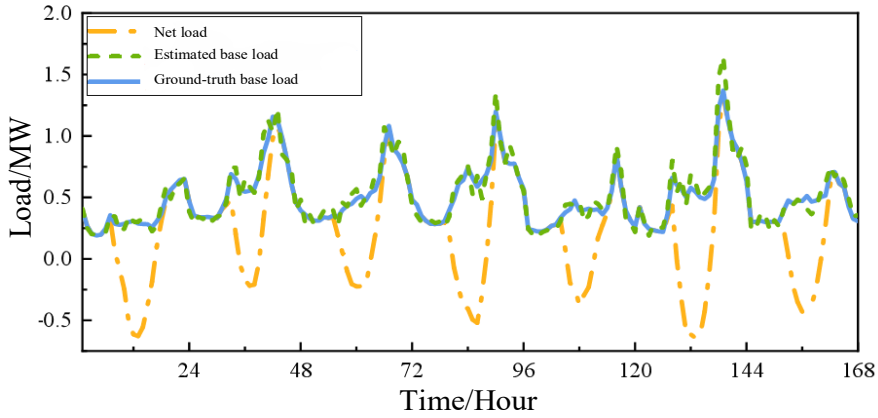

Figure 3. Comparison of load curves within a week.

The quantitative evaluation of the estimated base load yields an MAE of 0.1080 kW, an RMSE of 0.1602 kW, and a MAPE of 6.7%. These low error metrics demonstrate that the CSSS algorithm achieves high-precision separation of PV output, ensuring that the restored base load profile possesses high fidelity and credibility.

As observed, the disaggregated base load curve exhibits high alignment with the ground-truth curve in both profile morphology and magnitude. Notably, during daytime periods characterized by peak PV generation, the estimated curve accurately tracks the fluctuations of the actual base load. This indicates that the CSSS algorithm effectively decouples the stochasticity and intermittency inherent in PV generation, thereby restoring "purer" load features. This high-quality restoration provides a robust data foundation for the accurate extraction and identification of EV behavioral characteristics in the subsequent stages.

5.3.2. Comparison of Electric Vehicle Identification Results

To determine the fusion weights (α , β , γ , δ) for each feature within the Composite Charging Characteristic Index (CCCI), the Information Gain Ratio (IGR) was utilized to quantitatively evaluate the four categories of features extracted in Section 4.1. These evaluations, based on the training dataset, were subsequently normalized to derive the final fusion weights, as summarized in Table 2.

Table 2. Normalized fusion weights for EV charging features derived via Information Gain Ratio (IGR).

Feature	Description	Information Gain Ratio	Fusion Weight
F_1	Nighttime Load Volatility at Identical Time Intervals	0.152	0.18
F_2	Mean Maximum Step Change Rate	0.235	0.28
F_3	Nighttime High-Load Duration Ratio	0.201	0.24
F_4	Nighttime Total Energy Consumption	0.252	0.30

During the Stage 1 rapid initial screening, 31 user clusters without EVs were successfully identified through the CCCI calculation. These samples exhibited load patterns highly consistent with those of conventional households; therefore, they were directly classified as "Non-EV User Clusters" with high confidence. This allowed these samples to bypass the subsequent, more computationally intensive ensemble learning stage. Notably, the number of missed detections (false negatives) among

 doi:10.20944/preprints202603.1670.v1

these 31 clusters was zero, confirming the high reliability and robustness of the screening mechanism in preserving the integrity of the positive sample pool for Stage 2.

In the second stage, the sample-adaptive weighted ensemble learning model performs refined identification on the remaining user clusters. Table 2 presents the confusion matrix for the final identification results.

As observed from the matrix, the proposed method demonstrates significant identification performance for the "EV-containing" category. Only 7 cases of missed detections (False Negatives, FN) occurred, indicating that the model possesses high sensitivity to the presence of electric vehicles. Simultaneously, only 4 cases of false alarms (False Positives, FP) were recorded, suggesting that the model also exhibits strong specificity when identifying "non-EV clusters." Overall, the confusion matrix exhibits a pronounced diagonal dominance, signifying the robust and balanced identification capability of the model across both classification categories.

Table 3. Electric vehicle identification results.

	Predicted: EV-containing	Predicted: Non-EV
Actual: EV-containing	183	7
Actual: Non-EV	4	106

To evaluate the superiority of the proposed two-stage electric vehicle (EV) identification model, comparative experiments were conducted against several mainstream identification approaches. These benchmarks include individual classification models (SVM), ensemble learning models (GBDT, RF, and Traditional Stacking), and deep learning architectures (MLP and CNN-LSTM). All models were evaluated using identical training and testing datasets. The identification performance was assessed based on three metrics: Accuracy, Precision, and Recall, with the results illustrated in Figure 4.

As demonstrated, the proposed method outperforms all benchmark models across all three evaluation metrics:

Accuracy: The proposed method achieves an accuracy of 96.33%, indicating superior comprehensive identification capability in the overall classification task.

Precision: The performance is particularly outstanding in terms of precision, reaching 97.86%. This significantly surpasses the benchmark methods, demonstrating a remarkably low false-positive rate when designating a user cluster as "EV-containing." Such high decision confidence is crucial in practical applications, as it avoids the misidentification of non-EV users, thereby enhancing the targeted nature of grid dispatch and demand response strategies.

Recall: The recall rate of the proposed method is 96.32%, which is also superior to the other approaches. This indicates that the model can effectively identify the majority of actual EV-owning user clusters. This strong coverage capability is beneficial for fully tapping into the potential of EVs as flexible load resources and avoiding the loss of vehicle-to-grid (V2G) interaction opportunities caused by missed identifications.

 doi:10.20944/preprints202603.1670.v1

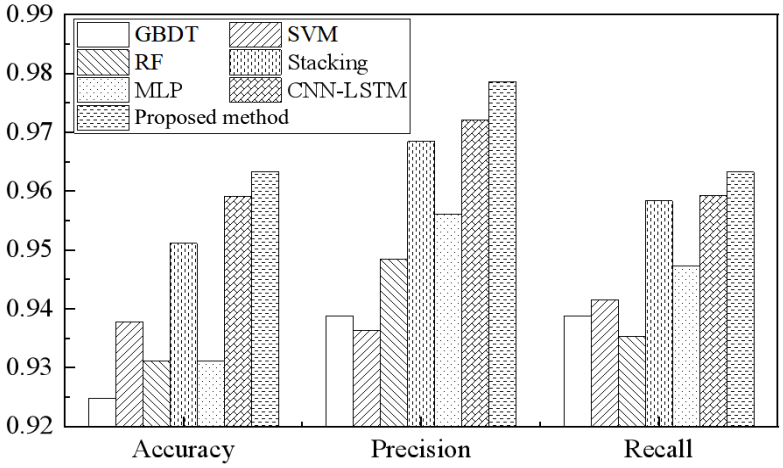

Figure 4. Comparison of electric vehicle recognition results of different methods.

5.3.3. Impact of Load Disaggregation on Identification Performance

To verify the specific impact of the load disaggregation stage on the final identification results, this study compares the model's performance under two distinct scenarios: "With Load Disaggregation" (the proposed framework) and "Without Load Disaggregation" (direct identification from net load). The results are summarized in Table 4.

Table 4. Comparison of identification results with and without load decomposition.

	Accuracy	Precision	Recall
With Load Disaggregation	0.9633	0.9786	0.9632
Without Load Disaggregation	0.9518	0.9578	0.9507

With the introduction of the load disaggregation process, the three primary metrics—Accuracy, Precision, and Recall—improved significantly to 96.33%, 97.86%, and 96.32%, respectively. This enhancement demonstrates that although distributed PV generation and EV charging loads may only overlap during limited temporal intervals, their temporal coupling and morphological confusion still severely compromise the characterization capability of EV charging features. This interference typically leads to constrained identification performance when using raw net load data.

By effectively decoupling the PV components via the CSSS algorithm, the intrinsic signatures of EV charging—such as high volatility, large-magnitude step changes, and sustained high-load durations—become more pronounced. Consequently, this leads to a significant boost in both the identification accuracy and the stability of the model when detecting electric vehicle behavior in high-PV-penetration scenarios.

5. Conclusions

This study proposed a two-stage electric vehicle (EV) identification model to address the challenges of identifying EV charging behavior in distribution networks with high-penetration distributed photovoltaics (PV). By constructing PV and base load feature vectors and employing a Context-aware Source-supervised Separation (CSSS) algorithm, the model achieves effective decoupling of PV output from the user cluster's base load, significantly mitigating the interference caused by behind-the-meter PV generation. On this basis, key features reflecting EV charging patterns were extracted, and a two-stage mechanism—combining Composite Charging Characteristic Index (CCCI) screening with sample-adaptive weighted Stacking ensemble learning—was designed. The simulation results lead to the following conclusions:

Exceptional Identification Performance: The proposed method demonstrates superior performance in EV identification, with Accuracy, Precision, and Recall reaching 96.33%, 97.86%, and

 doi:10.20944/preprints202603.1670.v1

96.32%, respectively. These results significantly surpass those of traditional individual classification models and unweighted ensemble methods, showcasing high decision confidence and robust coverage capability.

Effective Load Disaggregation: After isolating PV output using the CSSS algorithm, the estimated base load yielded an MAE of 0.1080 kW, an RMSE of 0.1602 kW, and a MAPE of 6.7%. The restoration of the base load makes EV charging signatures more pronounced, leading to a marked improvement in identification accuracy. This validates the critical importance of load disaggregation in enhancing EV detection precision.

Robust Feature Representation: A four-dimensional charging feature vector was constructed to account for the stochasticity and high-power characteristics of EV charging. Information Gain Ratio (IGR) analysis indicated that Total Nighttime Energy Consumption and the Mean Maximum Step Change Rate possess the strongest discriminative power for EV identification, with weight allocations of 0.30 and 0.28, respectively.

While the proposed model has achieved reliable results in scenarios without energy storage, future research will focus on joint identification methods involving energy storage systems (ESS). We aim to explore the modeling and fusion of storage behavior characteristics to improve the model's adaptability and accuracy in integrated "PV-Storage-Charging" coupled scenarios. Furthermore, the framework will be extended to address more complex user-side energy interaction landscapes.

Author Contributions: Conceptualization, Shuxian Yi and Guowu Li; methodology, Zhao Zhen; validation, Saining Yin and Zezhong Wang; resources, Xinsheng Ma; writing—original draft preparation, Zhao Zhen; funding acquisition, Shuxian Yi. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Science and Technology Project of State Grid Jibei Electric Power Co., Ltd. (Grant No. B3018K24000A).

Data Availability Statement: The data presented in this study are available in Ausgrid at https://www.ausgrid.com.au/, reference number [26, 27].

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

References

1. Kabalci, E.; Kabalci, Y.; Padmanaban, S.; et al. A Comprehensive Review of Solar PV Integration with Smart-Grids: Challenges, Standards, and Grid Codes. *Energies* **2024**, *17*, 1121.
2. Richardson, D. B. Electric vehicles and the electric grid: A review of modeling approaches, integrated assessment, and future research needs. *Renewable and Sustainable Energy Reviews* **2013**, *19*, 247–254.
3. Shu, Y.; Tang, Y.; Zhang, Z.; et al. Construction of New Distribution Network and Its Key Technologies. *CSEE Journal of Power and Energy Systems* **2024**, *10*, 15–28.
4. Zhuo, Z.; Zhang, N.; Kang, C.; et al. Quantitative Attribution Analysis Method of Power System Planning Scheme for Carbon Emission Peak and Carbon Neutrality Goals. *iEnergy* **2023**, *2*, 1–12.
5. Sovacool, B. K.; Hirsh, R. F. Beyond batteries: An examination of the benefits and barriers to introducing electric vehicles (EVs) and vehicle-to-grid (V2G) systems. *Energy Policy* **2009**, *37*, 1095–1103.
6. Kamoona, A.; Song, H.; et al. Effective identification of distributed energy resources using smart meter net-demand data. *IET Smart Grid* **2021**, *4*, 352–363.
7. Yan, X.; et al. Edge-Computing-Based Non-Intrusive Load Monitoring in Smart Grids: A Review. *IEEE Sensors Journal* **2023**, *23*, 16543–16558.
8. Qu, Z.; Ge, X.; Lu, J.; et al. Unsupervised disaggregation of aggregated net load considering behind-the-meter PV based on virtual PV sample construction. *Applied Energy* **2025**, *381*, 125007.
9. Khaleghian, S.; Doan, T. N.; Knox, J.; et al. Data-Driven Insights into EV Charging Patterns: Machine Learning Models Reveal Key Predictors of Station Utilization. *IEEE Access* **2024**, *12*, 42900–42915.
10. Ge, X.; Lu, J.; Wang, F.; et al. Optimal Bidding Model for Electric Vehicle Virtual Power Plant Considering Green Electricity Contract Decomposition in Spot Market. *IEEE Transactions on Smart Grid* **2025**, Early Access.

 doi:10.20944/preprints202603.1670.v1

11. Xiao, Z.; Liu, X.; Wang, X.; et al. Charging Load Prediction of Electric Vehicles with Multiple Power Levels in Large Communities. *Journal of Modern Power Systems and Clean Energy* **2025**, *13*, 45–58.
12. Wang, F.; Wang, G.; Xu, F. Review on Aggregation Characteristics and Trading Mechanisms of Virtual Power Plant for Enhancing System Response Capability. *Global Energy Interconnection* **2024**, *7*, 112–125.
13. Tabatabaei, S. M.; Dickert, J.; et al. Behind-the-Meter Load and PV Disaggregation via Deep Spatiotemporal Graph Generative Sparse Coding With Capsule Network. *IEEE Transactions on Neural Networks and Learning Systems* **2024**, *35*, 14573–14587.
14. Zhou, R.; Xiang, Y.; Wang, Y.; et al. Non-intrusive identification and load forecasting of household electric vehicle charging behavior based on smart meter data. *Power System Technology (English Edition)* **2022**, *46*, 1–12.
15. Hart, G. W. Nonintrusive appliance load monitoring. *Proceedings of the IEEE* **1992**, *80*, 1870–1891.
16. Barker, S.; Kalra, S.; et al. NILM Redux: The Case for Quiescent Filtering in Modern NILM. *Proceedings of the 1st ACM Conference on BuildSys* **2014**, 25–34.
17. Basu, K.; Debusschere, V.; Bacha, S. A review of non-intrusive load monitoring (NILM) for energy management systems. *Renewable and Sustainable Energy Reviews* **2021**, *146**, 111155.
18. Luan, W.; Ma, C.; Zhao, B.; et al. Non-intrusive Online Identification of Electric Bicycle Charging Load. *International Journal of Electrical Power & Energy Systems* **2022**, *142*, 108215.
19. Brefeld, T.; Lammert, S.; et al. Detection of Electric Vehicles and Photovoltaic Systems in Smart Meter Data. *Energies* **2022**, *15*, 4922.
20. Liang, Y.; Ding, Y.; et al. Urban EV fast-charging demand forecasting model combining data-driven approach with human decision-making behavior. *IEEE Transactions on Industry Applications* **2023**, *59*, 2345–2356.
21. Almutairi, A.; Ahmed, M.; et al. Data-Driven Modeling of Electric Vehicle Charging Sessions Based on Machine Learning Techniques. *Energies* **2025**, *16*, 1043.
22. Kong, W.; Dong, Z. Y.; et al. A Hierarchical Hidden Markov Model Framework for Home Appliance Modeling and Load Disaggregation. *IEEE Transactions on Smart Grid* **2018**, *9*, 3072–3081.
23. Qu, Z.; Ge, X.; Lu, J.; et al. Unsupervised Disaggregation of Aggregated Net Load Considering Behind-the-Meter PV Based on Virtual PV Sample Construction. *Applied Energy* 2025, *381*, 125007.
24. Yuvaraj, T.; Devabalaji, K.R.; Anish Kumar, J.; et al. A Comprehensive Review and Analysis of the Allocation of Electric Vehicle Charging Stations in Distribution Networks. *IEEE Access* 2024, *12*, 5404–5431.
25. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Francisco, CA, USA, 1993.
26. Bu, F.; Yuan, Y.; Wang, Z.; et al. A Time-Series Distribution Test System Based on Real Utility Data. In *Proceedings of the 2019 North American Power Symposium (NAPS)*, Wichita, KS, USA, 13–15 October 2019; pp. 1–6.
27. Muratori, M. Impact of Uncoordinated Plug-in Electric Vehicle Charging on Residential Power Demand. *Nat. Energy* 2018, *3*, 193–201.