ESJ Natural/Life/Medical Sciences Peer-reviewed

Machine Learning Techniques in Residential Electrical Load Forecasting: A PRISMA Review with LLM-Assisted Screening and Evidence Extraction

Nermin Siphocly, PhD
Misr International University, Egypt

Doi:10.19044/esj.2026.v22n6p1

Submitted: 08 January 2026
Accepted: 18 February 2026
Published: 28 February 2026

Cite As:
Siphocly, N. (2026). *Machine Learning Techniques in Residential Electrical Load Forecasting: A PRISMA Review with LLM-Assisted Screening and Evidence Extraction*. European Scientific Journal, ESJ, 22 (6), 1. https://doi.org/10.19044/esj.2026.v22n6p1

Abstract

This systematic review assesses machine learning (ML) techniques for residential electrical load forecasting, highlighting their effectiveness, methodologies, and challenges. Conducted under PRISMA 2020 guidelines, the review includes peer-reviewed Q1 and Q2 journal studies published between January 2020 and June 2025. These studies, sourced from the Web of Science, were selected based on their use of ML methods in residential contexts and focus on forecasting performance metrics and implementation. Extracted data covered forecasting horizons, ML algorithms, performance metrics, and limitations. Database searching yielded 712 records. After refinement and eligibility screening, 214 records were retained for title and abstract review, of which 105 were excluded. Only 93 full-text articles could be retrieved and assessed, seven of which were ultimately excluded due to methodological or contextual ineligibility, leaving the final 86 eligible studies. We present a novel multi-stage screening pipeline that incorporates semantic similarity models—specifically a zero-shot retrieval hybrid classifier—and large language models (ChatGPT-4o and Grok 3). Additionally, we examine their behavior, performance, and misclassification patterns throughout the screening process. We highlight key gaps in current literature-reproducibility issues, geographical imbalance, LLM bias, and limited use of explainable or privacy-preserving models—and suggest future research directions for residential load forecasting.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

Keywords: Machine Learning, Electrical Load Forecasting, Residential Areas, PRISMA, Systematic Review

Introduction

Electric energy forecasting plays a significant role in reducing energy waste by enabling more efficient management of electricity supply and demand. Accurate forecasts help utility companies anticipate energy needs, allowing them to optimize power generation and distribution Sakib et al., 2025). This reduces the need for excess energy production, which often leads to waste. Electric energy forecasting is a key tool in creating a more sustainable and efficient energy system.

In residential areas—where consumption patterns are shaped by diverse factors such as lifestyle, weather, and socioeconomic conditions—precise forecasting is particularly challenging yet vital. The International Energy Agency (IEA) projects that global electricity demand will grow significantly, with 85% of additional demand through 2026 expected from outside advanced economies International Energy Agency (IEA) (2022). This underscores the need for advanced forecasting techniques to address the complexities of residential energy use.

This paper presents a literature review on the state-of-the-art machine learning techniques used in electrical load forecasting in residential areas. Specifically, our survey addresses the following research questions:

- ***RQ1:*** *What methods and models have been used for short-term and long-term electrical load forecasting in residential areas?*
- ***RQ2:*** *What performance metrics are commonly used in the literature for evaluating the prediction of household electrical load?*
- ***RQ3:*** *What are the geographical and demographic regions most represented in residential electrical load forecasting research, and what gaps exist in terms of regional and population-based coverage?*
- ***RQ4:*** *How effective are semantic models and Large Language Models (LLMs) in supporting automated or semi-automated screening and classification in PRISMA-based systematic reviews?*
- ***RQ5:*** *What emerging trends, research gaps, and future directions can be identified in residential electrical load forecasting using machine learning?*

The main contributions of this study are summarized as follows:

- **Domain-specific focus:** This study provides a systematic review exclusively focused on residential electrical load forecasting using machine learning and deep learning approaches.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

- **Novel hybrid screening methodology:** We introduce an innovative multi-stage screening pipeline that integrates semantic similarity —a zero-shot retrieval hybrid classifier, large language models (ChatGPT-4o and Grok 3). We further analyze their behavior, performance, and misclassification patterns during the screening process.
- **Comprehensive categorization of forecasting techniques:** We systematically classify statistical, machine learning, hybrid, attention-based, transformer, and federated learning models, and map them to forecast horizons, data types, and application contexts, highlighting emerging trends such as privacy-preserving learning and spatio-temporal forecasting.
- **Evaluation metrics analysis:** We examine over 30 evaluation metrics, including traditional error-based metrics (RMSE, MAE, MAPE) and advanced metrics such as PTE, CRPS, DTW, ACE, and Pinball Loss, revealing a shift towards uncertainty-aware, time-sensitive, and robustness-oriented assessment practices.
- **Identification of open challenges and research gaps:** We highlight critical limitations in current literature, including geographical imbalance, LLM bias, and limited use of explainable or privacy-preserving forecasting architectures. We outline future research directions to advance residential load forecasting.

The rest of this paper is organized as follows. The Methods Section outlines the methodologies employed in this study, including the systematic literature review strategy, the inclusion and exclusion criteria for selecting manuscripts, and the PRISMA-based text analysis method. The Results Section presents, analyzes, and discusses the findings, providing further details on the PRISMA text analysis procedure. Finally, the Discussion Section discusses the implications of the results, highlights key research gaps, addresses this study's limitations, offers conclusions, and proposes directions for future research.

Methods

Eligibility Criteria

This study focuses exclusively on electrical load forecasting in residential areas, using machine learning techniques. Our review is limited to English-language journal articles published between January 2020 and June 2025, specifically those appearing in journals ranked in the first or second quartiles (Q1 and Q2) and top-tier conferences.

Information Sources

For our review, we use the WOS - Web of Science - Core collection database Clarivate (2025), which is a carefully curated database of high-

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

quality, peer-reviewed scholarly publications—including journals, books, and conference proceedings—spanning multiple disciplines. As a central component of the Web of Science platform, it is renowned for its stringent selection criteria and is widely regarded as a reliable resource by researchers and academics.

Search Strategy

We conducted our search using the WOS search engine and exported the results as a text file, which we then converted to a CSV file using the Zotero tool. We began our process using the general search terms: ”Electrical Load Forecasting in Residential Areas”, which yielded 712 search results. We subsequently refined our search using Boolean operators: (”load forecasting” OR ”energy prediction”) AND (”residential” OR ”household”) AND (”machine learning” OR ”deep learning” OR ”statistical model”). Additionally, we included the duplicate-removal feature. This refined search produced 374 results. We then filtered the results to include only papers published between 2020 and 2025 and written in English, which reduced the count to 368. To align with our inclusion standards, we retained only articles published in journals ranked Q1 or Q2 and top-tier conferences, as previously mentioned, resulting in the exclusion of 154 papers. Therefore, the number of papers we proceeded with for the next screening phases was 214. The different screening stages are discussed as follows.

Title and Keywords Screening: Our search and filtering process was conducted in several phases; the initial phase involved title screening, during which we excluded all papers not related to ”residential” areas or ”household”. We removed titles associated with ”factory”, ”industrial”, or ”commercial” contexts, as well as those focused on load forecasting in ”agricultural”, ”irrigation”, or ”Electric Vehicles (EV)”. Since our study does not focus on anomaly detection or load monitoring (e.g., non-intrusive load monitoring), titles addressing these topics were also excluded. Additionally, we excluded any papers that were reviews or surveys. The same criteria were applied to the keywords of each paper. Following this phase, the total number of papers was reduced to 193.

Abstract Screening: In the second phase, we conducted abstract screening, applying the same exclusion criteria used during the title screening phase. Moreover, we excluded papers that focused on ”thermal load forecasting” instead of electrical load forecasting. Similarly, we removed those concerning renewable energy such as wind or solar energy resources, as these are not widely accessible to most individuals Tigo (2023).

Instead of relying exclusively on keyword matching at this stage - a method that often proves unreliable due to its disregard for contextual meaning - we developed a semantic abstract screening tool, which is described in detail

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

in the subsequent subsection. Furthermore, we employed Large Language Models (LLMs), such as ChatGPT-4o and Grok 3, to interpret the content of each abstract and assess its relevance and conformity with our predefined inclusion criteria. The proposed methods are outlined in the following sub-sections.

Semantic Abstract Screening

To develop our semantic abstract screening tool, we experimented with three different NLP-based approaches, each designed to assess contextual relevance beyond keyword matching. The first approach used a sentence transformer model, where we created a descriptive reference paragraph reflecting inclusion and exclusion criteria, embedded both the reference and abstracts, and computed cosine similarity scores. However, this method produced very similar scores across most abstracts, making it ineffective for distinguishing relevant and irrelevant studies.

The second approach used a CrossEncoder fine-tuned for natural language inference (NLI), where each abstract was compared against predefined candidate labels (e.g., “residential load forecasting”, “anomaly detection”, “renewable energy forecasting”). The model generated entailment scores and assigned decisions such as “include,” “exclude,” or “maybe.” While this method captured contextual meaning better than the sentence transformer, it often struggled with abstracts that combined multiple forecasting topics, leading to inconsistent predictions.

The third approach implemented a zero-shot classification pipeline using a DistilBERT model trained on the Multi-Genre Natural Language Inference (MNLI) dataset, enabling semantic inference without task-specific training. It defines a set of domain-specific candidate labels, applying the classifier to infer the most probable label based solely on natural language understanding. For each abstract, it extracts the top-ranked label and its associated confidence score, storing these outputs along with the original metadata. This model demonstrated better contextual understanding, particularly in abstracts with overlapping terminology, and provided confidence scores that were useful for identifying borderline cases.

All three approaches were evaluated using a manually labeled subset of 35 abstracts sampled from the 193 abstracts entering the abstract screening phase. Each abstract was classified according to its primary application domain—such as residential electrical load forecasting, commercial or industrial energy use, renewable energy systems, thermal load forecasting, heating/cooling systems, or unclear/mixed focus—allowing structured assessment of contextual relevance.

Manual labeling yielded 25 exclusions, 6 inclusions, and 4 cases requiring further investigation. The predominance of exclusion categories

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

reflects the diversity of non-target themes encountered during screening and provides a realistic, imbalanced benchmark for evaluating abstract-screening performance.

When evaluated on this labeled subset, the zero-shot DistilBERT classifier demonstrated the highest reliability in distinguishing residential electrical load forecasting studies from non-target domains, particularly in cases involving overlapping terminology.

Although the 35 abstracts were not sufficient to fine-tune a domain-specific model (due to overfitting concerns), we used the same labeled set to enhance the zero-shot approach for future abstract screening by incorporating retrieval-based semantic support examples. We developed a hybrid screening strategy that combines zero-shot entailment scores with similarity-weighted label borrowing from the manually labeled samples. In this approach, abstracts are embedded using SPECTER2, compared to labeled support samples using cosine similarity, and the resulting retrieval-based logits are blended with zero-shot logits using Equation 1. This enabled the model to leverage both contextual inference and sample-based guidance, improving decision robustness for future abstract screening.

$$\mathbf{F} = \alpha\mathbf{z} + (1 - \alpha)\mathbf{r} \tag{1}$$

where $\mathbf{z}$ = zero-shot logits vector, $\mathbf{r}$ = retrieval-based logits vector, and $\alpha \in [0,1]$ = blending weight.

Of the 193 abstracts entering the abstract screening phase, 35 were manually labeled for evaluation, leaving 158 abstracts to proceed through the automated hybrid screening. In subsequent steps, we focused on screening these 158 abstracts, temporarily setting aside the manually reviewed subset. The outcomes of the manual and automated screening streams were later consolidated prior to initiating the full-text screening phase.

LLMs for Abstract Screening

Recent research has begun to explore the application of ChatGPT in conducting literature reviews Teperikidis et al. (2024). In this study, we extend this line of inquiry by evaluating the performance of Grok alongside ChatGPT for the specific task of abstract screening. The models utilized in this work are ChatGPT-4o and Grok 3.

Given that neither LLM was capable of processing all 158 abstracts in a single prompt—frequently omitting entries beyond the first 40—we segmented the dataset into batches of 40 abstracts each. The following prompt was employed for each batch:

This file has a list of research papers, I want you to read and understand the abstract of each paper (column C) and decide if it complies with the following criteria:

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

- *focuses on electrical load forecasting in residential areas only.*
- *doesn't include non-residential contexts in the study, such as commercial or manufacturing buildings.*
- *not related to agricultural energy systems*
- *doesn't study anomaly detection.*
- *doesn't study load monitoring as primary focus.*
- *not related to thermal load forecasting*
- *doesn't include electric vehicles.*
- *not a literature review or a survey paper*
- *renewable energy is not included in the study.*

You can write your decision in column E; whether this paper complies to ALL the previous criteria so it will be included or the paper fails to satisfy any of the criteria (please mention it) so the paper is considered excluded, or it is not clear from the abstract if all the criteria are satisfied.

Table 1: Breakdown of decisions of the three models regarding the papers under review

Included by all models	28
Excluded by all models	19
Included by GPT and Grok only	19
Excluded by GPT and Grok only	23
Included by our Semantic model and Grok only	3
Excluded by our Semantic model and Grok only	15
Included by our Semantic model and GPT only	19
Excluded by our Semantic model and GPT only	5
Total of unclear decision	27
Total	158

Fig. 1 visually summarizes the hybrid abstract screening workflow and the structured multi-model consensus and adjudication logic applied in this study. As illustrated in Fig. 1, abstracts were evaluated in parallel using the semantic hybrid model and two large language models (ChatGPT-4o and Grok 3). Decisions were compared to identify full consensus, partial disagreement, or unclear classifications. The numerical breakdown of these categories is presented in Table 1.

Decision-level consensus and rationale agreement*:* Following parallel evaluation (Fig. 1; Table 1), 28 abstracts were included through full consensus and advanced without further review. Nineteen abstracts were excluded by all three models; however, we examined whether this agreement extended to both the exclusion decision and its rationale.

Eight abstracts showed full agreement at both levels—that is, all models excluded them based on the same criterion (e.g., non-residential context, agricultural focus, or thermal load forecasting). These were removed without further review.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

The remaining 11 abstracts exhibited agreement in decision but divergence in stated reasons. These were manually reassessed to evaluate justification validity. In ten cases, ChatGPT and Grok provided identical and correct exclusion reasons. In the remaining case, Grok's reasoning was correct. In contrast, our model failed to make the correct decision for all 11 papers, suggesting that it requires additional fine-tuning to improve its classification accuracy.

Partial disagreement and adjudication outcomes: Cases with partial disagreement—where two models agreed and one differed—were manually adjudicated. Notably, in the group excluded by ChatGPT and Grok only (23 papers), the models agreed on exclusion reasons for 21 papers and disagreed on two; one disagreement favored Grok's reasoning, while in the other both models were incorrect and the paper was reinstated.

For papers excluded by the semantic model and Grok (15 papers), manual review showed that eight were wrongly excluded, two were unclear and deferred for further review, two were correctly excluded with consistent reasoning, and three were manually excluded due to lacking electrical load forecasting as their primary focus.

In the group excluded by the semantic model and ChatGPT (5 papers), two papers were wrongly excluded, one exclusion was correctly justified by ChatGPT, and two were flagged for further review due to abstract-level ambiguity.

At this stage, 111 papers remained from the initial 158 abstracts.

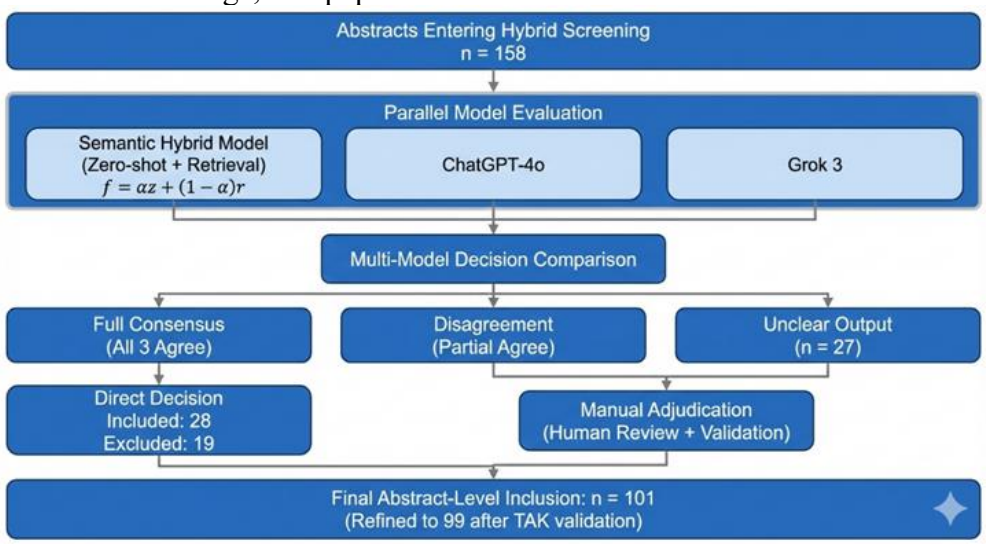


Figure 1: Hybrid Abstract Screening Workflow

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

Unclear classifications and final resolution: The 27 abstracts classified as "unclear" by at least one model were reviewed manually, leading to three additional exclusions and reducing the count to 108 papers.

Given the strong overall performance of ChatGPT and Grok, we revisited cases where only one LLM issued an exclusion. Many of these single-model exclusions were accompanied by low-confidence language (e.g., "maybe" or "likely"), warranting further validation. Among these, 17 papers had been excluded only by Grok and 3 only by ChatGPT. Reassessment resulted in seven additional exclusions: five correctly by Grok, one excluded by Grok for an incorrect reason, and one correctly excluded by ChatGPT. This yielded a final abstract-level inclusion of 101 studies.

Interestingly, when we queried ChatGPT again about some abstracts it had previously marked for inclusion—despite being excluded by Grok—it revised its decision to exclusion. For instance, in one case, ChatGPT recognized the mention of "lodging" in the abstract (originally flagged by Grok), and identified that the dataset used was the Great Energy Predictor III, which includes various building types (e.g., educational, commercial, residential, lodging, etc.), thus not focusing exclusively on residential contexts. This process also identified a number of borderline cases where the primary focus was unclear from the abstract alone, warranting further examin

ation in the next stage. Table 2 consolidates adjudication outcomes across disagreement categories, highlighting relative model reliability and error patterns.

Table 2: Adjudication outcomes across partial-disagreement groups

Exclusion Group	Total Papers	Both Correct	One Model Correct	Both Incorrect	Unclear
ChatGPT + Grok	23	21	1 (Grok)	1	0
Semantic + Grok	15	2	0	8	2
Semantic + ChatGPT	5	0	1 (ChatGPT)	2	2

LLMs for TAK screening

TAK, an abbreviation for Title, Abstract, and Keywords, refers to the combined use of these three elements for screening purposes. In this step, instead of asking the LLMs to determine whether a paper should be included or excluded based on predefined criteria, we instructed them to summarize the scope of each paper in a single phrase using its TAK. To define the task, we provided the following prompt:

This file has the TAK (title/abstract/keywords) of some research papers. The first column is a key identifying each research paper and the second column is its TAK. I want you to process each TAK and understand it and write its scope in one phrase, neglecting the details of implementation or algorithms used. My target is to know if the paper focuses on electrical load

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

forecasting in residential areas solely. Write the scopes on a downloadable table!

The analysis produced by Grok was notably informative, as it effectively captured the key aspects of each paper in a concise and meaningful phrase. This level of descriptive clarity enabled the identification and exclusion of two additional papers that focused on renewable energy—an aspect explicitly noted in Grok's output, such as in the following example: *"Focuses on net load forecasting for residential areas with high photovoltaic penetration."* Upon reviewing the abstracts of these two papers, we confirmed their emphasis on renewable energy, warranting their exclusion. In contrast, ChatGPT tended to generate generic summaries for nearly all entries, typically using the phrase: *"Residential electrical load forecasting"*, without providing further distinguishing details.

Of the 158 abstracts processed through hybrid screening, 99 were retained after adjudication and TAK validation. Incorporating the earlier manually screened subset of 35 abstracts (from the 193 entering abstract screening) added 6 inclusions and 4 papers flagged for further review, resulting in 109 papers advancing to full-text retrieval. Full texts were successfully obtained for 93 of these papers.

Full Paper Screening: We began the full-text screening phase by examining papers whose abstracts did not clearly indicate whether they were related to electrical load forecasting. During this stage, we made effective use of SCISPACE, a comprehensive platform designed to support the academic research process. Specifically, we utilized the "Ask the PDF" feature, an AI-powered tool that enables users to interact with research papers and other PDF documents through natural language queries.

We posed questions such as *"Is this paper concerned with electrical or thermal load forecasting?"* This tool proved especially helpful, as it not only provided answers but also highlighted the specific phrases and sections from which those answers were derived, allowing us to quickly focus on the most relevant content. Ten papers were previously identified as having abstracts with unclear focus. Upon further evaluation, four of these were excluded. In addition, we conducted full-text screening for all papers that had received an "unclear" decision from any of the models; however, no additional exclusions resulted from this group. Consequently, the final number of papers included after the full-text screening phase was 89.

The complete screening pathway and study reduction at each stage are visually summarized in Fig. 2. In total, 158 abstracts underwent hybrid screening, resulting in 101 abstract-level inclusions and 99 after TAK validation. Incorporation of the manually labeled subset added 10 abstracts (6 inclusions and 4 flagged for further review), bringing the total advancing to full-text retrieval to 109. Of these, 93 full texts were successfully obtained,

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

leading to 86 final studies. These transitions correspond directly to the sequential screening stages described above.

Data Collection Process

Initial data extraction was semi-automated using ChatGPT (OpenAI GPT-4o) and revised by the author. Primarily, the model was prompted to extract predefined data elements from each full-text article. The extracted fields included: ***Study design, Population or setting, Intervention or primary focus, Forecasting Models Used, Outcomes measured, Time Horizon, Measured Outcomes (Evaluation Metrics), and Relevance to the review question.***

A structured prompt template was applied uniformly across all included studies to ensure consistency in the extraction process. The full-text of each article was processed by the model, focusing on the abstract, methods, and results sections.

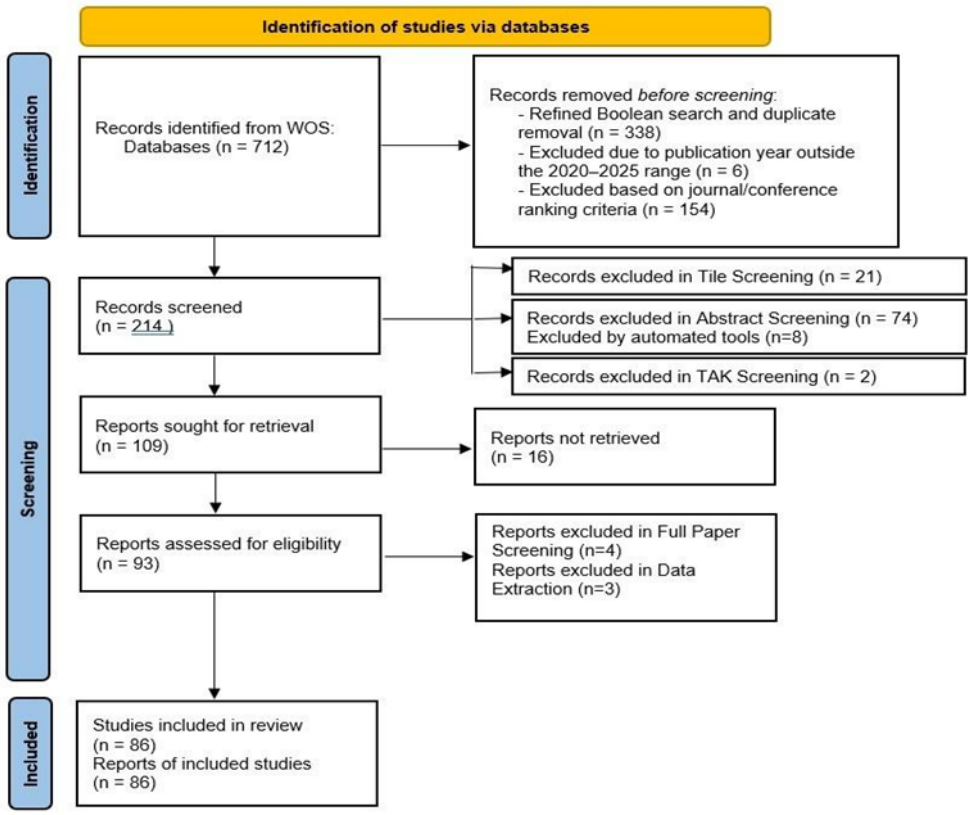


Figure 2: PRISMA 2020 flow diagram for study selection

A pilot test was conducted on a sample of five papers to assess ChatGPT's extraction accuracy and refine the prompt structure. We asked the model to act as a PRISMA reviewer and extract the required listed data. The

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

model parsed the first research paper correctly and extracted accurate data. However, in the second paper, it parsed a wrong title and a wrong dataset, so we asked the model what is the best prompt that would minimize these types of errors, and that is what it suggested:

"Act as a systematic review assistant following PRISMA guidelines. Read the full paper carefully, and extract the following information strictly based on the actual content, not assumptions or prior knowledge:

1) *Study design (e.g., experimental, simulation, case study)*
2) *Population or setting (with precise dataset names and geographic regions)*
3) *Intervention or primary focus (e.g., forecasting, modeling, disaggregation)*
4) *Time horizon of the analysis (short-term, long-term, or annual)*
5) *Outcomes measured (specific performance metrics)*
6) *Key findings*
7) *Relevance to the review question: "What methods and models have been used for short-term and long-term electrical load forecasting in residential areas?" — clearly state if the study is directly, partially, or not relevant based on the study's forecasting goal and data granularity.*

Do not infer or fill in missing data. Be cautious with datasets and time horizons — explicitly confirm them from the text. If something is not clearly stated, write "Not specified" or "Not applicable.""

The suggested prompt improved the quality of extracted data remarkably. We added the Forecasting Models Used as a separate point to put more focus on them. Following each automated extraction, the author reviewed and verified all extracted data for accuracy and completeness. Discrepancies or ambiguities were resolved through consensus discussion. In cases of uncertainty, the original full-text was re-reviewed manually. This approach allowed for a significant reduction in manual workload while maintaining rigorous data validation, consistent with PRISMA guidelines.

During the data extraction phase, three additional studies were excluded after full-text review showed they did not meet the inclusion criteria—due to using the wrong population, having an unrelated focus, and including renewable energy, respectively. As a result, 86 studies remained for analysis.

Synthesis Methods

We conducted a narrative synthesis of the included studies, structured around key characteristics such as study design, forecasting models used, time horizon (short-term, long-term, annual), data sources, and performance

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

metrics. Summary tables were used to organize and compare models, datasets, and outcomes across studies.

We grouped the included studies based on key characteristics to facilitate thematic and comparative analysis. These characteristics included the publication year, geographical location of the study population (countries or regions), study design (e.g., experimental, simulation-based), types of machine learning models employed, and the evaluation metrics reported. This categorization enabled us to identify trends over time, regional focus areas, methodological patterns, and variations in model performance assessment. This review adheres to the PRISMA 2020 guidelines for systematic reviews Page et al. (2021).

Results

The Results section progresses from screening outcomes (Fig. 2; Table 1) to study characteristics (Fig. 3; Table 3) and finally to methodological and evaluative trends (Figs. 4–7).

Study Selection

A total of 712 records were identified through WOS database searching. After Boolean search refinement and removing duplicates, 388 records were removed. An additional 6 records were removed due to being outside the publication range [2020-2025]. Records excluded based on Journal/Conference ranking criteria were 154. As a result, 214 records were screened by title and abstract. Of these, 105 were excluded based on eligibility criteria.

Ninety-three full-text articles were assessed for eligibility. Four were excluded for reasons such as non-residential setting, absence of forecasting models, or methodological irrelevance. Additionally, three studies were excluded during the data extraction phase after full-text re-evaluation revealed previously unrecognized ineligibility. In total, 86 studies were included in the final qualitative synthesis.

Study Characteristics

A total of 86 studies were included in qualitative synthesis. Forecasting horizons varied from short-term to long-term/annual projections. Short-term forecasting dominates the literature (85 studies), while medium-term (4), long-term (2), and annual (1) horizons appear only marginally and exclusively in combination with short-term analyses.

The included papers fall into two study design categories: Experimental and Simulation-based. Experimental studies constitute the majority, making up 74.4% of the total. The Simulation-based studies account for the remaining 25.6%, as shown in Fig. 3(a).

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

Fig. 3(b) presents the distribution of studies by publication year. It is shown that the majority of studies were published in 2025, accounting for 20.9% of the total, followed closely by 2022 with 19.8% and 2024 with 18.6%. The years 2023, 2021, and 2020 each contributed 10.5% of the studies. Earlier years saw considerably fewer publications, with 2018 representing 5.8%, 2019 contributing 2.3%, and 2016 the least at 1.2%.

Detailed study-level mappings are provided in the following link: https://github.com/nermin-siphocly/ml-residential-load-forecasting-prisma, in a file named: "Included Studies – Tables.pdf".

RQ1: What methods and models have been used for short-term and long-term electrical load forecasting in residential areas?

A comprehensive grouping of models by category is presented in Table 3, allowing readers to quickly identify methodological patterns across studies. The studies present a vast and diverse range of machine learning and deep learning models applied to time-series forecasting and load prediction, with Long Short-Term Memory (LSTM) and its variants (BiLSTM, Seq2Seq, LSTM-AE, Co-LSTM, etc.) being the most frequently employed. Numerous hybrid architectures are introduced, such as CNN-LSTM, CNN-GRU, CNN-BiGRU, and CNN-SE, often enhanced with attention mechanisms, skip connections, or domain adaptation techniques like transfer learning, DATN, and MK-MMD. Other recurrent models include GRU, BiGRU, RNN, and DRNN, while federated learning frameworks (FedAvg, FedSGD, Ditto, FedAAVG) integrate privacy-preserving collaborative training across LSTM-based models. Traditional machine learning regressors (Random Forest, XGBoost, LightGBM, CatBoost, SVR, KNN, Decision Trees) are widely used individually or in ensemble strategies (e.g., weighted, stacked, or convex combinations). Additional techniques include autoencoders (standard, sparse, denoising, pre-trained), graph-based models (STFAG-GCN, Graph WaveNet), transformers (JITtrans, Deep-Autoformer), and statistical baselines (ARIMA, Exponential Smoothing, Prophet). Optimization and preprocessing methods such as GA, BGA-PCA, mutual information, VMD, wavelets, and k-means clustering are often integrated for feature selection or input segmentation, resulting in highly customized, task-specific predictive architectures.

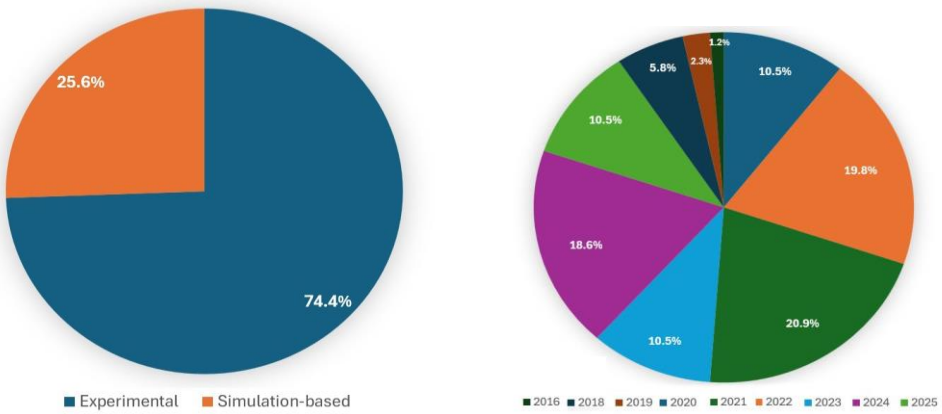
RQ2: What performance metrics are commonly used in the literature for evaluating the prediction of household electrical load?

Figs. 4–6 collectively visualize the distribution of evaluation metrics across the included studies, categorized into error-based, peak-specific, probabilistic, computational, statistical, and privacy-related measures. The studies employed a wide range of evaluation metrics to assess forecasting

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

performance, with the most frequently used being Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Additional commonly reported measures included Mean Squared Error (MSE), R-squared (R2), Normalized RMSE (NRMSE), and Coefficient of Variation (CV).

Several studies also incorporated advanced or task-specific metrics such as Dynamic Time Warping (DTW) error, Peak Timing Error (PTE), Spectrum Similarity (SS), Continuous Ranked Probability Score (CRPS), Quantile Score (QS), and probabilistic accuracy indicators like Pinball Loss, Interval Score (IS), and Average Coverage Error (ACE). Other considerations included computational efficiency (e.g., training time, runtime), model complexity (e.g., communication cost, model size), and statistical significance tests (e.g., Diebold-Mariano, Wilcoxon). This diversity reflects a comprehensive effort to evaluate both point and probabilistic forecast accuracy, robustness, and efficiency.



(a) Study Design Distribution. (b) Publication Years Distribution.

Figure 3: Study Design and Publication Year Distributions Among Studies

Table 3: Studies grouped by machine learning models used

Model	**Papers**
ADMIF	Fang and He (2023)
ANN	Fayaz and Kim (2018), Sulaiman et al. (2022), Truong et al. (2021), Harikrishnan et al. (2025), Dong et al. (2016)
ANFIS	Fayaz and Kim (2018)
Affinity Propagation	Dogra et al. (2023)
ATT-LSTM	Ozcan et al. (2021), Nguyen et al. (2020)
BiGRU	Aurangzeb et al. (2024), Irankhah et al. (2024)
BiLSTM	Aurangzeb et al. (2024), Kaur et al. (2024), Aurangzeb et al. (2024), Zhang et al. (2024)
CNN	Alhussein et al. (2020), Aurangzeb et al. (2024), Sakuma and Nishi (2022), Aouad et al. (2022), Jiang et al. (2021), Sajjad et al. (2020), Lotfipoor et al. (2024), Sinha et al. (2021), Shi and Wang

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

	(2022), Cheng et al. (2021), Acharya et al. (2024), Shi and Xu (2022), Ozdemir et al. (2025), Irankhah et al. (2024), Cao et al. (2025)
CatBoost	Muqtadir et al. (2025)
DA-LSTM	Ozcan et al. (2021)
DANN	Truong et al. (2021), Truong et al. (2021)
DATN	Zhu et al. (2024)
DBN	Fan et al. (2022), Fan et al. (2022)
Deep-Autoformer	Jiang et al. (2022)
DELM	Fayaz and Kim (2018)
DFNN	Al-Jamimi et al. (2023)
DI-RNN	Yuan et al. (2020), Kiprijanovska et al. (2020)
DMLP	Nguyen et al. (2020)
DRNN	Kiprijanovska et al. (2020), Shi et al. (2018)
DT	Moldovan and Slowik (2021), Ullah et al. (2021)
ELM	Sulaiman et al. (2022), Fayaz and Kim (2018)
Exponential Smoothing	Hribar et al. (2025), Yousaf et al. (2021)
FFNN	Kong et al. (2018), Kumaraswamy et al. (2024), Imani and Ghassemian (2019)
FedAvg	Park and Son (2023), Fekri et al. (2022), Dogra et al. (2023), Widmer et al. (2023), Qu et al. (2023)
FedSGD	Fekri et al. (2022)
GAN	Qu et al. (2023)
GBR	Ullah et al. (2021), Xia et al. (2024)
GMM	Dong et al. (2016)
GPR	Dong et al. (2016), Xia et al. (2024), Dab et al. (2023)
GRU	Aurangzeb et al. (2024), Sakuma and Nishi (2022), Sajjad et al. (2020), Khan et al. (2021)
Graph WaveNet	Lin et al. (2025)
JITtrans	Benali et al. (2024)
K-means	Han et al. (2021), Dogra et al. (2023), Kell et al. (2018), Khan et al. (2021), Acharya et al. (2024)
k-medoids	Dab et al. (2023)
KNN	Ullah et al. (2021), Moldovan and Slowik (2021), Forootani et al. (2022)
LS-SVM	Dong et al. (2016)
LSTM	Han et al. (2021), Alhussein et al. (2020), Lu et al. (2022), Fekri et al. (2022), Masood et al. (2022), Dogra et al. (2023), Hou et al. (2021), Yang et al. (2022), Ji et al. (2023), Aurangzeb et al. (2024), Xu et al. (2022), Nguyen et al. (2020), Li et al. (2025), Razghandi et al. (2021), Sakuma and Nishi (2022), Kong et al. (2018), Kumaraswamy et al. (2024), Fan et al. (2020), Kim and Cho (2019), Chen et al. (2024), Flor et al. (2021), Ullah et al. (2021), Park and Son (2023), Zhao et al. (2024), Jiang et al. (2021), Taik and Cherkaoui (2020), Imani and Ghassemian (2019), Ozcan et al. (2021), Sinha et al. (2021), Manandhar et al. (2024), Kell et al. (2018), Zhu et al. (2024), Dong et al. (2024), Shi and Wang (2022), Ismail et al. (2024), Shi and Xu (2022), Zang et al. (2021), Shahsavari-Pour et al. (2025), Gong et al.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

	(2022), Ozdemir et al. (2025), Bai (2024), Khan et al. (2021), Qu et al. (2023), Cao et al. (2025), Fan et al. (2022)
LightGBM	Muqtadir et al. (2025), Hribar et al. (2025)
LR	Li et al. (2025), Massidda and Marrocu (2018)
MLP	Nguyen et al. (2020), Sakuma and Nishi (2022), Kell et al. (2018), Jiang et al. (2022), Lin et al. (2025)
NBEATS	Li et al. (2025)
NHiTS	Li et al. (2025)
PDNN	Jeyaraj and Nadar (2021)
PDRNN	Shi et al. (2018)
PVS	Mansoor et al. (2021)
RNN	Fekri et al. (2021), Jagait et al. (2021), Flor et al. (2021), Yuan et al. (2020), Ozcan et al. (2021), Shi et al. (2018)
Random Forest	Chaianong et al. (2022), Moldovan and Slowik (2021), Nguyen et al. (2020), Li et al. (2025), Ullah et al. (2021), Xia et al. (2024), Lotfipoor et al. (2024), Manandhar et al. (2024), Kell et al. (2018), Massidda and Marrocu (2018)
Residual LSTM	Chen et al. (2024)
SVM	Pla and Jimenez Martinez (2023)
SVR	Nguyen et al. (2020), Sulaiman et al. (2022), Ullah et al. (2021), Kell et al. (2018), Ismail et al. (2024), Dong et al. (2016)
Seq2Seq	Masood et al. (2022), Razghandi et al. (2021), Sakuma and Nishi (2022), Aouad et al. (2022), Zhu et al. (2024), Dong et al. (2024)
Sparse Autoencoder	Cheng et al. (2021)
TCN	Li et al. (2025), Widmer et al. (2023)
TFT	Li et al. (2025)
TSMixer	Li et al. (2025)
TiDE	Li et al. (2025)
XGBoost	Yang et al. (2022), Li et al. (2025), Muqtadir et al. (2025), Hribar et al. (2025), Harikrishnan et al. (2025)

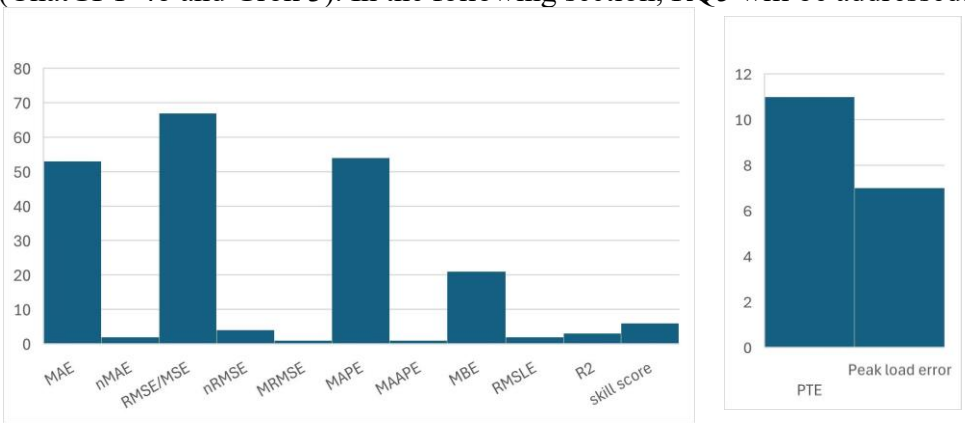
RQ3: What are the geographical and demographic regions most represented in residential electrical load forecasting research, and what gaps exist in terms of regional and population-based coverage?

Fig. 7 demonstrates the geographical distribution of studies based on the country or region of the dataset (population). The USA dominates the chart, contributing the largest share at 22.1% of the studies. This is followed by Canada at 11.6%, with Australia, China, and another unidentified region each contributing 10.5%. Several countries, including South Korea, Spain, and one other, each represent 5.8% of the total. Smaller contributions (ranging from 1.2% to 4.7%) come from a wide range of countries and regions, including the UK, Belgium, Switzerland, Morocco, UAE, Pakistan, Ecuador, Ireland, India, and others.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

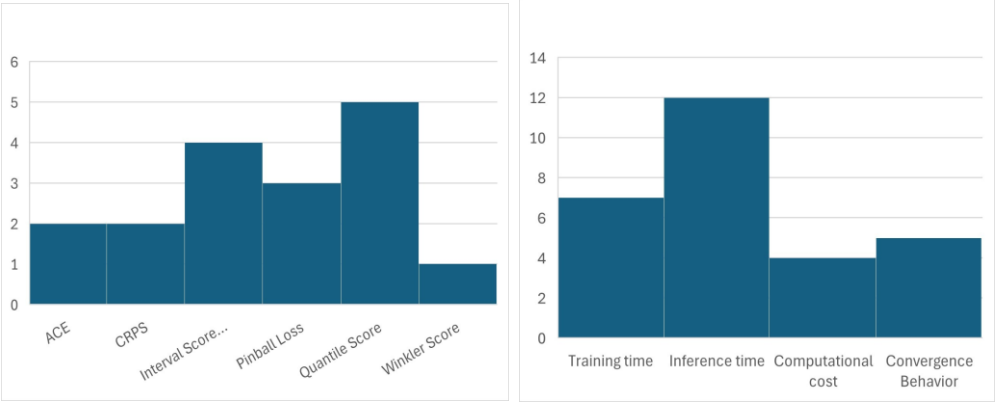
RQ4: How effective are semantic models and Large Language Models (LLMs) in supporting automated or semi-automated screening and classification in PRISMA-based systematic reviews?

This research question is answered within Section 2.3.2, where we compare the decisions of our semantic screening tool and the two LLMs (ChatGPT-4o and Grok 3). In the following section, RQ5 will be addressed.



(a) Error-Based Accuracy Metrics (b)Peak and Load-Specific Error Metrics

Figure 4: Distribution of Error-Based and Peak Metrics Among Studies



(a)Uncertainty & Probabilistic Forecasting (b)Computational Performance

Figure 5: Distribution of Probabilistic and Computation Metrics Among Studies

Discussion

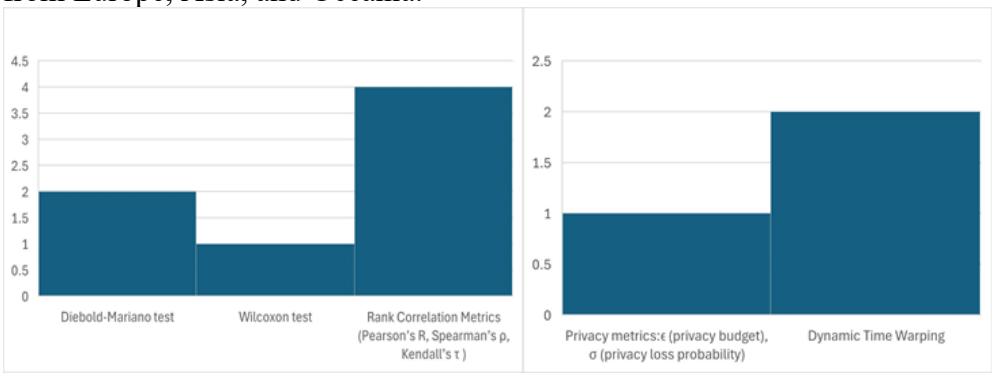
Interpretation of Key Findings

The distribution of studies by publication year highlights a growing research interest in the field, with over 70% of the studies emerging between 2022 and 2025.

The geographical distribution of studies indicates a global interest in the topic of residential electrical load forecasting, with a notable

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

concentration of research in North America and a diverse representation from Europe, Asia, and Oceania.



(a) Statistical Comparison Metrics (b)Privacy-Specific Metrics (for Federated /DP Models)

Figure 6: Distribution of Statistical and Privacy Metrics Among Studies

The current landscape of residential electrical load forecasting reveals a strong emphasis on deep learning models, particularly Long Short-Term Memory (LSTM) networks and their variants (e.g., CNN-LSTM, Seq2Seq, BiLSTM), reflecting the importance of modeling temporal dependencies in household energy consumption. The increasing use of hybrid and ensemble models—combining LSTM with attention mechanisms, convolutional layers, or traditional machine learning methods like Random Forest and XGBoost—suggests that no single model can capture the full complexity of load patterns. Additionally, clustering techniques and feature selection methods are often used as preprocessing steps to tailor models to specific household behaviors.

Challenges, Limitations, and Future Directions of Existing Literature

RQ5: What emerging trends, research gaps, and future directions can be identified in residential electrical load forecasting using machine learning?

Despite the breadth of existing research, several methodological weaknesses limit the maturity of residential load forecasting as a field. While the majority of studies are concentrated in countries like the USA, Canada, and Australia, regions such as Africa, South America, and parts of Asia remain significantly underrepresented, each contributing less than 2.5% of the total. This imbalance highlights a critical limitation in the global applicability of findings and underscores the need to expand research efforts in these underrepresented regions to ensure more inclusive and context-specific insights.

Emerging trends point to a growing interest in privacy-preserving approaches like federated learning, interpretable architectures using attention mechanisms, and early exploration of spatio-temporal and graph-based

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

models. However, gaps remain in standardization, benchmarking, and reproducibility. Future research should focus on developing explainable, modular, and geographically-aware forecasting systems, while also promoting transparency through open datasets and shared evaluation frameworks. These directions will support more robust, scalable, and trustworthy forecasting solutions for smart residential energy systems.

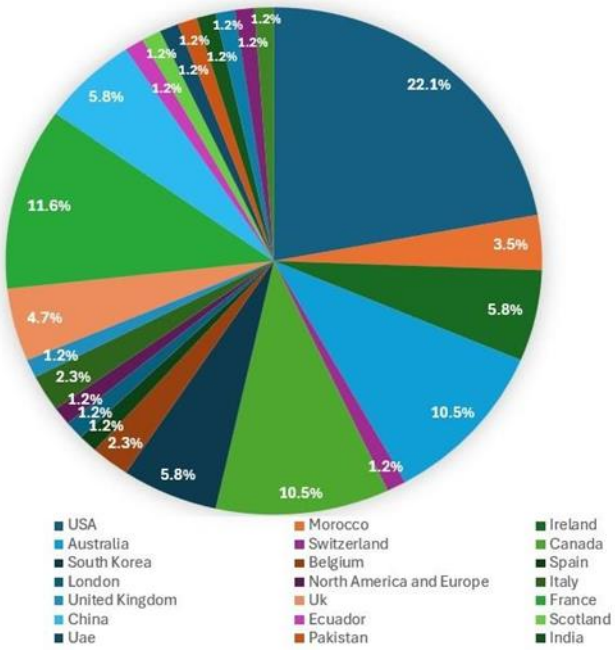


Figure 7: Studies distribution among countries

Although metrics provide a baseline understanding of model performance, they often fail to capture critical aspects like peak timing errors, uncertainty, and computational efficiency. A gradual shift is evident, however, with the emergence of specialized metrics—such as Peak Timing Error, Dynamic Time Warping (DTW), and probabilistic measures like Pinball Loss and Continuous Ranked Probability Score (CRPS)—reflecting growing interest in robustness, interpretability, and application-driven evaluation.

This trend indicates a field in transition, where future research would benefit from broader adoption of uncertainty-aware and efficiency-focused metrics, especially in real-time and resource-constrained settings. To enable meaningful model comparisons and support practical deployment, researchers are encouraged to standardize reporting practices, include peak and time-sensitive metrics, and account for computational cost and environmental impact. Ultimately, a multidimensional evaluation framework that integrates accuracy, robustness, uncertainty, and deployability is essential for advancing residential load forecasting research.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

The scarcity of long-term, high-resolution, and publicly accessible residential datasets makes it difficult to evaluate forecasting models in a consistent and scalable manner, especially across diverse climatic or socio-economic contexts. A large proportion of studies rely on small or short-duration datasets, often drawn from a single household or region, which increases the risk of overfitting and restricts the external validity of model performance claims. Another persistent challenge is the limited emphasis on robustness: most studies evaluate models under idealized conditions and few examine performance degradation under noise, sensor faults, data drift, or household behavioral variability. Cross-household generalization also remains problematic—models tailored to a specific home or region frequently fail to transfer to new settings without retraining. Finally, the computational complexity of many proposed architectures poses deployment challenges for real-time or edge-based environments, where limited processing power, memory constraints, and energy consumption often restrict the use of large deep learning models. These unresolved challenges highlight the need for more rigorous, standardized, and deployment-aware research practices moving forward.

Study Limitations

This review relied on a hybrid screening approach that combined semantic models, large language models (ChatGPT-4o and Grok 3), and manual adjudication. While this method improved efficiency and reduced initial screening effort, it introduced potential sources of inconsistency, particularly due to model stochasticity, potential hallucination, and prompt sensitivity. Although disagreement cases were manually reviewed, LLM-driven decisions may reflect context-driven bias.

Future research should adopt standardized benchmarks, domain-specific LLM fine-tuning, and explainable AI-based screening methods to enhance reproducibility, transparency, and comparability.

Data Availability

In Section 2.3, the search strategy, databases, keywords, and all inclusion and exclusion criteria are explicitly reported, and the PRISMA flow diagram documents each selection stage. The manually labeled set of 35 abstracts, used both for testing the semantic approaches and supporting the hybrid classifier, provides a transparent reference for screening decisions.

In the given link: https://github .com/nermin-siphocly/ml-residential-load-forecasting-prisma we provide the dataset of WOS research papers that we worked on in our systematic review, along with the codes for the Semantic Scores, Cross-Encoder, and DistilBERT Screening tools. We also provide the

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

manually labeled support list of abstracts in addition to the Zero-Shot Hybrid screening tool code.

Reproducibility is a key concern in systematic reviews, and it becomes more complex when using LLMs and semantic models for screening. While these tools improve efficiency and recall, their decision-making is not fully deterministic. Screening outcomes can vary due to prompt sensitivity, model updates, and stochastic behavior in LLM-generated outputs. Even when using identical abstracts and prompts, we observed small inconsistencies in decisions—particularly in borderline cases.

However, to enhance reproducibility, all screening steps were conducted using fixed model versions, documented all the prompts used, and clearly defined decision rules. The semantic screening tool and LLM classifiers (ChatGPT-4o and Grok 3) were applied using identical instructions to minimize variance.

Despite the stochastic nature of LLM outputs, maintaining a consistent prompting framework and recording all screening outcomes ensures methodological transparency and facilitates reproducibility. Therefore, reproducibility in LLM-assisted screening should be interpreted as process reproducibility (i.e., using the same workflow, prompts, and reasoning framework) rather than strict duplication of classification outputs. Future reviews can improve systematic review automation through logging rationales, and treating LLMs as decision-support tools rather than fully automated classifiers.

Threats to Validity

The use of LLMs in the screening process introduces potential threats to validity, particularly related to bias and potential hallucination. LLMs may incorrectly infer eligibility based on implicit assumptions rather than explicit content, leading to overinclusion of papers with loosely related terms (semantic bias) or unjustified exclusion due to misinterpreting domain-specific contexts. Additionally, hallucinations—where the model generates false or unsupported reasoning—can distort screening decisions, especially in abstracts with ambiguous or mixed-context content. To mitigate these risks, all the conflicting LLM-based decisions were cross-checked with human validation, and disagreement cases were resolved manually. However, the lack of fully transparent decision logic and potential for model drift remain inherent limitations.

Conclusions

Global electricity demand is increasing, underscoring the importance of accurate residential forecasting in enhancing energy efficiency and sustainability. This systematic review comprehensively examined 86 studies

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

on machine learning (ML) techniques for residential electrical load forecasting, adhering to the PRISMA 2020 framework. The review addressed two primary questions: the models and methods used in forecasting, and the evaluation metrics employed to assess their performance.

In the methodology section, a rigorous multi-stage screening process was outlined, beginning with Boolean search refinement, title and keyword filtering, and advanced abstract screening using semantic models and Large Language Models (LLMs) such as ChatGPT and Grok. A hybrid classification model combining zero-shot learning with retrieval-based similarity was developed and employed. Full-text screening and structured data extraction followed, guided by PRISMA-aligned prompts and validation from the author.

The results section detailed the final inclusion of 86 studies, most of which were experimental in design and focused on short-term forecasting. These studies spanned a diverse set of countries, though research was heavily concentrated in North America, with minimal representation from Africa and South America. A wide variety of ML models were applied, notably LSTM variants, hybrid CNN-based architectures, and ensemble methods. A diverse array of evaluation metrics was employed, including RMSE, MAE, MAPE, and more specialized metrics like PTE and CRPS, reflecting a growing attention to both accuracy and practical considerations such as efficiency and uncertainty.

The discussion emphasized the growing interest in deep learning and hybrid models, with increasing exploration into privacy-preserving, interpretable, and graph-based methods. However, it also underscored regional research gaps, limited reproducibility, and the need for broader adoption of uncertainty-aware and application-driven metrics. The field appears to be shifting toward more comprehensive, transparent, and context-sensitive forecasting frameworks.

Overall, this review consolidates current knowledge on ML-based residential load forecasting, identifies key methodological and geographical gaps, and highlights the need for standardized evaluation frameworks and more inclusive research efforts. These insights provide a foundation for advancing robust, scalable, and equitable forecasting solutions in the context of smart residential energy systems.

Conflict of Interest: The author reported no conflict of interest.

Data Availability: All data are included in the content of the paper.

Funding Statement: The author did not obtain any funding for this research.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

References:

1. Acharya, S. K., Yu, H., Wi, Y.-M., and Lee, J. (2024). Multihousehold load forecasting based on a convolutional neural network using moment information and data augmentation. ENERGIES, 17.0(4). Publisher: MDPI.
2. Al-Jamimi, H. A., Binmakhashen, G. M., Worku, M. Y., and Hassan, M. A. (2023). Advancements in household load forecasting: Deep learning model with hyperparameter optimization. ELECTRONICS, 12.0(24). Publisher: MDPI.
3. Alhussein, M., Aurangzeb, K., and Haider, S. I. (2020). Hybrid cnn-lstm model for short-term individual household load forecasting. IEEE ACCESS, 8.0:180544– 180557 Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC
4. Aouad, M., Hajj, H., Shaban, K., Jabr, R. A., and El-Hajj, W. (2022). A cnn-sequence-to-sequence network with attention for residential short-term load forecasting. ELECTRIC POWER SYSTEMS RESEARCH, 211.0. Publisher: ELSEVIER SCIENCE AS.
5. Aurangzeb, K., Haider, S. I., and Alhussein, M. (2024). Individual household load forecasting using bi-directional lstm network with time-based embedding. ENERGY REPORTS, 11.0:3963–3975. Publisher: ELSEVIER.
6. Bai, Z. (2024). Residential electricity prediction based on ga-lstm modeling. ENERGY REPORTS, 11.0:6223–6232. Publisher: ELSEVIER.
7. Benali, A. A. E., Cafaro, M., Epicoco, I., Pulimeno, M., and Schioppa, E. J. (2024). Just in time transformers. IEEE ACCESS, 12.0:178751–178767. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC .
8. Cao, W., Liu, H., Zhang, X., Zeng, Y., and Ling, X. (2025). Short-term residential load forecasting based on the fusion of customer load uncertainty feature extraction and meteorological factors. SUSTAINABILITY, 17.0(3). Publisher: MDPI.
9. Chaianong, A., Winzer, C., and Gellrich, M. (2022). Impacts of traffic data on short-term residential load forecasting before and during the covid-19 pandemic. ENERGY STRATEGY REVIEWS, 43.0. Publisher: ELSEVIER.
10. Chen, Y., Obrecht, C., and Kuznik, F. (2024). Enhancing peak prediction in residential load forecasting with soft dynamic time wrapping loss functions. INTEGRATED COMPUTER-AIDED ENGINEERING, 31.0(3):327–340. USA Publisher: SAGE PUBLICATIONS INC.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

11. Cheng, L., Zang, H., Xu, Y., Wei, Z., and Sun, G. (2021). Probabilistic residential load forecasting based on micrometeorological data and customer consumption pattern. IEEE TRANSACTIONS ON POWER SYSTEMS, 36.0(4):3762–3775. Publisher: IEEEINST ELECTRICAL ELECTRONICS ENGINEERS INC.
12. Clarivate (2025). Web of science core collection. Accessed: 2025-06-28.
13. Dab, K., Henao, N., Nagarsheth, S., Dube, Y., Sansregret, S., and Agbossou, K. (2023). Consensus-based time-series clustering approach to short-term load forecasting for residential electricity demand. ENERGY AND BUILDINGS, 299.0. Publisher: ELSEVIER SCIENCE SA.
14. Dogra, A., Anand, A., and Bedi, J. (2023). Consumers profiling based federated learning approach for energy load forecasting. SUSTAINABLE CITIES AND SOCIETY, 98.0. Publisher: ELSEVIER.
15. Dong, B., Li, Z., Rahman, S. M. M., and Vega, R. (2016). A hybrid model approach for forecasting future residential electricity consumption. ENERGY AND BUILDINGS, 117.0:341–351. Publisher: ELSEVIER SCIENCE SA.
16. Dong, H., Zhu, J., Li, S., Miao, Y., Chung, C. Y., and Chen, Z. (2024). Probabilistic residential load forecasting with sequence-to-sequence adversarial domain adaptation networks. JOURNAL OF MODERN POWER SYSTEMS AND CLEAN ENERGY, 12.0(5):1559–1571. Publisher: STATE GRID ELECTRIC POWER RESEARCH INST.
17. Fan, C., Li, Y., Yi, L., Xiao, L., Qu, X., and Ai, Z. (2022). Multi-objective lstm ensemble model for household short-term load forecasting. MEMETIC COMPUTING, 14.0(1):115–132. Publisher: SPRINGER HEIDELBERG.
18. Fan, L., Li, H., and Zhang, X.-P. (2020). Load prediction methods using machine learning for home energy management systems based on human behavior patterns recognition. CSEE JOURNAL OF POWER AND ENERGY SYSTEMS, 6.0(3):563–571. Publisher: CHINA ELECTRIC POWER RESEARCH INST.
19. Fang, L. and He, B. (2023). A deep learning framework using multi-feature fusion recurrent neural networks for energy consumption forecasting. APPLIED ENERGY, 348.0. Publisher: ELSEVIER SCI LTD.
20. Fayaz, M. and Kim, D. (2018). A prediction methodology of energy consumption based on deep extreme learning machine and comparative analysis in residential buildings. ELECTRONICS, 7.0(10). Publisher: MDPI.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

21. Fekri, M. N., Grolinger, K., and Mir, S. (2022). Distributed load forecasting using smart meter data: Federated learning with recurrent neural networks. INTERNATIONAL JOURNAL OF ELECTRICAL POWER & ENERGY SYSTEMS, 137.0. Publisher: ELSEVIER SCI LTD.
22. Fekri, M. N., Patel, H., Grolinger, K., and Sharma, V. (2021). Deep learning for load forecasting with smart meter data: Online adaptive recurrent neural network. APPLIED ENERGY, 282.0(A). Publisher: ELSEVIER SCI LTD.
23. Flor, M., Herraiz, S., and Contreras, I. (2021). Definition of residential power load profiles clusters using machine learning and spatial analysis. ENERGIES, 14.0(20). Publisher: MDPI.
24. Forootani, A., Rastegar, M., and Sami, A. (2022). Short-term individual residential load forecasting using an enhanced machine learning-based approach based on a feature engineering framework: A comparative study with deep learning methods. ELECTRIC POWER SYSTEMS RESEARCH, 210.0. Publisher: ELSEVIER SCIENCE SA.
25. Gong, H., Alden, R. E., Patrick, A., and Ionel, D. M. (2022). Forecast of community total electric load and hvac component disaggregation through a new lstm-based method. ENERGIES, 15.0(9). Publisher: MDPI.
26. Han, F., Pu, T., Li, M., and Taylor, G. (2021). Short-term forecasting of individual residential load based on deep learning and k-means clustering. CSEE JOURNAL OF POWER AND ENERGY SYSTEMS, 7.0(2):261–269. Publisher: CHINA ELECTRIC POWER RESEARCH INST.
27. Harikrishnan, G. R., Sreedharan, S., and Binoy, C. N. (2025). Advanced short-term load forecasting for residential demand response: An xgboost-ann ensemble approach. ELECTRIC POWER SYSTEMS RESEARCH, 242.0. Publisher: ELSEVIER SCIENCE SA
28. Hou, T., Fang, R., Tang, J., Ge, G., Yang, D., Liu, J., and Zhang, W. (2021). A novel short-term residential electric load forecasting method based on adaptive load aggregation and deep learning algorithms. ENERGIES, 14.0(22). Publisher: MDPI.
29. Hribar, J., Fortuna, C., and Mohorcic, M. (2025). The role of age of information in enhancing short-term energy forecasting. ENERGY, 318.0. Publisher: PERGAMON-ELSEVIER SCIENCE LTD.
30. Imani, M. and Ghassemian, H. (2019). Residential load forecasting using wavelet and collaborative representation transforms. APPLIED ENERGY, 253.0. Publisher: ELSEVIER SCI LTD

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

31. International Energy Agency (IEA) (2022). Egypt - countries & regions. https:// www.iea.org. Accessed: January 9, 2025.
32. Irankhah, A., Yaghmaee, M. H., and Ershadi-Nasab, S. (2024). Optimized short-term load forecasting in residential buildings based on deep learning methods for different time horizons. JOURNAL OF BUILDING ENGINEERING, 84.0. Publisher: ELSEVIER.
33. Ismail, L., Materwala, H., and Dankar, F. K. (2024). Machine learning data-driven residential load multi-level forecasting with univariate and multivariate time series models toward sustainable smart homes. IEEE ACCESS, 12.0:55632– 55668. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
34. Jagait, R. K., Fekri, M. N., Grolinger, K., and Mir, S. (2021). Load forecasting under concept drift: Online ensemble learning with recurrent neural network and arima. IEEE ACCESS, 9.0:98992–99008. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
35. Jeyaraj, P. R. and Nadar, E. R. S. (2021). Computer-assisted demand-side energy management in residential smart grid employing novel pooling deep learning algorithm. INTERNATIONAL JOURNAL OF ENERGY RESEARCH, 45.0(5):7961–7973. Publisher: WILEY.
36. Ji, X., Huang, H., Chen, D., Yin, K., Zuo, Y., Chen, Z., and Bai, R. (2023). A hybrid residential short-term load forecasting method using attention mechanism and deep learning. BUILDINGS, 13.0(1). Publisher: MDPI.
37. Jiang, L., Wang, X., Li, W., Wang, L., Yin, X., and Jia, L. (2021). Hybrid multitask multi-information fusion deep learning for household short-term load forecasting. IEEE TRANSACTIONS ON SMART GRID, 12.0(6):5362–5372. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
38. Jiang, Y., Gao, T., Dai, Y., Si, R., Hao, J., Zhang, J., and Gao, D. W. (2022). Very short-term residential load forecasting based on deep-autoformer. APPLIED ENERGY, 328.0. Publisher: ELSEVIER SCI LTD.
39. Kaur, S., Bala, A., and Parashar, A. (2024). Ga-bilstm: an intelligent energy prediction and optimization approach for individual home appliances. EVOLVING SYSTEMS, 15.0(2):413–427. Publisher: SPRINGER HEIDELBERG
40. Kell, A., McGough, A. S., and Forshaw, M. (2018). Segmenting residential smart meter data for short-term load forecasting. E-ENERGY'18: PROCEEDINGS OF THE 9TH ACM INTERNATIONAL CONFERENCE ON FUTURE ENERGY SYSTEMS, pages 91–96. Backup Publisher: Assoc Comp Machinery;

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

ACM SIGCOMM; Deutsche Forschungsgemeinschaft; ProCom; EnBW; TransnetBW.

41. Khan, A.-N., Iqbal, N., Rizwan, A., Ahmad, R., and Kim, D.-H. (2021). An ensemble energy consumption forecasting model based on spatial-temporal clustering analysis in residential buildings. ENERGIES, 14.0(11). Publisher: MDPI
42. Kim, J.-Y. and Cho, S.-B. (2019). Electric energy consumption prediction by deep learning with state explainable autoencoder. ENERGIES, 12.0(4). Publisher: MDPI.
43. Kiprijanovska, I., Stankoski, S., Ilievski, I., Jovanovski, S., Gams, M., and Gjoreski, H. (2020). Houseec: Day-ahead household electrical energy consumption forecasting using deep learning. ENERGIES, 13.0(10). Publisher: MDPI.
44. Kong, W., Dong, Z. Y., Hill, D. J., Luo, F., and Xu, Y. (2018). Short-term residential load forecasting based on resident behaviour learning. IEEE TRANSACTIONS ON POWER SYSTEMS, 33.0(1):1087–1088. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
45. Kumaraswamy, S., Subathra, K., Dattathreya, Geeitha, S., Ramkumar, G., Metwally, A. S. M., and Ansari, M. Z. (2024). An ensemble neural network model for predicting the energy utility in individual houses. COMPUTERS & ELECTRICAL ENGINEERING, 114.0. Publisher: PERGAMON-ELSEVIER SCIENCE LTD.
46. Li, H., Heleno, M., Zhang, W., Garcia, L. R., and Hong, T. (2025). A cross-dimensional analysis of data-driven short-term load forecasting methods with large-scale smart meter data. ENERGY AND BUILDINGS, 344.0. Publisher: ELSEVIER SCIENCE SA.
47. Lin, W., Wu, D., and Jenkin, M. (2025). Electric load forecasting for individual households via spatial-temporal knowledge distillation. IEEE TRANSACTIONS ON POWER SYSTEMS, 40.0(1):572–584. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
48. Lotfipoor, A., Patidar, S., and Jenkins, D. P. (2024). Deep neural network with empirical mode decomposition and bayesian optimisation for residential load forecasting. EXPERT SYSTEMS WITH APPLICATIONS, 237.0(A). Publisher: PERGAMON-ELSEVIER SCIENCE LTD
49. Lu, Y., Wang, G., and Huang, S. (2022). A short-term load forecasting model based on mixup and transfer learning. ELECTRIC POWER SYSTEMS RESEARCH, 207.0 Publisher: ELSEVIER SCIENCE SA.
50. Manandhar, P., Rafiq, H., Rodriguez-Ubinas, E., and Palpanas, T. (2024). New forecasting metrics evaluated in prophet, random forest,

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

and long short-term memory models for load forecasting. ENERGIES, 17.0(23). Publisher: MDPI

51. Mansoor, H., Rauf, H., Mubashar, M., Khalid, M., and Arshad, N. (2021). Past vector similarity for short term electrical load forecasting at the individual household level. IEEE ACCESS, 9.0:42771–42785. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
52. Masood, Z., Gantassi, R., Ardiansyah, and Choi, Y. (2022). A multi-step time-series clustering-based seq2seq lstm learning for a single household electricity load forecasting. ENERGIES, 15.0(7). Publisher: MDPI.
53. Massidda, L. and Marrocu, M. (2018). Smart meter forecasting from one minute to one year horizons. ENERGIES, 11.0(12). Publisher: MDPI.
54. Moldovan, D. and Slowik, A. (2021). Energy consumption prediction of appliances using machine learning and multi-objective binary grey wolf optimization for feature selection. APPLIED SOFT COMPUTING, 111.0 Publisher: ELSEVIER
55. Muqtadir, A., Li, B., Ying, Z., Songsong, C., and Kazmi, S. N. (2025). Nowcasting the next hour of residential load using boosting ensemble machines. SCIENTIFIC REPORTS, 15.0(1). Publisher: NATURE PORTFOLIO.
56. Nguyen, T. T. Q., Tran, T. P. T., Debusschere, V., Bobineau, C., and Rigo-Mariani, R. (2020). Comparing high accurate regression models for short-term load forecasting in smart buildings. IECON 2020: THE 46TH ANNUAL CONFERENCE OF THE IEEE INDUSTRIAL ELECTRONICS SOCIETY, pages 1962–1967. Backup Publisher: IEEE Ind Elect Soc and Nangyang Technol Univ and Inst Elect & Elect Engineers and Smart Grid Power Elect Consortium Singapore ISSN: 1553-572X
57. Ozcan, A., Catal, C., and Kasif, A. (2021). Energy load forecasting using a dualstage attention-based recurrent neural network. SENSORS, 21.0(21). Publisher: MDPI.
58. Ozdemir, S., Demir, Y., and Yildirim, O. (2025). The effect of input length on prediction accuracy in short-term multi-step electricity load forecasting: A cnnlstm approach. IEEE ACCESS, 13.0:28419–28432. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
59. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ, 372:n71.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

60. Park, K.-J. and Son, S.-Y. (2023). Residential load forecasting using modified federated learning algorithm. IEEE ACCESS, 11.0:40675–40691. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
61. Pla, E. and Jimenez Martinez, M. (2023). Dealing with change: Retraining strategies to improve load forecasting in individual households under covid-19 restrictions. ENERGY REPORTS, 9.0(11):82–89. Publisher: ELSEVIER
62. Qu, X., Guan, C., Xie, G., Tian, Z., Sood, K., Sun, C., and Cui, L. (2023). Personalized federated learning for heterogeneous residential load forecasting. BIG DATA MINING AND ANALYTICS, 6.0(4, SI):421–432. Publisher: TSINGHUA UNIV PRESS.
63. Razghandi, M., Zhou, H., Erol-Kantarci, M., and Turgut, D. (2021). Short-term load forecasting for smart home appliances with sequence-to-sequence learning. IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS (ICC 2021). Backup Publisher: IEEE; Telus; Huawei; Ciena; Nokia; Samsung; Qualcomm; Cisco; Google Cloud ISSN: 1550-3607
64. Sajjad, M., Khan, Z. A., Ullah, A., Hussain, T., Ullah, W., Lee, M. Y., and Baik, S. W. (2020). A novel cnn-gru-based hybrid approach for short-term residential load forecasting. IEEE ACCESS, 8.0:143759–143768. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
65. Sakib, M., Siddiqui, T., Mustajab, S., Alotaibi, R. M., Alshareef, N. M., and Khan, M. Z. (2025). An ensemble deep learning framework for energy demand forecasting using genetic algorithm-based feature selection. PLOS ONE, 20(1):e0310465.
66. Sakuma, Y. and Nishi, H. (2022). Hierarchical multiobjective distributed deep learning for residential short-term electric load forecasting. IEEE ACCESS, 10.0:69950– 69962. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC
67. Shahsavari-Pour, N., Heydari, A., Keynia, F., Fekih, A., and Shahsavari-Pour, A. (2025). Building electrical consumption patterns forecasting based on a novel hybrid deep learning model. RESULTS IN ENGINEERING, 26.0. Publisher: ELSEVIER.
68. Shi, H., Xu, M., and Li, R. (2018). Deep learning for household load forecasting-a novel pooling deep rnn. IEEE TRANSACTIONS ON SMART GRID, 9.0(5):5271– 5280 Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC
69. Shi, J. and Wang, Z. (2022). A hybrid forecast model for household electric power by fusing landmark-based spectral clustering and deep learning. SUSTAINABILITY, 14.0(15). Publisher: MDPI.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

70. Shi, Y. and Xu, X. (2022). Deep federated adaptation: An adaptative residential load forecasting approach with federated learning. SENSORS, 22.0(9). Publisher: MDPI.
71. Sinha, A., Tayal, R., Vyas, A., Pandey, P., and Vyas, O. P. (2021). Forecasting electricity load with hybrid scalable model based on stacked non linear residual approach. FRONTIERS IN ENERGY RESEARCH, 9.0. Publisher: FRONTIERS MEDIA SA.
72. Sulaiman, S. M., Jeyanthy, P. A., Devaraj, D., and Shihabudheen, K., V. (2022). A novel hybrid short-term electricity forecasting technique for residential loads using empirical mode decomposition and extreme learning machines. COMPUTERS & ELECTRICAL ENGINEERING, 98.0. Publisher: PERGAMON-ELSEVIER SCIENCE LTD.
73. Taik, A. and Cherkaoui, S. (2020). Electrical load forecasting using edge computing and federated learning. ICC 2020 - 2020 IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS (ICC). Backup Publisher: IEEE; Huawei; ZTE; Qualcomm ISSN: 1550-3607
74. Teperikidis, L., Boulmpou, A., Papadopoulos, C., and Biondi-Zoccai, G. (2024). Using chatgpt to perform a systematic review: a tutorial. Minerva cardiology and angiology, 72.
75. Tigo (2023). A solar powered future: Residential adoption on a global scale. Accessed: 2025-07-01.
76. Truong, L. H. M., Chow, K. H. K., Luevisadpaibul, R., Thirunavukkarasu, G. S., Seyedmahmoudian, M., Horan, B., Mekhilef, S., and Stojcevski, A. (2021). Accurate prediction of hourly energy consumption in a residential building based on the occupancy rate using machine learning approaches. APPLIED SCIENCES-BASEL, 11.0(5). Publisher: MDPI.
77. Ullah, F. U. M., Khan, N., Hussain, T., Lee, M. Y., and Baik, S. W. (2021). Diving deep into short-term electricity load forecasting: Comparative analysis and a novel framework. MATHEMATICS, 9.0(6). Publisher: MDPI.
78. Widmer, F., Nowak, S., Bowler, B., Huber, P., and Papaemmanouil, A. (2023). Datadriven comparison of federated learning and model personalization for electric load forecasting. ENERGY AND AI, 14.0. Publisher: ELSEVIER.
79. Xia, Z., Zhang, R., Ma, H., and Saha, T. K. (2024). Day-ahead electricity consumption prediction of individual household capturing peak consumption pattern. IEEE TRANSACTIONS ON SMART GRID, 15.0(3):2971–2984. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.

ISSN: 1857-7881 (Print)
e-ISSN 1857-7431 (Online)

80. Xu, C., Li, C., and Zhou, X. (2022). Interpretable lstm based on mixture attention mechanism for multi-step residential load forecasting. ELECTRONICS, 11.0(14). Publisher: MDPI.
81. Yang, W., Shi, J., Li, S., Song, Z., Zhang, Z., and Chen, Z. (2022). A combined deep learning load forecasting model of single household resident user considering multi-time scale electricity consumption behavior. APPLIED ENERGY, 307.0. Publisher: ELSEVIER SCI LTD.
82. Yousaf, A., Asif, R. M., Shakir, M., Rehman, A. U., and S. Adrees, M. (2021). An improved residential electricity load forecasting using a machine-learning-based feature selection approach and a proposed integration strategy. SUSTAINABILITY, 13.0(11). Publisher: MDPI.
83. Yuan, L., Ma, J., Gu, J., Wen, H., and Jin, Z. (2020). Featuring periodic correlations via dual granularity inputs structuredrnnsensemble load forecaster. INTERNATIONAL TRANSACTIONS ON ELECTRICAL ENERGY SYSTEMS, 30.0(11). Publisher: WILEY.
84. Zang, H., Xu, R., Cheng, L., Ding, T., Liu, L., Wei, Z., and Sun, G. (2021). Residential load forecasting based on lstm fusing self-attention mechanism with pooling. ENERGY, 229.0. Publisher: PERGAMON-ELSEVIER SCIENCE LTD
85. Zhang, X.-Y., Cordoba-Pachon, J.-R., Guo, P., Watkins, C., and Kuenzel, S. (2024). Privacy-preserving federated learning for value-added service model in advanced metering infrastructure. IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, 11.0(1):117–131. Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC.
86. Zhao, W., Li, T., Xu, D., and Wang, Z. (2024). A global forecasting method of heterogeneous household short-term load based on pre-trained autoencoder and deep-lstm model. ANNALS OF OPERATIONS RESEARCH, 339.0(1-2, SI):227–259. Publisher: SPRINGER.
87. Zhu, J., Miao, Y., Dong, H., Li, S., Chen, Z., and Zhang, D. (2024). Short-term residential load forecasting based on k-shape clustering and domain adversarial transfer network. JOURNAL OF MODERN POWER SYSTEMS AND CLEAN ENERGY, 12.0(4):1239–1249. Publisher: STATE GRID ELECTRIC POWER RESEARCH INST