A Review on Deep Learning-based Electrical Load Forecasting: From Perspectives of Learning Paradigms and Foundation Models

Pengfei Zhao, Weihao Hu, Di Cao, Zhaochen Dong, Yaqi Zeng, Qi Huang, and Zhe Chen

Abstract—**Electrical load forecasting (ELF) plays a critical role in the planning and operation of modern power systems. As energy demand patterns grow more complex, deep learning (DL) techniques, and more recently, foundation models (FMs), have emerged as powerful tools for modeling temporal dynamics and integrating heterogeneous inputs. In practice, the effectiveness of these models depends not only on their architectures but also on the learning paradigms that determine how they are trained, adapted, and deployed. However, most existing surveys focus solely on network architectures, with limited attention to the underlying paradigms. To this end, we survey the DL-based ELF from perspectives of learning paradigms and FMs. It organizes the literature into four orthogonal paradigms: task-tuned offline learning, adaptive DL, collaborative DL, and general-purpose DL. This paradigm-centric perspective enables a unified understanding of how DL methods evolve to meet the challenges of ELF. It also provides a natural framework to incorporate FMs as the latest advancement in this trajectory. Finally, key challenges are provided, and research opportunities are highlighted to inform future directions.**

Index Terms—**Electrical load forecasting, power system, deep learning, learning paradigm, foundation model.**

I. Introduction

POWER systems are undergoing a profound transformation in response to the global transition toward cleaner and more decentralized energy sources. On the supply side, the growing penetration of renewable energy sources introduces substantial variability and intermittency to the grid. It is reported that global renewable power capacity increased by about 585 GW in 2024, which lifted the total to 4448 GW and marked an annual growth rate of 15.1% [1]. In parallel, the proliferation of distributed energy resources such as rooftop solar and community microgrids alters power flow dynamics and complicates system-level coordination. On the demand side, the electrification of sectors such as transportation, heating, and industrial processes reshapes consumption patterns, introducing new spatiotemporal variability. In 2024, the global electricity consumption grew by about 4.3%, nearly doubling the decade-average rate [2]. Moreover, the increasing adoption of residential loads with stochastic behavior and the growing participation in demand response programs further amplify the complexity and uncertainty of load profiles [3].

As a result, the grid becomes more dynamic, less predictable, and increasingly reliant on timely, data-driven decision-making. In this context, electrical load forecasting (ELF) has become more critical than ever. ELF refers to the process of predicting future electricity demand based on historical consumption patterns and relevant contextual information. It is a fundamental tool for ensuring the efficient, stable, and economic operation of modern power systems. Based on published analyses, for a power system with a peak load of 5000 MW operating at approximately 5% mean absolute percentage error (MAPE), each 1% improvement in forecasting accuracy delivers about 0.6-1.6 million US dollars per year, equivalent to a 0.12%-0.30% reduction in variable generation cost [4]. Achieving high-precision forecasts, however, is far from straightforward. It requires the ability to model intricate temporal dependencies and contextual influences such as weather variability, calendar effects, and consumer behavior [5]. In addition, the increasing heterogeneity and decentralization of power systems necessitate forecasting methods that are scalable, adaptive, and capable of handling high-resolution multi-source data.

Fortunately, recent advancements in data collection infrastructure such as smart meters and Internet of Things (IoT)-enabled sensors have made large volumes of fine-grained consumption data readily available. At the same time, the rapid progress of machine learning (ML) and artificial intelli-

Manuscript received: August 14, 2025; revised: November 18, 2025; accepted: January 23, 2025. Date of CrossCheck: January 23, 2025. Date of online publication: February 27, 2026.

This work was supported in part by the National Natural Science Foundation of China (No. 72401055), in part by the China Postdoctoral Science Foundation (No. 2023M730495), and in part by the Central University Basic Research Fund Project (No. ZYGX2024J014).

P. Zhao, W. Hu, D. Cao (corresponding author), and Z. Dong are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: zhaopf@std.uestc.edu.cn; whu@uestc.edu.cn; caodi@uestc.edu.cn;202521040209@std.uestc.edu.cn).

Y. Zeng is with the Glasgow College, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: 2024190905032@std.uestc.edu.cn).

Q. Huang is with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and he is also with the Institute of Scientific and Technical Information of China, Beijing 100038, China (e-mail: hwong@uestc.edu.cn).

Z. Chen is with the Department of Energy Technology, Aalborg University, DK-9220 Aalborg, Denmark (e-mail: zch@et.aau.dk).

DOI: 10.35833/MPCE.2025.000740

gence (AI), especially deep learning (DL), has empowered researchers to move beyond handcrafted models toward data-driven methods that can automatically extract complex patterns and hidden representations [6]. In principle, the more effectively a model can learn such representations, the better its forecasting performance. Yet the difficulty in learning useful representations varies widely across ELF scenarios. Some applications must work with limited data, others contend with continuous distribution shifts, and many face privacy restrictions across regions and utilities. These differences mean that improving ELF is not simply a matter of designing better neural architectures, which also requires thoughtful choices about how forecasting models learn, adapt, and generalize from data in the presence of these challenges [7]. As a result, a diverse set of learning paradigms has emerged in ELF. However, existing surveys have paid little attention to the learning paradigms.

A diverse array of forecasting methods has been developed over the past decades. These methods can be broadly classified into statistical methods, classical ML models, and DL methods. Early statistical methods, including autoregressive integrated moving average (ARIMA), exponential smoothing, and linear regression, are widely used due to their simplicity and interpretability [8]-[10]. These methods are effective in capturing seasonal trends and short-term fluctuations under stationary conditions, but they struggle with nonlinear interactions between input variables and output targets. Classical ML methods such as support vector machines, decision trees, and ensemble methods (e.g., random forests and gradient boosting) relax these assumptions and enable better modeling of nonlinear relationships [11], [12]. These demonstrate better performance across various forecasting tasks, but the predictive accuracy often depends on carefully engineered features and domain expertise. Moreover, most of these methods lack explicit mechanisms to capture temporal dependencies, which makes them less suitable for long-sequence forecasting scenarios.

DL methods has emerged as a powerful paradigm for ELF. Unlike classical ML methods, DL methods can automatically learn meaningful representations from raw data and capture long-term dependencies through sequence modeling. A wide range of DL methods have been explored in the literature, including long short-term memory (LSTM) networks [13], gated recurrent units (GRUs) [14], convolutional neural networks (CNNs) [15], temporal convolutional networks (TCNs) [16], Transformers [17], and graph neural networks (GNNs) [18], [19]. These have shown strong performance across different forecasting horizons and input modalities. Most DL methods are tailored to specific ELF tasks, with architecture, input representation, and training strategy designed according to the characteristics of the target problem. While this specialization allows for flexibility and task-specific optimization, it may also limit the reusability and scalability of models across different contexts.

In recent years, inspired by the success of foundation models (FMs) in fields such as natural language processing (NLP) and computer vision (CV), the idea of building general-purpose models for ELF has started to gain attention. FMs such as large language models (LLMs) in NLP [20] are typically pre-trained on large-scale and diverse datasets to learn transferable representations that support a wide range of downstream tasks with minimal adaptation. Applied to the ELF domain, FMs in ELF (termed as ELFFMs) aim to develop unified forecasting frameworks that can effectively incorporate heterogeneous data sources such as historical load profiles, numerical weather prediction (NWP) outputs, calendar information, and even satellite imagery. Early studies suggest that ELFFMs have the potential to enhance generalization across geographic regions and temporal scales, reduce dependence on manual feature engineering, and support multi-objective forecasting tasks [21]-[23].

Given the breadth and rapid development of ELF, a number of surveys have aimed to synthesize the field by classifying modeling strategies, application domains, and evaluation protocols. Broadly, these existing surveys fall into three groups. The first group focuses on methodological taxonomies. References [24] and [25] examine the progression from traditional statistical models to AI techniques. More comprehensive surveys [26]-[28] extend these discussions by comparing single and hybrid models across different dimensions. Specific attention is given to probabilistic forecasting in [29], which provides a tutorial-style synthesis of modeling principles and evaluation metrics. The second group organizes surveys around application-specific challenges. Reference [30] investigates ELF methods in integrated energy systems (IESs), where interactions between electricity, heating, and gas introduce new layers of complexity. References [31] and [32] focus on microgrid environments, highlighting the role of forecast resolution and decentralized data availability. Others such as [33] and [34] analyze forecasting in commercial buildings and smart cities. Meanwhile, [35] and [36] tackle emerging themes like interpretability, which are increasingly crucial for deploying AI models in mission-critical settings. References [37] and [38] contextualize ELF in low-voltage and smart grid environments, respectively. The third group centers on DL methods. References [5], [39], and [40] provide detailed comparisons of RNN, LSTM, CNN, and attention-based architectures, covering their design principles, data preprocessing strategies, and limitations in generalization.

Despite these valuable contributions, existing surveys on DL-based ELF often focus narrowly on model architectures while overlooking key learning paradigms and training strategies. A paradigm-based perspective offers a deeper understanding of how forecasting systems learn and adapt. In fact, model performance in ELF depends not only on architectural design but also on the way learning is conducted. It may be trained once on historical data, updated continuously with new observations, adapted collaboratively across regions, or pre-trained for generalization across multiple tasks. Viewing existing studies through this lens clarifies why certain methods succeed under specific conditions but fail elsewhere, and reveals the trade-offs among accuracy, adaptability, scalability, and cost [41]. In addition, FMs have shown promise in improving generalization, reducing dependence on handcrafted features, and handling multiple forecasting objectives, but little effort has been made to systematically summarize the

FMs in ELF. This survey aims to bridge these gaps by mapping the field from the lens of learning paradigms of DL. The FMs are naturally included as a general-purpose DL paradigm. A detailed explanation of the paradigm-based taxonomy is provided in Supplementary Material A. We provide the comprehensive survey that organizes DL-based forecasting through the lens of learning paradigms, a perspective that has not been systematically explored in ELF or in other computer science domains. The main contributions are as follows:

1) A comprehensive and up-to-date survey is provided that reviews DL-based ELF from lens of learning paradigms. This allows us to move beyond architecture-centric taxonomies and instead classify methods by how they are trained, adapted, and deployed in response to diverse forecasting challenges.

2) A first systematic synthesis of FMs in ELF is provided, detailing the architectures, pre-training strategies, and adaptation methods. This allows us to understand how FMs can reshape ELF from task-specific solutions to general-purpose frameworks.

3) An analysis of challenges and future research opportunities is provided, including issues related to human and environmental uncertainties, downstream task integration, and DL reliability. This provides a roadmap for developing more adaptive, robust, and trustworthy forecasting systems.

The remainder of this paper is organized as follows. Sections II-V survey task-tuned offline learning, adaptive DL, collaborative DL, and general-purpose DL, respectively. Section VI discusses challenges and opportunities. Section VII concludes the paper.

II. Task-tuned Offline Learning

Task-tuned offline learning is the basic learning paradigm where forecasting models are trained offline on historical data for a single and fixed forecasting task without knowledge sharing across different tasks. This setting assumes that the underlying load data distribution remains stationary or changes slowly enough that a static model can generalize well during deployment. The forecasting models are trained end-to-end using supervised learning, with the objective of minimizing a predefined loss function [42], [43].

A. Learning Paradigm of Task-tuned Offline Learning

In the task-tuned offline learning, the forecasting model is trained using supervised learning on a fixed dataset for a specific task. Given the k-step historical load sequence $\{\boldsymbol{y}_{t-k}, \boldsymbol{y}_{t-k+1}, \ldots, \boldsymbol{y}_t\}$ and auxiliary covariates $\{\boldsymbol{v}_{t-k}, \boldsymbol{v}_{t-k+1}, \ldots, \boldsymbol{v}_{t+h}\}$, the forecasting model f_θ, parameterized by θ, learns to estimate the future h-step load values $\{\hat{\boldsymbol{y}}_{t+1}, \hat{\boldsymbol{y}}_{t+2}, \ldots, \hat{\boldsymbol{y}}_{t+h}\}$ as:

$$\hat{\boldsymbol{y}}_{t+1:t+h} = f_\theta\left(\boldsymbol{y}_{t-k:t}, \boldsymbol{v}_{t-k:t+h}\right) \tag{1}$$

The forecasting model is optimized by minimizing a loss function between the predicted and true future loads over the training set as:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} l\left(f_\theta\left(\boldsymbol{y}^{(i)}_{t-k:t}, \boldsymbol{v}^{(i)}_{t-k:t+h}\right), \boldsymbol{y}^{(i)}_{t+1:t+h}\right) \tag{2}$$

where $\mathcal{L}(\cdot)$ is the empirical training loss of the forecasting model; $l(\cdot)$ is the point-wise loss function; N is the number of training samples; $\boldsymbol{y}^{(i)}_{t-k:t}$ denotes the historical load sequence for the i^{th} sample; $\boldsymbol{v}^{(i)}_{t-k:t+h}$ denotes the associated exogenous variable for the i^{th} sample; and $\boldsymbol{y}^{(i)}_{t+1:t+h}$ denotes the ground-truth future load trajectory for the i^{th} sample.

In practice, the design of the loss function $l(\cdot)$ is not trivial, as it determines what aspect of forecasting accuracy the model prioritizes. In most forecasting studies, l is chosen as a standard regression loss such as mean squared error (MSE) or mean absolute error (MAE), which measures the average deviation between predicted and true loads. Variants such as weighted MSE or MAPE are also used to emphasize peak hours or critical periods where accurate prediction is particularly important. When uncertainty quantification is desired, probabilistic losses are adopted instead. Typical examples include the pinball loss for quantile regression and the negative log-likelihood (NLL) for parametric distribution modeling [44]. Moreover, recent research has explored cost-aware or decision-focused loss functions, where the conventional statistical objective is replaced by an operational cost surrogate so that model training directly optimizes downstream decision performance rather than mere prediction accuracy [45].

θ is updated iteratively using gradient-based optimization. Specifically, backpropagation is used to compute the gradient of the loss with respect to the model parameters, as shown in (3). $\nabla_\theta \mathcal{L}(\theta)$ is obtained via automatic differentiation through the model layers.

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta) \tag{3}$$

where η is the learning rate; and ∇_θ is the gradient operator. The training typically proceeds in mini-batches with stochastic gradient descent (SGD) or its variants (e.g., Adam) [46].

Since f_θ is task-tuned, a critical step is to carefully design the neural network architecture in accordance with the specific characteristics of the forecasting task. A critical factor that should be considered is the nature of the input data of the specific task. If the forecasting task involves time series data with temporal dependencies, architectures that can effectively capture sequential patterns are critical. LSTMs [47], [48], GRUs [49], TCN [50], [51], or attention networks [52]-[54] are well-suited for such tasks, as they are specifically designed to process and model sequences over time. If the data involve multiple related inputs across different regions or contexts, CNNs and GNNs can be helpful. CNNs are effective at detecting spatial relationships and can process data with multiple spatial features such as time-series data from different locations [54]. On the other hand, GNNs excel in capturing the dependencies between nodes in a graph, making them ideal for tasks where data are inherently structured as a graph [55]. If the task involves high-dimensional input features or requires the integration of different data modalities, hybrid forecasting models may be more suitable. These hybrid forecasting models combine multiple architectural components to leverage the strengths of different methods. For example, combining CNNs or GNNs for spatial feature extraction with RNNs, TCNs, or Transformers for capturing temporal dependencies can be highly effective. By focusing on the relationships between all features simultaneously, complex dependencies can be modeled across both space and time.

Once the task-tuned forecasting model is trained and validated, its parameters are fixed, and it is then integrated into the production system. During deployment, the task-tuned forecasting model receives real-time input data, but its weights and architecture remain unchanged. The deployed model operates in inference mode, where it generates predictions based on the input data it receives, without further modification to the parameters unless retrained in the future.

B. Discussion on Task-tuned Offline Learning

1) Reliability in Stable Environment

Task-tuned offline learning is one of the most widely used paradigms in DL-based ELF. Its appeal lies in the simplicity and effectiveness of training models offline on a fixed dataset for a specific forecasting task. By directly optimizing the model for a well-defined objective, reliable performance is ensured, when the underlying data distribution is relatively stable [56]. In practical settings where historical patterns dominate and disruptions are infrequent, task-tuned forecasting models can achieve high accuracy with minimal complexity. The straightforward nature of the end-to-end supervised learning process also allows for efficient integration into existing forecasting systems.

2) Vulnerability to Concept Drift and Lack of Task Flexibility

The most notable challenge of task-tuned offline learning is its reliance on the assumption that the data distribution remains stationary. In practice, this is rarely the case, particularly in modern power systems, which are subject to rapid changes driven by various factors. Once trained, the task-tuned forecasting model lacks the ability to adapt in real time, making it vulnerable to performance degradation under distribution shifts [57]. Online learning (OL) can address this challenge by continuously updating model parameters as new data arrive. In addition, task-tuned offline learning often struggles to generalize when the forecasting task changes, e. g., being transferred to a different region, horizon, or energy type. Models trained under fixed statistical conditions cannot easily transfer learned dependencies to new environments, especially when only limited samples are available in emerging tasks [58]. Transfer learning (TL) and FMs are therefore regarded as potential solutions to this generalization problem.

3) Limited Cross-task Information Sharing

Because task-tuned offline learning optimizes each forecasting task in isolation, it depends solely on its own task-specific data and overlooks the correlations that often exist among related tasks. These shared patterns, however, carry important information that can improve both data efficiency and predictive accuracy. For example, in IES, electricity, heating, cooling, and gas are interconnected [59]. Capturing the coupling among multiple energy loads allows collaborative forecasting models to utilize shared information and thereby improve forecasting accuracy. Collaborative learning offers a natural framework for this purpose, as it facilitates knowledge sharing across related tasks while maintaining task-specific distinctions.

III. Adaptive DL

Task-tuned forecasting models are typically trained offline on sufficient historical data and assume a stationary mapping between inputs and future demand. However, this assumption often breaks down in practice. On the one hand, electricity consumption patterns are inherently dynamic and shaped by evolving weather conditions, seasonal trends, user behaviors, and disruptive events such as extreme weather or policy shifts. As a result, static models struggle to maintain accuracy once deployed. This challenge becomes more acute in modern power systems characterized by high renewable penetration and low inertia, where accurate short-term forecasting is critical for grid stability [60]. On the other hand, many deployment scenarios such as newly built households suffer from sparse historical data. This makes it difficult to train specialized models from scratch. These two challenges necessitate learning paradigms that can adapt across time and tasks, rather than rely on one-shot training.

Adaptive DL addresses this need by equipping forecasting models with mechanisms to respond to changing environments, encompassing two paradigms: ① TL focuses on cross-task generalization, which allows forecasting models to transfer knowledge from well-trained domains to novel domains; and ② OL emphasizes continual adaptation, enabling forecasting models to incrementally update their parameters as new data streams in [61]. This section, therefore, reviews these two adaptive DL paradigms. A literature summary of adaptive DL is shown in Table I.

TABLE I
LITERATURE SUMMARY OF ADAPTIVE DL

Ref.	Paradigm	Framework	Application
[62]	TL	Residual network (ResNet) + biphase weight adjusting	Residential ELF with limited data
[63]	TL	DL + RL	ELF with insufficient data
[64]	TL	LSTM + fine-tune	ELF with limited data
[65]	TL	Sequence-to-Sequence (Seq2 Seq) + domain adaptation	Residential ELF with limited data
[66]	TL	Clustering + domain adaptation + fine-tune	Residential ELF with limited data
[67]	TL	Adversarial network	Prediction of masked load
[68]	TL	Model-agnostic meta-learning (MAML)	Residential ELF with limited data
[69]	TL	ResNet + Gaussian process (GP)	Load pattern changes during abnormal events
[70]	TL	Graph neural network (GNN) + attention	Residential ELF with limited data
[71]	OL	Least absolute shrinkage and selection operator (LASSO)-LSTM + buffering-tuning	Concept drift during abnormal events
[72]	OL	Echo state network	Online learning and forecasting of residential load for edge computing
[73]	OL	Multi-fidelity data fusion network	Data gap
[74]	OL	Offline soft spiking network + online GP	Load pattern fluctuations of residents
[75]	OL	Ensemble of DL and ML models	Incremental training with new samples

A. TL

For newly-deployed smart meters or emerging microgrids, collecting sufficient labeled data is often a slow and costly process. This data scarcity becomes a major bottleneck for training DL models, which typically rely on large datasets to capture complex patterns [54]. A natural workaround is to ask: if similar forecasting tasks have already been solved in other domains, can we transfer some of that knowledge to the current scenario? This intuition underpins TL, a learning paradigm that allows a model to reuse knowledge from a source domain to adapt to a target domain [76]. In TL, the model is first trained on a source domain where data are plentiful, and then fine-tuned or adapted to a different but related target domain with limited data. This part reviews how TL has been used in ELF to address generalization and data limitations.

1) Learning Paradigm of TL

TL encompasses a broad range of strategies that reuse knowledge from a source domain to improve generalization in a target domain. Let the source domain be denoted as $\mathcal{D}_S = \left\{\left(\boldsymbol{x}_{i,S}, \boldsymbol{y}_{i,S}\right)\right\}_{i=1}^{N_S}$ and the target domain as $\mathcal{D}_T = \left\{\left(\boldsymbol{x}_{j,T}, \boldsymbol{y}_{j,T}\right)\right\}_{j=1}^{N_T}$, where $\boldsymbol{x}_{i,S}$ and $\boldsymbol{y}_{i,S}$ represent the input feature and forecasting target of the i^{th} sample of the source domain, respectively; $\boldsymbol{x}_{j,T}$ and $\boldsymbol{y}_{j,T}$ represent the input feature and forecasting target of the j^{th} sample of the target domain, respectively; and N_S and N_T are the numbers of samples in the source and target domains, respectively, with $N_T \ll N_S$. TL begins by training a forecasting model f_{θ_S} on the source domain. Then, part or all of the learned parameters θ_S are transferred to the target domain, and a new model f_{θ_T} is adapted by minimizing the forecasting loss on $\mathcal{D}_T$. In ELF, the following strategies for TL are often used.

1) Fine-tuning: fine-tuning remains the most widely adopted strategy in ELF due to its simplicity and effectiveness. The basic idea is to initialize the target model with parameters from a pre-trained source model, then continue training on limited target-domain data. This strategy is useful when the source and target share similar dynamics but differ in specific features such as consumption behavior. For example, [62] introduces a Bayesian fine-tuning that adaptively assigns weights for forecasting models that are trained on different source households. Reference [63] uses reinforcement learning to choose among multiple candidate models trained on different load data sources before fine-tuning the best one. Personalized ensemble methods also exist, where forecasting models pre-trained on users with similar patterns are fine-tuned and aggregated for better robustness [64].

In fine-tuning, the loss function is usually kept consistent with the source task to preserve the learned feature representations, but its role becomes more subtle. When the source and target domains are highly similar, the same statistical loss (e.g., MSE or MAE) is retained and only a small learning rate is applied to avoid catastrophic forgetting. When domain gaps exist, additional regularization terms are often introduced to stabilize transfer and prevent overfitting to the small target dataset. Typical examples include L_2 weight penalties on deviations from the source parameters [77], feature-level alignment losses that match hidden representations between domains [78], [79], or task-weighted composite losses that balance old and new objectives.

2) Adversarial TL: in many real-world cases, the source and target domains differ not just in size but also in distributional properties. Fine-tuning alone may fail to bridge such gaps. Adversarial TL tackles this by explicitly aligning the feature distributions of source and target through a game-theoretic training setup [80]. Specifically, a shared encoder f_{θ_e} with parameter θ_e is trained to extract domain-invariant representations, a task predictor f_{θ_c} with parameter θ_c performs forecasting, and a domain discriminator f_{θ_d} with parameter θ_d attempts to distinguish between domains. The resulting min-max optimization problem is formulated as:

$$\min_{\theta_e,\theta_c} \max_{\theta_d} \left(\mathcal{L}_{task} + \lambda \mathcal{L}_{adv}\right) \tag{4}$$

$$\mathcal{L}_{task} = \sum_{i=1}^{N_S} \mathcal{L}\left(f_{\theta_c}\left(f_{\theta_e}\left(\boldsymbol{x}_{i,S}\right)\right), \boldsymbol{y}_{i,S}\right) \tag{5}$$

$$\mathcal{L}_{adv} = \sum_i \ln f_{\theta_d}\left(f_{\theta_e}\left(\boldsymbol{x}_{i,S}\right)\right) + \sum_j \ln\left(1 - f_{\theta_d}\left(f_{\theta_e}\left(\boldsymbol{x}_{j,T}\right)\right)\right) \tag{6}$$

where $\mathcal{L}_{task}$ is the source-domain forecasting loss; $\mathcal{L}_{adv}$ is the adversarial loss for domain discrimination; λ controls the trade-off between predictive accuracy and domain alignment; and i and j are the samples from $\mathcal{D}_S$ and $\mathcal{D}_T$, respectively. In the context of short-term ELF, [65] applies this strategy by coupling an LSTM-based Seq2Seq model with adversarial learning, which effectively aligns load features across provinces with differing climate and consumption profiles. Recognizing that not all source regions are equally useful for transfer, [66] introduces a K-shape clustering technique to pre-select those source domains whose load dynamics are most similar to the target domain. By narrowing the transfer scope to structurally compatible clusters, the technique avoids negative transfer and improves warm-start training effectiveness. At a more granular level, [67] explores intra-user domain shift, where the unmasked and masked versions of a user's residential load are treated as distinct domains. An adversarial training scheme is used to bridge the internal shift, enabling the forecasting model to remain accurate even when critical signals are unavailable or distorted.

3) Meta-learning: TL often involves transferring knowledge from many source domains to a target domain. This raises a key question: can we build a forecasting model capable of quickly adapting to a new domain with very few labeled samples? The solution is meta-learning, which goes beyond simply applying knowledge from one task to another and enables models to "learn how to learn," such that the forecasting model can adapt rapidly to new tasks with minimal data. MAML is one of the most widely used meta-learning strategy [81]. It optimizes an initialization of parameters that can be rapidly fine-tuned across a distribution of tasks.

In ELF, MAML has been applied to the scenarios such as newly-built households [68] and regional cold-start forecasting [14]. Other studies [82]-[84] combine meta-learning with ensemble or voting strategies to automate model selection and adaptation based on evolving load patterns.

2) Discussion of TL

Some key issues of applying TL to ELF are discussed below.

1) Knowledge transfer and efficiency: when acquiring vast amounts of labeled data can be expensive and time-consuming, TL offers a pragmatic solution by enabling the use of models that have been pre-trained on similar datasets. This not only reduces the data requirements but also accelerates the model development process, making it more efficient and cost-effective [69]. This is beneficial for applications such as microgrid forecasting or newly-deployed smart meters, where historical data are often limited [70].

2) Data availability: by reusing representations learned from data-rich source domains, TL can help bootstrap accurate forecasting models even when the target domain lacks sufficient historical records. Namely, most existing methods still assume that at least a small amount of labeled data is available for fine-tuning. In more extreme cases, however, even a small number of labeled samples is unavailable [85]. This could cause the failure of the traditional strategies for TL. Recent advances in FMs offer a new way forward. Instead of transferring from a specific source domain, FMs are trained on diverse data at scale and designed to generalize across tasks. With appropriate prompting or lightweight adaptation, they could potentially forecast unseen targets without any additional labels [86].

3) Data privacy: another important consideration in real-world deployment is data privacy. Many strategies for TL promise to improve forecasting accuracy without requiring centralized access to user data. Typically, the forecasting model is pre-trained on source domains and only its parameters are transferred. However, recent research highlights that even forecasting model parameters may leak private information [87]. This becomes a concern in ELF, where patterns in weights or gradients may indirectly reflect user behaviors. To mitigate such risks, privacy-preserving strategies for TL are gaining attention [88].

B. OL

OL is a dynamic training paradigm where the forecasting model is updated incrementally as new data arrive. This allows it to dynamically adjust to changes in the underlying data distribution.

1) Learning Paradigm of OL

At each time t, the forecasting model receives the input $\boldsymbol{x}_t$, which represents the features or observations at time t, and generates a prediction $\hat{\boldsymbol{y}}_t = f_{\theta_t}(\boldsymbol{x}_t)$. After observing the true output $\boldsymbol{y}_t$, the forecasting model incurs a prediction error, quantified by the loss function $\mathcal{L}(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t)$. The parameters are then updated to minimize this forecasting error, using an update rule that can be derived from various optimization methods, including but not limited to gradient-based methods. Specifically, the parameters are adjusted as [89]:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \mathcal{L}\left(f_{\theta_t}(\boldsymbol{x}_t), \boldsymbol{y}_t\right) \tag{7}$$

where η_t is the learning rate at time t. The learning rate may be constant, decay over time, or be adaptively adjusted based on the model performance. This update rule allows the forecasting model to gradually learn and adapt as new data become available. The objective of OL is to minimize the cumulative loss over time.

In online learning, the instantaneous loss typically follows conventional regression forms such as MSE or MAE. What distinguishes OL from static training is how these losses are accumulated and weighted over time. A common strategy is to apply recency-weighted or exponentially decaying losses that assign greater importance to recent samples, thereby capturing concept drift and evolving load patterns in real time [90]. Some studies further reinterpret the online learning objective through the lens of regret minimization, where the loss quantifies the deviation from an oracle predictor rather than merely measuring past prediction errors [91].

The challenge in OL lies in balancing two competing objectives: ① plasticity, which is the forecasting ability of the model to quickly adapt to new information; and ② stability, which is the ability to maintain previously learned knowledge without overfitting to recent data [61]. To achieve such a trade-off, different methods are used. One is online-only learning, where the forecasting model is trained directly on real-time data without involving offline training [71], [72]. However, a forecasting model that is too shallow may fail to capture complex patterns, whereas an overly deep forecasting model can become prone to overfitting to recent data. To address these challenges, techniques like Hedge Backpropagation have been introduced, which adaptively adjust the model capacity and balance shallow and deep network representations as the learning process evolves [73]. A main limitation of online-only learning is that it fails to fully exploit the benefits of offline learning, which means that valuable insights from historical data are not fully utilized. To address this, offline-online methods have been introduced, which combine the advantages of both offline and online learning paradigms. In [74], the parameters of a spiking neural network trained offline are used as the initial parameters for kernel learning online. Reference [75] first trains various basic models offline and then assigns adaptive weights online to the offline-trained models.

2) Discussion on OL

Some key issues of OL are discussed below.

1) Adaptation to dynamic environments: OL is particularly useful for forecasting tasks where the consumption patterns are subject to frequent changes due to factors like unexpected demand spikes, policy changes, or environmental events. OL allows for continuous adaptation to such changes and ensures the forecasting model remains up-to-date with minimal retraining [72].

2) Risk of forgetting and overfitting: OL operates in dynamic load patterns where both long-term trends and short-term anomalies coexist. Historical data often encapsulate valuable structural patterns, and real-time data streams reflect the most recent shifts. Ideally, the forecasting model should retain useful long-term knowledge while remaining responsive to meaningful short-term variations. However, achieving this balance is difficult [92]. The models that adapt too aggressively to incoming data may quickly forget

important historical patterns. Conversely, forecasting models that rely too heavily on past knowledge may fail to respond promptly to new trends. To address this issue, recent research works in other fields explore mechanisms that combine memory retention and adaptive plasticity [93], [94]. Exponential decay weighting, dual-memory structures, or meta-learning-based update strategies have shown promise in mediating this trade-off [95].

IV. Collaborative DL

In the context of ELF, it is increasingly common to encounter a collection of related tasks. While each task may differ in spatial coverage, temporal granularity, and data availability, they are often influenced by a similar set of external drivers such as weather conditions, policy shifts, seasonal demand cycles, and collective human behavior [96]. For instance, a regional cold wave can lead to synchronized spikes in both residential and commercial electricity usage. Rather than developing separate models for each setting, a growing body of research has explored collaborative DL as a way to leverage shared patterns across tasks. In particular, multi-task learning (MTL) and FL have gained traction as practical frameworks for this purpose. Although these differ in their assumptions and implementation, they all aim to improve forecasting performance by allowing forecasting models to exchange information. This section explores how collaborative DL help make ELF models more data-efficient, generalizable, and robust to the deployment in heterogeneous environments.

A. MTL

In practice, tasks may differ in trends, physical mechanisms, and temporal scales, yet they often exhibit underlying couplings. Modeling these relationships can lead to better generalization and improved forecasting accuracy. This motivates the use of MTL, a general framework where a single model is trained to solve multiple related tasks simultaneously by sharing certain components or representations across tasks. The underlying assumption is that learning multiple tasks together enables the forecasting model to extract commonalities and transfer useful knowledge [97]. In practice, MTL is often used more narrowly to refer to joint learning over a set of heterogeneous forecasting objectives such as the joint learning of regression and classification tasks [98].

It is important to distinguish MTL from multi-output forecasting. The latter focuses on predicting multiple outputs of a single task where all outputs share the same temporal context and physical meaning, e.g., forecasting the loads of different substations or regions at the same future time step. Although some research works also refer to this as MTL, in this paper it is classified under task-tuned offline learning. In contrast, MTL deals with multiple forecasting tasks that may differ in data modality or physical meaning. MTL has been primarily adopted for multi-energy ELF in IES, where electricity, heating, cooling, and gas loads are coupled but influenced by partially shared drivers [59], [99]. MTL enables the joint modeling of these interdependent yet distinct behaviors and allows the forecasting model to learn shared latent factors while preserving task-specific nuances. A literature summary of MTL is shown in Table II.

TABLE II
Literature Summary of MTL

Ref.	Knowledge sharing module	Training balance
[15]	CNN-BiGRU-attention	Homoscedastic uncertainty weighting
[17]	Transformer	Homoscedastic uncertainty weighting
[18]	Multi-scale spatiotemporal GNN	Arithmetic mean
[59]	Complex neural network	Geometric loss function
[96]	Spatiotemporal-coupling informer	Arithmetic mean
[99]	CNN-BiGRU-attention	Arithmetic mean
[100]	BiLSTM	Weighted sum
[101]	CNN	Arithmetic mean
[102]	TCN-BiGRU-attention	Arithmetic mean
[103]	TCN-GRU	Homoscedastic uncertainty and dynamic weight averaging
[104]	BiMTL LSTM	Arithmetic mean in two shared layers
[105]	BiLSTM	Arithmetic mean
[106]	LSTM	Weighted sum
[107]	Attention layer	Arithmetic mean
[108]	Kolmogorov-Arnold network	Arithmetic mean
[109]	Encoder in Transformer	Weighted sum
[110]	Temporal top windowed attention	Arithmetic mean
[111]	Task attention	Weighted sum

1) Learning Paradigm of MTL

Since MTL is primarily applied to multi-energy ELF in IES, we formulate the learning paradigm accordingly. Suppose there are K related forecasting tasks, each associated with a distinct energy type. For task $k \in 1, 2, \ldots, K$, we define the input sequence $\boldsymbol{x}_t^{(k)}$ and the corresponding target sequence $\boldsymbol{y}_t^{(k)}$. The goal of MTL is to jointly train a forecasting model that captures both the shared structure across tasks and the task-specific behaviors by minimizing the cumulative forecasting loss over all tasks:

$$\min_{\theta_s, \{\theta_k\}_{k=1}^{K}} \sum_{k=1}^{K} \lambda_k \sum_{t=1}^{T_k} \mathcal{L}^{(k)}\Big(f_{\theta_s, \theta_k}\Big(\boldsymbol{x}_t^{(k)}\Big), \boldsymbol{y}_t^{(k)} \Big) \tag{8}$$

where θ_s is the shared parameter across all tasks; θ_k is the task-specific parameter for task k; λ_k is a task-specific weighting factor that controls the contribution of each task to the overall loss; T_k is the data length or the number of samples for task k; and $\mathcal{L}^{(k)}$ is the task-specific loss function.

In most existing literature on ELF, MTL is implemented via hard parameter sharing, which enforces all tasks to share a common feature extraction backbone, while keeping separate task-specific heads for final predictions.

In existing literature on multi-energy ELF, a wide range of deep architectures have been adopted under the hard parameter sharing paradigm. These include CNNs [15], [59], [99]-[101], GRUs [102], [103], LSTMs [104]-[106], TCNs

[107], Kolmogorov-Arnold networks [108], transformers [17], [109]-[111], and GNNs [18], [96].

During the training of MTL, determining how to balance the optimization of each task is also important since naive equal weighting of task losses often leads to biased learning. Specifically, tasks with larger loss magnitudes or noisier gradients may dominate the shared representation, causing the forecasting model to over-fit to certain modalities while underperforming on others [98]. To address this, various task-balancing strategies have been explored. These include static manual tuning, uncertainty-based dynamic weighting [112], gradient-based normalization [113], and multi-objective optimization [114] strategies. These strategies have been extensively studied in other domains such as CV. In the context of multi-energy ELF, the adoption of these strategies remains limited. Most existing studies rely on static manual tuning, where task weights are set heuristically or via grid search and remain fixed throughout training. A small but growing body of work has begun to explore uncertainty-based dynamic weighting [17], which adjusts task importance based on predictive variance or model confidence. Reference [59] also employs geometric loss to dynamically balance the training of different forecasting tasks. In contrast, more advanced methods such as gradient-based normalization and multi-objective optimization have not yet been applied to MTL.

2) Discussion on MTL

The application of MTL in multi-energy ELF presents several opportunities and challenges that need to be addressed to fully realize its potential in IES.

1) Enhanced generalization through shared learning: in IES, different energy modalities such as electricity, heating, cooling, and gas often exhibit complex interdependencies [105]. MTL leverages these shared drivers to extract commonalities between these energy types while also accounting for their distinct characteristics. By simultaneously training on multiple related tasks, MTL allows the forecasting model to capture cross-modal correlations that might be overlooked in traditional single-task forecasting models [104], [106].

2) Hard versus soft parameter sharing: the majority of existing frameworks of MTL rely on hard parameter sharing, which, while effective and easy to implement, may not be optimal in all settings. Hard parameter sharing assumes a fixed structure where all tasks use the same backbone encoder, but this may struggle to capture the nuanced differences across heterogeneous energy types [115]. Soft parameter sharing, where each task maintains its own forecasting model but shares information through regularization or structured interaction, can offer a more adaptive and task-aware solution [116]. Incorporating mechanisms such as cross-task attention, task affinity modeling, or modular network design could strike a better balance between generalization and specialization in MTL-based forecasting models [117].

3) Interpretability in task sharing: interpretability remains an open challenge in understanding the effectiveness of task sharing. While shared representations are assumed to capture common factors across tasks, there is limited evidence on what is actually being transferred and how this influences forecasting performance. In some cases, shared features may contribute positively to learning; in others, especially when tasks are only weakly correlated, they may introduce noise or conflict [118]. Advancing our understanding of task relationships and shared knowledge could lead to more selective and theoretically grounded sharing strategies. Recent efforts such as task affinity measurement, representation similarity analysis, and visualization techniques can help reveal which features are beneficial to share and under what conditions [119].

4) Data privacy: data privacy is pronounced in IES settings, where different energy modalities are often managed by separate entities or subsystems. Although MTL offers a powerful mechanism to exploit inter-modal correlations for improved forecasting, its standard implementation assumes centralized access to multi-modal data, which may not be feasible in practice [120]. To address this, recent studies have begun exploring decentralized MTL that maintains task collaboration without raw data sharing [121], and federated MTL is one of the decentralized MTL.

B. Federated Learning (FL)

While MTL provides a way to improve forecasting accuracy by sharing representations across related tasks, it typically requires centralized access to all training data. This is often unrealistic in practical ELF settings due to privacy issues. In practice, we face a double-challenge: we want to make use of shared patterns across different users to improve generalization, yet we also need to respect data ownership and preserve privacy [122]. This is where FL comes in. Rather than collecting data in one place, FL allows each data owner to train a forecasting model locally. These local forecasting models then share only parameter updates with a central coordinator, which aggregates them into a global forecasting model [123]. In this way, FL makes it possible to build collaborative forecasting systems that learn from distributed experiences without exposing sensitive raw data. A literature summary of FL for ELF is shown in Supplementary Material B Table SBI [124]-[131].

1) Learning Paradigm of FL

FL encompasses a family of decentralized training paradigms that enable multiple data holders to collaboratively train ML models without sharing raw data. Depending on how the data are distributed across participants, FL can be broadly categorized into three types: horizontal FL, vertical FL, and hybrid FL [98]. Among these, horizontal FL is the most common paradigm in ELF, where each client typically has the same types of features but records data from different users. In this setting, each client trains a local forecasting model on its private dataset $\mathcal{D}_k$ [124].

After local updates, clients periodically send model parameters (or gradients) to a central server, which aggregates them (most commonly) via weighted averaging $\theta \leftarrow \sum_{m=1}^{M} \frac{n_m}{n} \theta_m$, where n_m is the number of samples on client m, and $n = \sum_m n_m$ is the total sample size. The updated θ is then broadcast back to all clients, and the process continues for multiple rounds. This iterative protocol is often referred to as federated averaging (FedAvg) and remains the foundation of most FL implementations in ELF [98]. In most current FL-based ELF studies, the local objective $\mathcal{L}$ follows the same forms used in centralized training. In essence, FL does not redefine what is optimized, but rather changes how optimization is carried out. The standard FedAvg can therefore be viewed as minimizing a global surrogate loss, which aggregates the weighted local objectives from all clients.

Vanilla FedAvg assumes a homogeneous and balanced data distribution across clients. In practice, however, consumption patterns can vary widely across user types, geographic regions, and load profiles, leading to non-independent and identically distributed (non-IID) data [88]. To overcome this limitation, various extensions of FL have been proposed and applied in the ELF. One major enhancement is to decouple the global forecasting model from client-specific adaptations, which allows the model to balance global knowledge sharing with local customization. Rather than enforcing a one-size-fits-all solution, each client m maintains a hybrid model composed of a shared global encoder and a client-specific local module.

The global module typically captures generalized representations or forecasting trends common across clients, whereas the local module adjusts the model to reflect the client's own consumption patterns. Training can be done in two ways: either jointly, where both global and local modules are updated during local training and only global parameters are shared; or in a decoupled fashion, where the global model is trained via FedAvg and the local modules are fine-tuned separately [132]. For example, [125] proposes an orthogonal decomposition framework that explicitly enforces a separation between the shared and private components in the representation space of load features. A lightweight alternative to this structural decoupling is residual personalization, where each client learns a residual correction module on top of the received global model.

Instead of altering the global training process or splitting the architecture, this method allows each client to personalize the post-aggregation of global model by learning a small residual function using local data. The methods explored in [121] and [126] offer a clean trade-off between federated efficiency and personalization capacity.

Another effective method for handling heterogeneity of load data is to form client groups based on task similarity and train separate models for each group. Rather than enforcing a single global model shared across all clients, this method maintains multiple cluster-specific models, each tailored to a subset of clients with similar data characteristics. The training process typically involves two alternating steps: ① grouping clients into clusters based on similarity in their model updates, gradients, or metadata; and ② aggregating updates within each cluster to refine a cluster-specific model. Only clients within the same cluster communicate with each other through the server, and each cluster evolves its own version of the model over time. This learning paradigm strikes a balance between full personalization and full generalization. The methods like clustered FL [133] and similarity-aware aggregation [127] have shown that such grouping can significantly improve forecasting accuracy in heterogeneous ELF scenarios.

Although FL offers stronger privacy guarantees than centralized training by keeping raw data local, it does not eliminate all risks of information leakage [134]. In particular, model updates may still expose sensitive information through indirect reconstruction attacks. To mitigate such risks, several privacy-preserving strategies have been proposed for ELF. One of the most widely adopted is secure aggregation, which ensures that the server receives only the aggregated result of multiple clients' updates without accessing any individual contribution [128]. Another strategy is gradient-level obfuscation, which reduces the information exposed by each client. For example, sign-based communication, as explored in [129] and [135], transmits only the sign of each gradient dimension. In this way, the amount of recoverable information can be drastically limited. More recently, researchers have begun exploring blockchain integration as a complementary privacy and accountability mechanism [130]. By storing model update hashes or metadata on an immutable ledger, blockchain-based systems provide verifiable traceability of client contributions and training dynamics.

2) Discussion on FL

The practical application of FL introduces several important considerations and opportunities for improvement.

1) Enhanced privacy and scalability: in a typical centralized setup, raw data from different users must be collected, which can be problematic when dealing with sensitive information like energy consumption patterns. FL offers a solution by keeping data local to individual devices or servers, sharing only model updates instead of raw data. This not only keeps individual data private but also aligns with growing privacy regulations [131]. It is a major step forward for creating systems that respect user autonomy and confidentiality without sacrificing the power of collaborative DL. Scalability is another major advantage of FL, especially for large-scale ELF tasks [136]. In a centralized method, managing and processing data from millions of users can become a logistical and computational nightmare. FL, on the other hand, allows each client to participate in the learning process without the need for overwhelming central infrastructure.

2) Client heterogeneity and forecasting consistency: one of the most pervasive issues is the non-uniformity of client data distributions. Different users exhibit vastly different consumption patterns [98]. This non-IID nature can result in unstable or biased global models that favor dominant clients or overfit to shared patterns while neglecting minority behaviors. While personalization techniques and client clustering help alleviate this, they raise a new question: how much sharing is optimal? Over-personalization risks isolate clients and lose generalizable knowledge, while insufficient personalization may fail to capture local patterns. Developing adaptive mechanisms to balance personalization and global consistency remains an ongoing research frontier [137].

3) Communication efficiency versus forecasting granularity: ELF requires high-resolution temporal predictions. However, FL setups suffer from bandwidth constraints and limit-

ed computation [138]. Striking a balance between communication frequency and model accuracy is non-trivial. Frequent communication may yield better synchronization but increases latency and energy costs. Sparse updates reduce overhead but risk model drift or delayed adaptation. Emerging techniques such as update compression, asynchronous aggregation, and event-triggered communication offer partial solutions [139], but their integration into forecasting pipelines remains underexplored.

V. General-purpose DL

A. Motivation of General-purpose DL

In practice, ELF is rarely a single, well-defined task. Different application roles place different demands on the forecasting process [140]. For example, a utility may need 15-min-ahead predictions to support real-time frequency control, whereas an energy market operator cares primarily about day-ahead accuracy for bidding. Forecasting requirements also vary across spatial and aggregation scales, since residential, feeder-level, and system-level loads support different types of operation and planning decisions. Data conditions introduce another layer of complexity. Some regions have access to high-resolution AMI measurements and detailed weather observations, whereas others rely on coarse, infrequent, or partially missing records collected by legacy metering infrastructures. As a result, both the available features and their statistical properties can differ substantially across locations and customer categories. These differences raise a simple but fundamental question: how can we design forecasting models that adapt across such heterogeneous downstream tasks and data environments without being rebuilt from scratch each time?

Unfortunately, most DL models used today are still built in a highly task-specific fashion. A model trained for one forecasting horizon, one region, or one type of input often cannot be reused directly in another scenario. Whenever the forecasting horizon changes, new covariates become available, or the data quality deteriorates, practitioners usually retrain or redesign the forecasting model. Adaptive and collaborative paradigms such as TL, OL, MTL, and FL alleviate some of these issues, but they are fundamentally designed around a fixed task set and do not support open-ended reuse of knowledge [141]. Their performance also degrades when labeled data are scarce, or when the inputs come from multiple modalities at different temporal resolutions. In other words, the growing diversity of forecasting requirements is increasingly at odds with the narrow scope of traditional DL pipelines.

These limitations have motivated increasing interest in what the computer science community refers to as general-purpose DL. In mainstream AI domains such as NLP and CV, this paradigm denotes a shift from building models for single, narrow tasks to training large networks on extensive, heterogeneous datasets so that they can learn task-agnostic and transferable representations [142]. The core idea is to let the model absorb broad structural patterns like linguistic, visual, or temporal during pre-training, and then adapt these representations to diverse downstream tasks with only lightweight updates. When this paradigm is transferred to ELF, its meaning becomes more concrete. Instead of building many separate models, general-purpose DL for ELF aims to build a single pre-trained model that captures universal load-weather-context relationships. Such a forecasting model learns common temporal structures, behavioral regularities, environmental drivers, and cross-modal correlations that consistently appear across modern power systems [143]. Once this backbone is obtained, downstream forecasting tasks can be supported through efficient adaptation mechanisms without rebuilding a new architecture. In this regard, FMs are the concrete realization of the general-purpose DL. ELFFM serves exactly this purpose by providing a unified backbone that can be efficiently adapted to diverse forecasting scenarios. A literature summary of general-purpose DL is shown in Table III. In the remainder of this section, we review how recent advances in FMs reshape the learning paradigms in ELF.

TABLE III
Summary of General-purpose DL

Ref.	Model	Architecture	Pre-training	Adaptation
[21]	Energy GPT	Decoder-only + dynamic attention	Multi-energy autoregressive	Low-rank adaptation (LoRA) + fine-tuning on per energy type
[22]	MMGPT4 LF	Dual-stream Transformer	Multi-modal contrastive	Prompt alignment + fine-tuning
[23]	ELFLLM	Transformer + LoRA	Autoregressive	LoRA + fine-tuning
[143]	Trace	LLM with chain of thought (CoT) reasoning + multi-agent consistency	Pre-trained LLM + CoT reasoning	Thought refinement + agent collaboration
[144]	Time GPT	Decoder-only Transformer	Autoregressive on diverse time series	Fine-tuning on local grids
[145]	BlackInter	Transformer + external adapters	Pretrained LLM	Adapter tuning + external prompt
[146]	BuildProg	LLM-based code generation	Language modeling	Prompt programming for code synthesis
[147]	GPT-3.5	Decoder-only LLM	Instruction tuning + code corpora	Prompt engineering for code synthesis
[148]	Chronos	Bidirectional Transformer encoder	Masked modeling on multi-domains	Zero-shot + fine-tuning
[149]	DiffLoad	Diffusion-based Transformer	Denoising diffusion modeling	Conditional sampling
[150]	Energy forecasting-LLM	Multi-channel LLM	Language modeling	Function calling + external model

B. Model Architecture

1) Transformer-based Model

Transformer architectures have become the default backbone for state-of-the-art ELFFMs because they can capture long-range temporal dependencies and fuse heterogeneous inputs within a unified end-to-end framework [151]. At the heart of the Transformer is the self-attention mechanism, which enables each time step to aggregate information from all other positions in the sequence. In practice, the forecasting model first projects the input embeddings into three latent representations corresponding to queries, keys, and values through learned linear mappings. It then computes similarity scores between queries and keys, applies a normalization operation to obtain attention weights, and uses these weights to form a weighted combination of the value vectors. Through this process, each output position can selectively focus on the most relevant temporal contexts regardless of their distance in the sequence. Compared with recurrent architectures that rely on step-by-step processing, self-attention performs these interactions in parallel, thereby improving computational efficiency and scalability for long sequences and large-scale forecasting tasks [152].

In ELFFMs, one prominent line of work follows the decoder-only autoregressive Transformer design, as shown in TimeGPT [144]. The model is pre-trained on a massive corpus of diverse time series and generates future sequences one step at a time. A similar decoder structure underlies EnergyGPT [21], which adapts GPT-2 for IES forecasting and introduces a dynamic self-attention module to adjust the temporal focus across different energy types. TRACE [143] also follows a similar autoregressive method, but it further incorporates CoT reasoning with multi-agent consistency for refining energy forecasts across distributed resources. Other models, most notably Chronos, Lag-Llama, TimesFM, and Moirai [148], [153], adopt a bidirectional encoder-style Transformer. These models are trained to embed entire time series windows using full self-attention, which makes them more suitable for downstream zero-shot or fine-tuned ELF.

2) Non-Transformer-based Model

While Transformer architectures have become the de facto backbone for most recent ELFFMs, a number of alternative paradigms, particularly those based on feedforward, convolutional, and generative mechanisms, have gained traction in the broader time series community. These non-Transformer-based models are typically designed to be lightweight, interpretable, and more inductively biased toward temporal structures [154].

Recent examples include the time series mixer (TSMixer) [155] and temporal token MLPs (TTMs) [154], which adopt channel-mixing strategies to remove the need for attention altogether. These models enable fast, memory-efficient inference without sacrificing performance. Similarly, TimesNet [156] introduces a temporal inception block that transforms time series into 2D feature maps to learn multi-periodic temporal patterns, achieving strong results across general forecasting tasks. A parallel line of exploration centers around diffusion-based FMs, which are increasingly recognized for their capacity to model complex data distributions via progressive denoising processes. Applied to time series, such models like DiffSTG [157] have demonstrated promising capabilities in probabilistic forecasting.

However, it is important to note that most of these non-Transformer-based models have not yet been systematically explored in the context of ELF. Their design intuitions appear inherently compatible with load patterns, but their effectiveness under real-world energy scenarios remains an open question.

C. Learning Paradigm of General-purpose DL

1) Pretraining

Pretraining serves as the foundation of general-purpose DL by decoupling knowledge acquisition from downstream forecasting tasks. Its core objective is to learn a task-agnostic representation function f_θ that captures generic temporal dependencies, recurring seasonal patterns, and multi-modal interactions from a large and diverse corpus $\mathcal{D}_{pretrain}$. Unlike traditional supervised learning, pre-training uses broader objectives that reflect the statistical regularities of time series without requiring labeled outputs [158].

What distinguishes this paradigm is its emphasis on universality and transferability. Instead of training separate models for every target region or forecasting horizon, pre-training builds a shared backbone that learns to encode the "grammar" of electricity consumption, which can be adapted with minimal supervision to downstream tasks [158]. Generative pre-training, contrastive pre-training, and masked pre-training are the three most popular pre-training strategies in ELFFMs.

1) Generative pre-training is designed to model the conditional distribution of a future sequence conditioned on its past observations.

$$p_\theta\left(\boldsymbol{x}_{t+1:t+H}\middle|\boldsymbol{x}_{1:t}\right)=\prod_{h=1}^{H}p_\theta\left(\boldsymbol{x}_{t+h}\middle|\boldsymbol{x}_{1:t+h-1}\right) \tag{9}$$

where $p_\theta(\cdot)$ denotes the conditional probability distribution; $\boldsymbol{x}_{1:t}$ denotes the historical context; and $\boldsymbol{x}_{t+1:t+H}$ denotes the future sequence over the forecasting horizon H.

This factorizes the joint distribution over future time steps into a product of one-step-ahead conditional distributions, where each prediction depends only on all previous values. In practice, the forecasting model is trained by minimizing the NLL of the next-step prediction, as shown in (10), which encourages the model to assign high probability to the observed trajectory given its historical context. When the conditional distribution $p_\theta\left(\boldsymbol{x}_{t+1}\middle|\boldsymbol{x}_{1:t}\right)$ is assumed to follow a Gaussian form, this loss becomes equivalent to a MSE objective. For Laplacian or quantile assumptions, it corresponds to MAE or pinball loss.

$$\mathcal{L}_{gen}=-\sum_{t=1}^{T-1}\ln p_\theta\left(\boldsymbol{x}_{t+1}\middle|\boldsymbol{x}_{1:t}\right) \tag{10}$$

where $\mathcal{L}_{gen}$ is the generative pre-training loss.

Such an autoregressive structure is suited to ELF as the model leverages historical load profiles to generate the next prediction, and then recursively uses its own outputs to inform subsequent steps. In training, the model is typically optimized to predict the next step given the past using teacher forcing. During inference, the model operates autoregressively, using its own predictions to roll out future trajectories [159]. Architecturally, this is often implemented with decoder-only Transformer backbones, which are pre-trained across large-scale multivariate time series from diverse domains. This generative perspective has recently gained traction in

ELF. The models such as TimeGPT [144], EnergyGPT [21], TSMixer [155], and TTMs [154] all adopt generative pre-training.

2) Contrastive pre-training provides a fundamentally different inductive bias compared with generative pre-training. Rather than modeling temporal generation, it focuses on learning robust and discriminative representations by comparing different time series segments. The core idea is to pull together representations of similar inputs and push apart representations of dissimilar inputs [160]. This contrastive principle encourages the model to learn invariant features that capture underlying patterns while being insensitive to noise, temporal shifts, or missing values. In practice, constructing informative and consistent views is critical. Common strategies include applying temporal cropping, masking, jittering, or even modality-wise perturbations [161]. These augmentations help the models focus on semantically meaningful structures rather than superficial noise. The models like MMGPT4LF [22] have begun exploring this direction by aligning electricity patterns with weather variables. In the broader time-series literature, TS2Vec [162] and SimMTM [163] show that contrastive representations can improve forecasting performance.

3) Masked pre-training: inspired by the success of masked language models in NLP, masked modeling for time series randomly removes parts of the input and tasks the model with reconstructing them [164]. This encourages the model to reason over contextual dependencies and learn latent structures across time.

PowerPM [165], for instance, combines masked time-step reconstruction with dual-view contrastive objectives to build robust representations that are aware of both short-term fluctuations and long-term trends. References [166] and [167] extend this idea to hierarchical masking or multi-resolution contexts, encouraging the model to infer both fine-grained and coarse-grained temporal relationships.

2) Adaptation

Following pre-training, adaptation plays a pivotal role in aligning general-purpose FMs with the specific requirements of ELF. Unlike conventional DL methods that train models from scratch for each use case, ELFFMs adopt more efficient mechanisms to transfer pre-trained knowledge to downstream tasks with minimal data or computation. Broadly, adaptation strategies can be categorized into four classes: full fine-tuning, parameter-efficient tuning, prompt-based adaptation, and black-box or adapter-based adaptation. A straightforward adaptation strategy is to fine-tune the entire model on the target dataset $\mathcal{D}_{target}$.

While effective in tailoring the model to specific spatial or temporal regimes [168], [169], full fine-tuning can be computationally intensive and prone to overfitting. To address the limitation of full fine-tuning, many ELFFMs adopt parameter-efficient tuning strategies such as LoRA [170]. It injects low-rank trainable matrices into attention or feedforward layers and keeps the original weights frozen. For example, LFLLM [23] and EnergyGPT [21] leverage LoRA to adapt large Transformer backbones with only a few thousand parameters per task.

A more lightweight and modular method is prompting, where the FM remains frozen, and the adaptation is achieved by injecting external conditioning signals.

Prompts may be handcrafted or learned from data. In the ELFFM, prompt-based design is often used for task routing, temporal context conditioning, or domain adaptation. For instance, BuildProg [146] uses prompts to guide the generation of building-specific forecasting pipelines, while TEMPO [171] maintains a pool of continuous prompts and uses similarity-based retrieval to select the most relevant one for each forecast instance.

In scenarios where FMs are deployed as closed-source application programming interfaces or when internal parameters are inaccessible, black-box adaptation becomes critical. One effective solution is to train external adapter modules that steer the behavior of frozen model. For example, BlackInter [145] introduces trainable adapters that modulate input embeddings or output logits, enabling transfer to unseen buildings or tasks with no direct parameter access.

D. Benefits of FMs

1) Scalability

FMs bring a new level of scalability to ELF. Instead of training a new model for each task, ELFFMs learn once and adapt to different scenarios through lightweight fine-tuning or prompting. Recent empirical evidence illustrates this scalability. TimeGPT, for example, employs a large pre-trained backbone and applies minimal task-specific tuning to campus-level loads at the University of Texas at Austin, USA. Even under scarce historical data, it delivers more than a 10% improvement in RMSE compared with a range of widely used ML and DL baselines [144]. EnergyGPT adopts the same philosophy but is specialized for multi-energy systems. By attaching lightweight adapters to a GPT-2 backbone, the model can be efficiently fine-tuned for electricity, cooling, heating, and combined load forecasting on the Tempe campus of Arizona State University, USA. EnergyGPT achieves MAPE values of 1.85%, 3.21%, 2.60%, and 1.95%, respectively, outperforming all DL baselines across four seasons (spring, summer, autumn, and winter) [21]. At the building level, BlackInter further demonstrates that FM-based representations remain effective even when little or no target data are available. By injecting building-specific information through a compact adapter on top of a pre-trained language model, it reduces prediction errors to 42.88% under few-shot and 69.44% under zero-shot of the best alternative method [145]. These results show that once an FM has been pre-trained, it can be rapidly adapted to heterogeneous assets, horizons, and data regimes [172].

2) Flexibility Under Data Scarcity

FMs offer strong flexibility when labeled data are scarce or partially missing, which is common in newly-built stations or emerging regions. By learning generalizable temporal patterns and contextual dependencies from large-scale heterogeneous datasets, ELFFMs provide powerful inductive priors that can be quickly adapted to new forecasting tasks with limited samples [173].

3) Unified Multimodal Representation

FMs enable seamless integration of heterogeneous information sources that characterize modern power systems. ELF inherently involves diverse and asynchronous signals such as historical consumption, NWPs, calendar effects, and sensor telemetry. Traditional models often rely on ad hoc feature engineering or independent preprocessing pipelines, which limits their scalability and forecasting accuracy. By contrast, FM provides a unified framework to fuse these multi-modal inputs in a single architecture. Through shared representation learning and cross-modal attention, it captures complex interactions across temporal and spatial dimensions, leading to more accurate and stable forecasts.

E. Challenges of FMs in ELF

1) Computational Cost

The ambition of general-purpose models comes with substantial computational demands, especially during pre-training. Modern FM-style architectures for time-series forecasting often contain hundreds of millions to over one billion parameters, and training them requires processing multi-year, multi-region load and weather data with long input sequences [174]. This places heavy pressure not only on graphics processing unit (GPU) computation but also on memory bandwidth, data loading pipelines, and distributed training infrastructure. Transformer-based models, in particular, face quadratic scaling with sequence length, which severely limits its batch size and increases the need for expensive memory-optimization techniques such as model parallelism, gradient checkpointing, or flash attention [152]. As a result, pre-training typically requires access to high-end GPU clusters (e.g., 8 to 32 NVIDIA A100 or H100 GPUs) running continuously for several days and in some cases several weeks [175]. Even once trained, large FMs impose nontrivial inference costs when deployed for rolling forecasting or across many spatial nodes. These computational requirements far exceed what most utilities, system operators, or academic research groups can sustain. The mismatch between the scale needed for pre-training and the resources available in typical power system environments remains a major barrier to the widespread adoption of FMs in ELF.

2) Data Curation and Governance

Beyond computation, the more persistent challenge lies in data curation and governance. Pretraining an FM requires not only large volumes of data but also well-organized, high-quality datasets. In practice, raw load data contain missing values, sensor noises, abrupt discontinuities from meter replacements, and inconsistencies across AMI systems. Multi-source inputs such as weather, market variables, and event records often follow different temporal resolutions and metadata formats, making temporal alignment and schema harmonization a significant engineering burden [176].

Data governance further complicates the process. Utilities and regional operators typically store data under different standards for timestamping, aggregation, anomaly handling, and privacy protection. The access to granular load profiles may be restricted due to confidentiality concerns, while cross-utility data sharing is often limited by regulatory boundaries. Ensuring consistent preprocessing, feature definitions, metadata documentation, and auditability across such heterogeneous sources can require more effort than model training itself [177]. Large organizations may amortize these costs through centralized data platforms, but smaller utilities and research groups face an inherent imbalance: they bear the inference cost of FM-scale models without having the capacity to construct or maintain the high-quality datasets required for pre-training. Until standardized data pipelines and shared governance protocols mature within the power sector, data curation will remain a major barrier to the broader adoption of domain-specific FMs in ELF.

F. Trade-off Between Fine-tuning LLMs and Building Domain-specific FMs

Given the ELF requirements and data conditions discussed above, practitioners face two broad pathways when adopting general-purpose DL: fine-tuning the existing LLM or developing a domain-specific FM.

Fine-tuning a pre-trained LLM is attractive for obvious reasons. It is fast, efficient, and remains effective even when data or computing resources are limited. With minimal effort, one can obtain a reasonably good model that also handles mixed inputs such as weather descriptions, event notes, or other textual context. This makes fine-tuning useful when quick deployment matters more than squeezing out every bit of accuracy. Namely, the limitations become apparent once we look at the nature of load data. LLMs are built to understand language, not to model the temporal dynamics that drive electricity demand. The tokenization and training objectives of LLM are optimized for semantics, not for representing continuous numerical signals. As a result, they struggle with long-horizon forecasting, periodic fluctuations, and uncertainty calibration. Reference [178] reports that removing the language component from LLM-based time series models often has little impact on performance. This implies that the linguistic structure itself adds limited value to numerical prediction.

A domain-specific FM is developed using the opposite method. It requires more time, data, and computational cost upfront, but the payoff is a model that understands the behavioral patterns underlying electricity use. By pre-training on large-scale load datasets, the model can internalize patterns that reflect how people actually consume energy. The main challenge, of course, is that pre-training such a model demands diverse, high-quality datasets and a nontrivial engineering investment, which may not always be available in every region or organization.

In practice, the choice between fine-tuning LLMs and building domain-specific FMs depends on operational priorities. LLMs provide flexibility and rapid adaptation, while domain-specific FMs offer higher fidelity for numerical forecasting [141], [172], [179]. System operators can select the method that best matches their data conditions, accuracy requirements, and resource constraints.

VI. Challenges and Opportunities

Although various learning paradigms have been studied to tackle different challenges in load forecasting, several critical challenges still remain. This section identifies key obstacles and outlines promising research opportunities.

A. Modeling Impact of Human Behavior

One of the most difficult aspects of ELF is capturing the nuances of human power consumption behavior, which is influenced by a wide range of factors like technological advancements, societal shifts, and even personal habits. Taking electric vehicles (EVs) as an example, as more people adopt EVs, EV charging occurs at diverse time and in patterns that were difficult to anticipate just a few years ago. The problem here is not just about accounting for new variables, but about understanding how individuals and communities adapt to a changing world. People's choices around energy consumption are often a reaction to broader societal forces, like rising electricity prices, environmental concerns, or government incentives to reduce energy consumption. But these reactions are highly personal and can vary widely depending on the lifestyle, location, and even personal values. Predicting these shifts requires moving beyond historical consumption trends and actively incorporating behavioral insights into forecasting models.

Real-time data from smart meters, connected home devices, and even wearable technologies give us a chance to see how people's behaviors are shifting. For instance, an individual decision to turn down the thermostat in response to higher energy prices can now be tracked and used to refine forecasts. This is where models that can dynamically adjust to behavior come into play. At the heart of this is the idea that ELF does not just have to be about numbers, but about understanding individual choices and how those choices interact with the grid. The more we can integrate human factors into forecasting, the better we can manage energy systems in ways that are both efficient and flexible.

B. Modeling Impact of Extreme Weather

The challenge of extreme weather events in ELF stems from both the unpredictability of these events and the lack of sufficient historical data to train forecasting models. While standard weather patterns can typically be predicted with reasonable accuracy, extreme events such as heatwaves, storms, or cold fronts often deviate significantly from these patterns, resulting in sudden surges or drops in demand that traditional forecasting models struggle to capture. The rarity of these extreme events means that historical data on their impact are limited, making it difficult for forecasting models to learn from past occurrences. Moreover, the impact of extreme weather on electricity consumption is not just driven by the environment but is also influenced by human behavior. These factors add a layer of complexity that makes forecasting under such conditions particularly difficult.

To address this challenge, one promising method is to generate synthetic scenarios of extreme weather events and the corresponding human responses using data augmentation and scenario generation techniques. For example, [180] uses classifier-guided diffusion models combined with extreme value theory to generate realistic and varied extreme load scenarios. Furthermore, integrating adaptive DL such as RL and OL offers a dynamic solution. These techniques allow the models to continuously update their predictions based on real-time data, reflecting actual behavioral changes in response to extreme weather.

C. ELF in Downstream Tasks

The ultimate goal of ELF is not merely to predict future energy demand, but to enable more effective decision-making in the energy systems. However, many current forecasting models focus solely on improving prediction accuracy without considering how these forecasts can be translated into actionable insights for decision-making. This disconnection leads to a critical challenge: while these forecasting models perform well in terms of error metrics, they may not be good enough for operational decisions.

A promising solution is to develop task cost-based forecasting models, where forecasts are optimized not only for accuracy but also for minimizing specific operational costs. The optimizer in such forecasting models may be differentiable, thus allowing for end-to-end training. Several studies such as [181] and [182] have explored this direction. However, the integration of task cost-based objectives with forecasting models requires careful consideration of the trade-offs between accuracy and operational goals. In many cases, optimizing for one objective may degrade performance on another. Moreover, the complexity of incorporating real-time operational constraints and dynamic grid behaviors into the learning process remains a challenge.

D. Is DL Truly Effective in Every Forecasting Scenario?

Researchers are actively exploring DL, practitioners are implementing it, and it is often seen as the go-to solution for forecasting problems. But here is a question that has not been asked enough: "is DL truly the right tool for every forecasting scenario?" It is easy to get caught up in the hype and assume that because DL has shown remarkable results in some areas, it should automatically work everywhere. However, this mindset can lead to blind application of a technology without truly assessing whether it aligns with the specific needs of the task at hand.

In the real world, the application of DL in ELF is far more nuanced than simply applying the most sophisticated model available. DL thrives in situations where there is abundant high-quality data and complex relationships to capture. But what happens when the data are sparse, noisy, or fundamentally limited? What if the problem at hand involves rare events that are not well-represented in historical data? In these cases, the predictive power of DL is not guaranteed, and its data-hungry nature becomes a liability. Moreover, DL models can sometimes be overly complex for simpler forecasting tasks where traditional models might do just as well. For example, in environments with well-established patterns or in tasks that do not require nuanced pattern recognition, simpler methods might be more efficient, interpretable, and computationally feasible. Why overcomplicate things with DL if the task does not demand it?

The key takeaway here is that DL should not be used blindly. It is not a one-size-fits-all solution. When deciding whether to apply DL to a forecasting task, we need to take a step back and consider the specific nature of the task. Does the problem involve a complex, non-linear relationship that is difficult for traditional models to capture? Do we have enough data to support training a DL model? Is interpretability important for the stakeholders? These are the types of questions that need to be asked before automatically defaulting to DL.

VII. Conclusion

Accurate ELF has been a cornerstone of power system management. Over the last decade, DL has revolutionized this task, enabling forecasting models that can adapt to complex data patterns. In this survey, we examine the progression of ELF through the lens of learning paradigms. Four representative learning paradigms are reviewed and discussed, including task-tuned offline learning, adaptive DL, collaborative DL, and general-purpose DL. These learning paradigms reflect the increasing complexity and diversity of power systems, with each one offering improvements but also introducing new challenges. Looking ahead, the key challenge for the field is clear: forecasting models must go beyond simply improving accuracy. The real goal is to make forecasts more actionable and aligned with operational decisions.

References

[1] International Renewable Energy Agency. (2025, Mar.). Record-breaking annual growth in renewable power capacity. [Online]. Available: https://www. irena. org/News/pressreleases/2025/Mar/Record-Breaking-Annual-Growth-in-Renewable-Power-Capacity

[2] International Energy Agency. (2025, Jan.). Global energy review 2025: electricity analysis. [Online]. Available: https://www. iea. org/reports/global-energy-review-2025/electricity

[3] Y. Eren, İ. Küçükdemiral, İ. Üstoğlu *et al.*, "A comprehensive review on deep learning approaches for short-term load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 189, p. 114031, Jan. 2024.

[4] B. F. Hobbs, S. Jitprapaikulsarn, S. Konda *et al.*, "Analysis of the value for unit commitment of improved load forecasts," *IEEE Transactions on Power Systems*, vol. 14, no. 4, pp. 1342-1348, Nov. 1999.

[5] S. Aslam, H. Herodotou, S. M. Mohsin *et al.*, "A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids," *Renewable and Sustainable Energy Reviews*, vol. 144, p. 110992, Jul. 2021.

[6] X. Liu, Z. Zhang, and Z. Song, "A comparative study of the data-driven day-ahead hourly provincial load forecasting methods: from classical data mining to deep learning," *Renewable and Sustainable Energy Reviews*, vol. 119, p. 109632, Mar. 2020.

[7] V. Golovko, "Deep learning: an overview and main paradigms," *Optical Memory and Neural Networks*, vol. 26, no. 1, pp. 1-17, Jan. 2017.

[8] G. Tziolis, C. Spanias, M. Theodoride *et al.*, "Short-term electric net load forecasting for solar-integrated distribution systems based on Bayesian neural networks and statistical post-processing," *Energy*, vol. 271, p. 127018, May 2023.

[9] S. Karamolegkos and D. E. Koulouriotis, "Advancing short-term load forecasting with decomposed Fourier ARIMA: a case study on the greek energy market," *Energy*, vol. 325, p. 135854, Jun. 2025.

[10] M. I. El-Afifi, A. A. Eladl, B. E. Sedhom *et al.*, "Enhancing EV charging station integration: a hybrid ARIMA-LSTM forecasting and optimization framework," *IEEE Transactions on Industry Applications*, vol. 61, no. 3, pp. 4924-4935, May 2025.

[11] Z. Zhao, Y. Zhang, Y. Yang *et al.*, "Load forecasting via grey model-least squares support vector machine model and spatial-temporal distribution of electric consumption intensity," *Energy*, vol. 255, p. 124468, Sept. 2022.

[12] L. Zhang and D. Jánošík, "Enhanced short-term load forecasting with hybrid machine learning models: CATBoost and XGBoost approaches," *Expert Systems with Applications*, vol. 241, p. 122686, May 2024.

[13] O. Rubasinghe, X. Zhang, T. K. Chau *et al.*, "A novel sequence to sequence data modelling based CNN-LSTM algorithm for three years ahead monthly peak load forecasting," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 1932-1947, Jan. 2024.

[14] E. Yang and C.-H. Youn, "Temporal data pooling with meta-initialization for individual short-term load forecasting," *IEEE Transactions on Smart Grid*, vol. 14, no. 4, pp. 3246-3258, Jul. 2023.

[15] D. Niu, M. Yu, L. Sun *et al.*, "Short-term multi-energy load forecasting for integrated energy systems based on CNN-BIGRU optimized by attention mechanism," *Applied Energy*, vol. 313, p. 118801, May 2022.

[16] Z. Xu, Z. Yu, H. Zhang *et al.*, "Phacia-TCNs: short-term load forecasting using temporal convolutional networks with parallel hybrid activated convolution and input attention," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 1, pp. 427-438, Jan. 2024.

[17] C. Wang, Y. Wang, Z. Ding *et al.*, "Probabilistic multi-energy load forecasting for integrated energy system based on Bayesian transformer network," *IEEE Transactions on Smart Grid*, vol. 15, no. 2, pp. 1495-1508, Mar. 2024.

[18] W. Zhuang, J. Fan, M. Xia *et al.*, "A multi-scale spatialtemporal graph neural network-based method of multienergy load forecasting in integrated energy system," *IEEE Transactions on Smart Grid*, vol. 15, no. 3, pp. 2652-2666, May 2024.

[19] P. Zhao, Z. Shen, D. Cao *et al.*, "Missing data-aware robust electrical load forecasting based on hierarchical downsampling-upsampling spatiotemporal graph network," *Applied Energy*, vol. 405, p. 127209, Jan. 2026.

[20] W. Zhao, K. Zhou, J. Li *et al.* (2023, Mar.). A survey of large language models. [Online]. Available: https://arxiv.org/abs/2303.18223

[21] M. Liang, Y. Hu, H. Weng *et al.*, "EnergyGPT: fine-tuning large language model for multi-energy load forecasting," *Renewable Energy*, vol. 251, p. 123313, May 2025.

[22] M. Gao, S. Zhou, W. Gu *et al.*, "MMGPT4LF: leveraging an optimized pre-trained GPT-2 model with multi-modal cross-attention for load forecasting," *Applied Energy*, vol. 392, p. 125965, Jun. 2025.

[23] G. Liu, Y. Bai, K. Wen *et al.* (2024, May). LFLLM: a large language model for load forecasting. [Online]. Available: https://doi.org/10.36227/techrxiv.170475236.64005369/v1.

[24] C. Kuster, Y. Rezgui, and M. A. Mourshed, "Electrical load forecasting models: a critical systematic review," *Sustainable Cities and Society*, vol. 35, pp. 257-270, Nov. 2017.

[25] M. Q. Raza and A. Khosravi, "A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings," *Renewable and Sustainable Energy Reviews*, vol. 50, pp. 1352-1372, Oct. 2015.

[26] I. K. Nti, M. Teimeh, O. Nyarko-Boateng *et al.*, "Electricity load forecasting: a systematic review," *Journal of Electrical Systems and Information Technology*, vol. 7, pp. 1-19, Dec. 2020.

[27] M. A. Hammad, B. Jereb, B. Rosi *et al.*, "Methods and models for electric load forecasting: a comprehensive review," *Logistics, Supply Chain, Sustainability and Global Challenges*, vol. 11, no. 1, pp. 51-76, Dec. 2020.

[28] A. Al Mamun, M. Sohel, N. Mohammad *et al.*, "A comprehensive review of the load forecasting techniques using single and hybrid predictive models," *IEEE Access*, vol. 8, pp. 134911-134939, Jul. 2020.

[29] T. Hong and S. Fan, "Probabilistic electric load forecasting: a tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914-938, Jul. 2016.

[30] J. Zhu, H. Dong, W. Zheng *et al.*, "Review and prospect of data-driven techniques for load forecasting in integrated energy systems," *Applied Energy*, vol. 321, p. 119269, Sept. 2022.

[31] V. Kondaiah, B. Saravanan, P. Sanjeevikumar *et al.*, "A review on short-term load forecasting models for microgrid application," *The Journal of Engineering*, vol. 2022, no. 7, pp. 665-689, Jul. 2022.

[32] R. Wazirali, E. Yaghoubi, M. S. S. Abujazar *et al.*, "State-of-the-art review on energy and load forecasting in microgrids using artificial neural networks, machine learning, and deep learning techniques," *Electric Power Systems Research*, vol. 225, p. 109792, Dec. 2023.

[33] B. Yildiz, J. I. Bilbao, and A. B. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 1104-1122, Jun. 2017.

[34] X. Wang, H. Wang, B. Bhandari *et al.*, "Ai-empowered methods for smart energy consumption: a review of load forecasting, anomaly detection and demand response," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 11, no. 3, pp. 963-993, May 2024.

[35] M. Hasan, Z. Mifta, S. J. Papiya *et al.*, "A state-of-the-art comparative

review of load forecasting methods: characteristics, perspectives, and applications," *Energy Conversion and Management: X*, vol. 25, p. 100922, Jan. 2025.

[36] L. Baur, K. Ditschuneit, M. Schambach *et al.*, "Explainability and interpretability in electric load forecasting using machine learning techniques – a review," *Energy and AI*, vol. 18, p. 100358, Dec. 2024.

[37] S. Haben, S. Arora, G. Giasemidis *et al.*, "Review of low voltage load forecasting: methods, applications, and recommendations," *Applied Energy*, vol. 304, p. 117798, Dec. 2021.

[38] A. R. Khan, A. Mahmood, A. Safdar *et al.*, "Load forecasting, dynamic pricing and DSM in smart grid: a review," *Renewable and Sustainable Energy Reviews*, vol. 54, pp. 1311-1322, Feb. 2016.

[39] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: a review and evaluation," *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 44-55, Feb. 2001.

[40] F. M. Bianchi, E. Maiorino, M. C. Kampffmeyer *et al.* (2017, May). An overview and comparative analysis of recurrent neural networks for short term load forecasting. [Online]. Available: https://arxiv.org/abs/1705.04378

[41] W. Guo, F. Zhuang, X. Zhang *et al.*, "A comprehensive survey of federated transfer learning: challenges, methods and applications," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186356, Dec. 2024.

[42] J. Gong, Z. Qu, Z. Zhu *et al.*, "Ensemble models of TCN-LSTM-lightGBM based on ensemble learning methods for short-term electrical load forecasting," *Energy*, vol. 318, p. 134757, Mar. 2025.

[43] J. Xing, J. Su, Y. Xue *et al.*, "A unifying framework for short-term load forecasting via EM-CCGPBOATT methodology," *IEEE Transactions on Power Systems*, vol. 40, no. 2, pp. 1414-1426, Mar. 2025.

[44] A. Jadon, A. Patil, and S. Jadon, "A comprehensive survey of regression-based loss functions for time series forecasting," in *Proceedings of International Conference on Data Management, Analytics & Innovation*, Pune, India, Jan. 2024, pp. 117-147.

[45] H. Zhang, R. Li, Q. Du *et al.*, "Decision-focused learning for power system decision-making under uncertainty," *IEEE Transactions on Power Systems*, vol. 40, no. 1, pp. 1-18, Jan. 2025.

[46] D. P. Kingma and J. Ba. (2014, Dec.). Adam: a method for stochastic optimization. [Online]. Available: https://arxiv.org/abs/1412.6980

[47] V. Moudgil, R. Sadiq, J. Brar *et al.*, "Dual-channel encoded bidirectional LSTM for multi-building short-term load forecasting," *Journal of Cleaner Production*, vol. 486, p. 144555, Jan. 2025.

[48] X. Dong, Y. Luo, S. Yuan *et al.*, "Building electricity load forecasting based on spatiotemporal correlation and electricity consumption behavior information," *Applied Energy*, vol. 377, p. 124580, Jan. 2025.

[49] Z. Chen, C. Du, B. Zhang *et al.*, "A cybersecure distribution-free learning model for interval forecasting of power load under cyberattacks," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 3, pp. 2540-2549, Mar. 2025.

[50] Y. Wang, J. Chen, X. Chen *et al.*, "Short-term load forecasting for industrial customers based on TCN-lightGBM," *IEEE Transactions on Power Systems*, vol. 36, no. 3, pp. 1984-1997, May 2021.

[51] F. Cheng and H. Liu, "Multi-step electric vehicles charging loads forecasting: an autoformer variant with feature extraction, frequency enhancement, and error correction blocks," *Applied Energy*, vol. 376, p. 124308, Dec. 2024.

[52] O. Bouhamed, M. Dissem, M. Amayri *et al.*, "Transformer-based deep probabilistic network for load forecasting," *Engineering Applications of Artificial Intelligence*, vol. 152, p. 110781, Jun. 2025.

[53] Z. Fazlipour, E. Mashhour, and M. Joorabian, "A deep model for short-term load forecasting applying a stacked autoencoder based on LSTM supported by a multi-stage attention mechanism," *Applied Energy*, vol. 327, p. 120063, Dec. 2022.

[54] Z. Tian, W. Liu, W. Jiang *et al.*, "CNNs-transformer based day-ahead probabilistic load forecasting for weekends with limited data availability," *Energy*, vol. 293, p. 130666, Apr. 2024.

[55] Y. Lv, L. Wang, D. Long *et al.*, "Multi-area short-term load forecasting based on spatiotemporal graph neural network," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109398, Dec. 2024.

[56] J. Xiao, X. Cui, X. Liu *et al.*, "Improved 3-D LSTM: a video prediction approach to long sequence load forecasting," *IEEE Transactions on Smart Grid*, vol. 16, no. 2, pp. 1885-1896, Mar. 2025.

[57] G. He, D. Jin, L. Dai *et al.*, "Online learning of temporal association rule on dynamic multivariate time series data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 1-14, Dec. 2024.

[58] T. Liu, S. Chen, S. Liang *et al.*, "Fast adaptive gradient RBF networks for online learning of nonstationary time series," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2015-2030, Mar. 2020.

[59] P. Zhao, D. Cao, W. Hu *et al.*, "Geometric loss-enabled complex neural network for multi-energy load forecasting in integrated energy systems," *IEEE Transactions on Power Systems*, vol. 39, no. 4, pp. 5659-5671, Jul. 2024.

[60] T. U. Badrudeen, N. I. Nwulu, and S. L. Gbadamosi, "Low-inertia control of a large-scale renewable energy penetration in power grids: a systematic review with taxonomy and bibliometric analysis," *Energy Strategy Reviews*, vol. 52, p. 101337, Mar. 2024.

[61] X. Wang, H. Wang, S. Li *et al.*, "A reinforcement learning-based online learning strategy for real-time short-term load forecasting," *Energy*, vol. 305, p. 132344, Sept. 2024.

[62] Z. Zhang, P. Zhao, P. Wang *et al.*, "Transfer learning featured short-term combining forecasting model for residential loads with small sample sets," *IEEE Transactions on Industry Applications*, vol. 58, no. 4, pp. 4279-4288, Jul. 2022.

[63] L. Xiao, Q. Bai, and B. Wang, "A dynamic multi-model transfer based short-term load forecasting," *Applied Soft Computing*, vol. 159, p. 111627, Jul. 2024.

[64] C. Peng, Y. Tao, Z. Chen *et al.*, "Multi-source transfer learning guided ensemble LSTM for building multi-load forecasting," *Expert Systems with Applications*, vol. 202, p. 117194, Sept. 2022.

[65] H. Dong, J. Zhu, S. Li *et al.*, "Probabilistic residential load forecasting with sequence-to-sequence adversarial domain adaptation networks," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 5, pp. 1559-1571, Sept. 2024.

[66] J. Zhu, Y. Miao, H. Dong *et al.*, "Short-term residential load forecasting based on *k*-shape clustering and domain adversarial transfer network," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 4, pp. 1239-1249, Jul. 2024.

[67] Z. Zhou, Y. Xu, and C. Ren, "A transfer learning method for forecasting masked-load with behind-the-meter distributed energy resources," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4961-4964, Nov. 2022.

[68] Y. He, F. Luo, and G. Ranzi, "Transferrable model-agnostic meta-learning for short-term household load forecasting with limited training data," *IEEE Transactions on Power Systems*, vol. 37, no. 4, pp. 3177-3180, Jul. 2022.

[69] P. Zhao, D. Cao, Y. Wang *et al.*, "Gaussian process-aided transfer learning for probabilistic load forecasting against anomalous events," *IEEE Transactions on Power Systems*, vol. 38, no. 3, pp. 2962-2965, May 2023.

[70] D. Wu and W. Lin, "Efficient residential electric load forecasting via transfer learning and graph neural networks," *IEEE Transactions on Smart Grid*, vol. 14, no. 3, pp. 2423-2431, May 2023.

[71] C. Cao and Y. He, "An online probability density load forecasting against concept drift under anomalous events," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 5241-5252, Apr. 2024.

[72] Y. Fujimoto, M. Fujita, and Y. Hayashi, "Deep reservoir architecture for short-term residential load forecasting: an online learning scheme for edge computing," *Applied Energy*, vol. 298, p. 117176, Sept. 2021.

[73] P. Zhao, W. Hu, D. Cao *et al.*, "Online probabilistic load forecasts considering the data gaps," *CSEE Journal of Power and Energy Systems*, vol. 11, no. 2, pp. 1-12, Mar. 2025.

[74] Y. Li, F. Zhang, Y. Liu *et al.*, "Residential load forecasting: an online-offline deep kernel learning method," *IEEE Transactions on Power Systems*, vol. 39, no. 2, pp. 4264-4278, Mar. 2024.

[75] L. Von Krannichfeldt, Y. Wang, and G. Hug, "Online ensemble learning for load forecasting," *IEEE Transactions on Power Systems*, vol. 36, no. 1, pp. 545-548, Jan. 2021.

[76] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[77] J. Kirkpatrick, R. Pascanu, N. Rabinowitz *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521-3526, Mar. 2017.

[78] M. Long, Y. Cao, J. Wang *et al.*, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, Jul. 2015, pp. 97-105.

[79] X. Zhang, J. Chen, Y. Wang *et al.*, "Cross-building transfer learning for deep building load forecasting," *Applied Energy*, vol. 325, p. 119837, Nov. 2022.

[80] C. Yu, J. Wang, Y. Chen *et al.*, "Transfer learning with dynamic adversarial adaptation network," in *Proceedings of 2019 IEEE International Conference on Data Mining*, Beijing, China, Nov. 2019, pp. 778-786.

[81] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of International Conference on Machine Learning*, Sydney, Australia, Aug. 2017, pp.

1126-1135.

[82] Y. Li, S. Zhang, R. Hu *et al.*, "A meta-learning based distribution system load forecasting model selection framework," *Applied Energy*, vol. 294, p. 116991, Jul. 2021.

[83] X. Xiao, H. Mo, Y. Zhang *et al.*, "Meta-ANN – a dynamic artificial neural network refined by meta-learning for short-term load forecasting," *Energy*, vol. 246, p. 123418, May 2022.

[84] J. Chen, C. Zhang, X. Li *et al.*, "An integrative approach to enhance load forecasting accuracy in power systems based on multivariate feature selection and selective stacking ensemble modeling," *Energy*, vol. 326, p. 136337, Jul. 2025.

[85] P. Zhao, W. Hu, D. Cao *et al.*, "Causal mechanism-enabled zero-label learning for power generation forecasting of newly-built PV sites," *IEEE Transactions on Sustainable Energy*, vol. 16, no. 1, pp. 392-406, Jan. 2025.

[86] R. Chen, Y. Liu, L. Kong *et al.*, "Towards label-free scene understanding by vision foundation models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 75896-75910, Dec. 2023.

[87] C. Song and A. Raghunathan, "Information leakage in embedding models," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, Virtual Event, Nov. 2020, pp. 377-390.

[88] Y. He, F. Luo, M. Sun *et al.*, "Privacy-preserving and hierarchically federated framework for short-term residential load forecasting," *IEEE Transactions on Smart Grid*, vol. 14, no. 6, pp. 4409-4423, Nov. 2023.

[89] V. Álvarez, S. Mazuelas, and J. A. Lozano, "Probabilistic load forecasting based on adaptive online learning," *IEEE Transactions on Power Systems*, vol. 36, no. 4, pp. 3668-3680, Jul. 2021.

[90] A. Bifet and R. Gavalda, "Adaptive learning from evolving data streams," in *Proceedings of International Symposium on Intelligent Data Analysis*, Lyon, France, Aug. 2009, pp. 249-260.

[91] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107-194, Apr. 2012.

[92] K. Malialis, C. G. Panayiotou, and M. M. Polycarpou, "Online learning with adaptive rebalancing in nonstationary environments," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4445-4459, Oct. 2021.

[93] Q. Pham, C. Liu, D. Sahoo *et al.*, "Learning fast and slow for online time series forecasting," in *Proceedings of 7th International Conference on Learning Representations*, Kigali, Rwanda, May 2023, pp. 1-12.

[94] Y. Zeng, P. Zhao, D. Cao *et al.*, "Online spatiotemporal ensemble learning for load forecasting against anomalous events," *IEEE Transactions on Power Systems*, vol. 40, no. 2, pp. 1-4, Mar. 2025.

[95] S. Dey, S. Ali, and C. Park, "Weighted exponential distribution: properties and different methods of estimation," *Journal of Statistical Computation and Simulation*, vol. 85, no. 18, pp. 3641-3661, Dec. 2015.

[96] F. Liu, Y. Huang, Y. Wang *et al.*, "A two-stage day-ahead multi-energy consumption forecasting method based on multi-task spatiotemporal coupling informer and enhanced spider wasp optimization," *IEEE Transactions on Smart Grid*, vol. 16, no. 3, pp. 1-14, May 2025.

[97] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30-43, Jan. 2018.

[98] Y. Zhang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586-5609, Dec. 2022.

[99] D. Peng, Y. Liu, D. Wang *et al.*, "Multi-energy load forecasting for integrated energy system based on sequence decomposition fusion and factors correlation analysis," *Energy*, vol. 308, p. 132796, Nov. 2024.

[100] K. Li, Y. Mu, F. Yang *et al.*, "A novel short-term multi-energy load forecasting method for integrated energy system based on feature separation-fusion technology and improved CNN," *Applied Energy*, vol. 351, p. 121823, Dec. 2023.

[101] S. Zhu, H. Ma, L. Chen *et al.*, "Short-term load forecasting of an integrated energy system based on STL-CPLE with multitask learning," *Protection and Control of Modern Power Systems*, vol. 9, no. 6, pp. 71-92, Nov. 2024.

[102] H. Chen, H. Huang, Y. Zheng *et al.*, "A load forecasting approach for integrated energy systems based on aggregation hybrid modal decomposition and combined model," *Applied Energy*, vol. 375, p. 124166, Dec. 2024.

[103] N. Huang, S. Ren, J. Liu *et al.*, "Multi-task learning and single-task learning joint multi-energy load forecasting of integrated energy systems considering meteorological variations," *Expert Systems with Applications*, vol. 267, p. 128269, Mar. 2025.

[104] C. Liao, M. Tan, K. Li *et al.*, "Sequence signal prediction and reconstruction for multi-energy load forecasting in integrated energy systems: a bi-level multi-task learning method," *Energy*, vol. 313, p. 133960, Dec. 2024.

[105] J. Shi, J. Teh, B. Alharbi *et al.*, "Load forecasting for regional integrated energy system based on two-phase decomposition and mixture prediction model," *Energy*, vol. 297, p. 131236, Jun. 2024.

[106] K. Wu, J. Gu, L. Meng *et al.*, "An explainable framework for load forecasting of a regional integrated energy system based on coupled features and multi-task learning," *Protection and Control of Modern Power Systems*, vol. 7, no. 2, pp. 1-14, Apr. 2022.

[107] H. Guo, B. Huang, and J. Wang, "Probabilistic load forecasting for integrated energy systems using attentive quantile regression temporal convolutional network," *Advances in Applied Energy*, vol. 14, p. 100165, Jun. 2024.

[108] X. Liu, Z. Yang, Y. Guo *et al.*, "A novel correlation feature self-assigned Kolmogorov-Arnold networks for multi-energy load forecasting in integrated energy systems," *Energy Conversion and Management*, vol. 325, p. 119388, Feb. 2025.

[109] C. Wang, Y. Wang, Z. Ding *et al.*, "A transformer-based method of multienergy load forecasting in integrated energy system," *IEEE Transactions on Smart Grid*, vol. 13, no. 4, pp. 2703-2714, Jul. 2022.

[110] J. Fan, W. Zhuang, M. Xia *et al.*, "Optimizing attention in a transformer for multihorizon, multienergy load forecasting in integrated energy systems," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 8, pp. 10238-10248, Aug. 2024.

[111] P. Zhao, W. Hu, D. Cao *et al.*, "Probabilistic multienergy load forecasting based on hybrid attention-enabled transformer network and Gaussian process-aided residual learning," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 6, pp. 8379-8393, Jun. 2024.

[112] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, Jun. 2018, pp. 7482-7491.

[113] Z. Chen, V. Badrinarayanan, C.-Y. Lee *et al.*, "Gradnorm: gradient normalization for adaptive loss balancing in deep multitask networks," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, Jul. 2018, pp. 794-803.

[114] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proceedings of the 32nd Conference on Neural Information Processing Systems*, Montreal, Canada, Dec. 2018, pp. 527-538.

[115] L. Zhang, Q. Yang, X. Liu *et al.*, "Rethinking hard-parameter sharing in multi-domain learning," in *Proceedings of 2022 IEEE International Conference on Multimedia and Expo*, Taipei, China, Jul. 2022, pp. 1-6.

[116] N. Pahari and K. Shimada, "Multi-task learning using BERT with soft parameter sharing between layers," in *Proceedings of 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems*, Ise, Japan, Dec. 2022, pp. 1-6.

[117] X. Sun, R. Panda, R. Feris *et al.*, "Adashare: learning what to share for efficient deep multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8728-8740, Dec. 2020.

[118] S. Chi, Y. Tian, F. Wang *et al.*, "Deep semisupervised multitask learning model and its interpretability for survival analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3185-3196, Aug. 2021.

[119] D. Li, A. Sharma, and H. R. Zhang, "Scalable multitask learning using gradient-based estimation of task affinity," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona, Spain, Aug. 2024, pp. 1542-1553.

[120] Y. Zhang, Q. Cui, L. Shi *et al.*, "PPenergyNET: privacy-preserving multi-energy load forecasting in energy internet considering energy coupling," *IEEE Transactions on Power Systems*, vol. 39, no. 5, pp. 6235-6248, Sept. 2024.

[121] H. Wu and Z. Xu, "Multi-energy load forecasting in integrated energy systems: a spatial-temporal adaptive personalized federated learning approach," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 10, pp. 12262-12274, Oct. 2024.

[122] H. Wang, S. Xing, C. Si *et al.*, "Multi-gradient-descent federated learning with parity for cooperative short-term load forecasting," *IEEE Transactions on Power Systems*, vol. 40, no. 2, pp. 1199-1213, Mar. 2025.

[123] C. Huang, S. Bu, W. Chen *et al.*, "Deep reinforcement learning-assisted federated learning for robust short-term load forecasting in electricity wholesale markets," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 5, pp. 5073-5086, Sept. 2024.

[124] X. Song, Z. Chen, J. Wang *et al.*, "A multi-stage LSTM federated forecasting method for multi-loads under multi-time scales," *Expert Sys-*

tems with Applications, vol. 253, p. 124303, Nov. 2024.

[125] N. Lu, S. Liu, Q. Wen *et al.*, "Federated domain separation for distributed forecasting of non-IID household loads," *IEEE Transactions on Smart Grid*, vol. 15, no. 4, pp. 4271-4283, Jul. 2024.

[126] Y. Wang, N. Gao, and G. Hug, "Personalized federated learning for individual consumer load forecasting," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 1, pp. 326-330, Jan. 2023.

[127] G. Singh and J. Bedi, "A federated and transfer learning based approach for households load forecasting," *Knowledge-Based Systems*, vol. 299, p. 111967, Sept. 2024.

[128] Y. Dong, Y. Wang, M. Gama *et al.*, "Privacy-preserving distributed learning for residential short-term load forecasting," *IEEE Internet of Things Journal*, vol. 11, no. 9, pp. 16817-16828, May 2024.

[129] M. A. Husnoo, A. Anwar, N. Hosseinzadeh *et al.*, "A secure federated learning framework for residential short-term load forecasting," *IEEE Transactions on Smart Grid*, vol. 15, no. 2, pp. 2044-2055, Mar. 2024.

[130] S. M. Danish, A. Hameed, A. Ranjha *et al.*, "Block-fedl: electric vehicle charging load forecasting using federated learning and blockchain," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 2, pp. 2048-2056, Feb. 2025.

[131] J. D. Fernández, S. P. Menci, C. M. Lee *et al.*, "Privacy-preserving federated learning for residential short-term load forecasting," *Applied Energy*, vol. 326, p. 119915, Nov. 2022.

[132] P. Kairouz, H. B. McMahan, B. Avent *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1, pp. 201-210, Jun. 2021.

[133] D. Qin, Q. Wen, Z. Zhou *et al.*, "Deciding when to use a personalized model for load forecasting," *IEEE Transactions on Smart Grid*, vol. 16, no. 1, pp. 411-423, Jan. 2025.

[134] T. Wang, T. Tang, Z. Cai *et al.*, "Federated learning-based information leakage risk detection for secure medical internet of things," *ACM Transactions on Internet Technology*, vol. 24, no. 4, pp. 1-22, Nov. 2024.

[135] X. Qu, C. Guan, G. Xie *et al.*, "Personalized federated learning for heterogeneous residential load forecasting," *Big Data Mining and Analytics*, vol. 6, no. 4, pp. 421-432, Dec. 2023.

[136] Q. Mao, L. Wang, Y. Long *et al.*, "A blockchain-based framework for federated learning with privacy preservation in power load forecasting," *Knowledge-Based Systems*, vol. 284, p. 111338, Jan. 2024.

[137] A. Tan, H. Yu, L. Cui *et al.*, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9587-9603, Dec. 2023.

[138] S. S. Shinde, A. Bozorgchenani, D. Tarchi *et al.*, "On the design of federated learning in latency and energy constrained computation offloading operations in vehicular edge computing systems," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 2041-2057, Feb. 2022.

[139] Y. Chen, X. Sun, and Y. Jin, "Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 4229-4238, Oct. 2020.

[140] P. Li, Z. Hu, Y. Shen *et al.*, "Short-term electricity load forecasting based on large language models and weighted external factor optimization," *Sustainable Energy Technologies and Assessments*, vol. 82, p. 104449, Feb. 2025.

[141] A. Das, W. Kong, R. Sen *et al.*, "A decoder-only foundation model for time-series forecasting," in *Proceedings of Forty-first International Conference on Machine Learning*, Vienna, Austria, Jul. 2024, pp. 1-12.

[142] W. F. Wiggins and A. S. Tejani, "On the opportunities and risks of foundation models for natural language processing in radiology," *Radiology: Artificial Intelligence*, vol. 4, no. 4, p. e220119, Jul. 2022.

[143] Y. Chen and H. Xie, "Trace: unlocking the potential of LLMs in time series forecasting for distributed energy resources," *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 2, pp. 1-14, Feb. 2025.

[144] W. Liao, S. Wang, D. Yang *et al.*, "TimeGPT in load forecasting: a large time series model perspective," *Applied Energy*, vol. 379, p. 124973, Feb. 2025.

[145] Y. Zhou and M. Wang, "Empower pre-trained large language models for building-level load forecasting," *IEEE Transactions on Power Systems*, vol. 40, no. 3, pp. 1-12, May 2025.

[146] Y. Deng, D. Xie, R. Liang *et al.*, "Buildprog: program generation for testing ML-based building load forecasting models via LLM and prompt engineering," in *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, Hangzhou, China, Nov. 2024, pp. 248-249.

[147] C. Zhang, J. Zhang, Y. Zhao *et al.*, "Automated data-driven building energy load prediction method based on generative pre-trained transformers (GPT)," *Energy*, vol. 318, p. 134824, Mar. 2025.

[148] M. Meyer, D. Zapata, S. Kaltenpoth *et al.* (2024, Oct.). Benchmarking time series foundation models for short-term household electricity load forecasting. [Online]. Available: https://arxiv.org/abs/2410.09487

[149] Z. Wang, Q. Wen, C. Zhang *et al.* (2023, Jun.). DiffLoad: uncertainty quantification in load forecasting with diffusion model. [Online]. Available: https://arxiv.org/abs/2306.01001

[150] Z. Qiu, C. Li, Z. Wang *et al.* (2024, Nov.). EF-LLM: energy forecasting LLM with AI-assisted automation, enhanced sparse prediction, hallucination detection. [Online]. Available: https://arxiv.org/abs/2411.00852

[151] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *Proceedings of Advances in Neural Information Processing Systems*, Long Beach, USA, Dec. 2017, pp. 5998-6008.

[152] A. de S. Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6037-6124, Dec. 2022.

[153] J. Li, D. Li, Y. Yang *et al.* (2025, Feb.). Zero-shot load forecasting for integrated energy systems: a large language model-based framework with multi-task learning. [Online]. Available: https://arxiv.org/abs/2502.16896

[154] V. Ekambaram, A. Jati, N. H. Nguyen *et al.* (2024, Jan.). TTMS: fast multi-level tiny time mixers for improved zero-shot and few-shot forecasting of multivariate time series. [Online]. Available: https://arxiv.org/abs/2401.03955

[155] V. Ekambaram, A. Jati, N. Nguyen *et al.* (2023, Jun.). TSMixer: lightweight MLP-mixer model for multivariate time series forecasting. [Online]. Available: https://arxiv.org/abs/2306.09364

[156] H. Wu, T. Hu, Y. Liu *et al.*, "TimesNet: temporal 2D-variation modeling for general time series analysis," in *Proceedings of The Eleventh International Conference on Learning Representations*, Kigali, Rwanda, May 2023, pp. 1-12.

[157] H. Wen, Y. Lin, Y. Xia *et al.*, "DiffSTG: probabilistic spatio-temporal graph forecasting with denoising diffusion models," in *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, Hamburg, Germany, Nov. 2023, pp. 1-12.

[158] C. Zhou, Q. Li, C. Li *et al.*, "A comprehensive survey on pretrained foundation models: a history from Bert to ChatGPT," *International Journal of Machine Learning and Cybernetics*, vol. 15, pp. 1-65, Jan. 2024.

[159] Q. Sun, Q. Yu, Y. Cui *et al*. (2023, Jul.). EMU: generative pretraining in multimodality. [Online]. Available: https://arxiv.org/abs/2307.05222

[160] Z. Jiang, T. Chen, T. Chen *et al.*, "Robust pre-training by adversarial contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16199-16210, Dec. 2020.

[161] X. Wang, R. Zhang, C. Shen *et al.*, "Dense contrastive learning for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, Jun. 2021, pp. 3024-3033.

[162] Z. Yue, Y. Wang, J. Duan *et al.*, "TS2VEC: towards universal representation of time series," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, Virtual Event, Feb. 2022, pp. 8980-8987.

[163] J. Dong, H. Wu, H. Zhang *et al.*, "SIMMTM: a simple pre-training framework for masked time-series modeling," *Advances in Neural Information Processing Systems*, vol. 36, pp. 29996-30025, Dec. 2023.

[164] C. Wei, H. Fan, S. Xie *et al.*, "Masked feature prediction for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, Jun. 2022, pp. 14668-14678.

[165] S. Tu, Y. Zhang, J. Zhang *et al.*, "PowerPM: foundation model for power systems," *Advances in Neural Information Processing Systems*, vol. 37, pp. 115233-115260, Dec. 2024.

[166] X. Zhang, Z. Zhao, T. Tsiligkaridis *et al.*, "Self-supervised contrastive pre-training for time series via time-frequency consistency," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3988-4003, Dec. 2022.

[167] E. Eldele, M. Ragab, Z. Chen *et al.*, "Self-supervised contrastive representation learning for semi-supervised time-series classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15604-15618, Dec. 2023.

[168] Y. Park, F. Germain, J. Liu *et al.* (2024). Probabilistic forecasting for building energy systems: are time-series foundation models the answer? [Online]. Available: https://arxiv.org/abs/2410.14217

[169] H. Xue and F. D. Salim, "Utilizing language models for energy load forecasting," in *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, Istanbul, Turkey, Nov. 2023, pp. 224-227.

[170] E. Hu, Y. Shen, P. Wallis *et al.* (2022, Jun.). Lora: low-rank adaptation of large language models. [Online]. Available: https://arxiv.org/abs/2106.09685
[171] D. Cao, F. Jia, S. O. Arik *et al.* (2023, Oct.). TEMPO: prompt-based generative pre-trained transformer for time series forecasting. [Online]. Available: https://arxiv.org/abs/2310.04948
[172] K. Rasul, A. Ashok, A. R. Williams *et al.* (2023, Oct.). Lag-LLAMA: towards foundation models for time series forecasting. [Online]. Available: https://arxiv.org/abs/2310.08278
[173] Y. Liang, H. Wen, Y. Nie *et al.*, "Foundation models for time series analysis: a tutorial and survey," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona, Spain, Aug. 2024, pp. 6555-6565.
[174] Z. Li, X. Qiu, P. Chen *et al.*, "TSFM-bench: a comprehensive and unified benchmark of foundation models for time series forecasting," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Toronto, Canada, Aug. 2025, pp. 5595-5606.
[175] A. Garza, C. Challu, and M. Mergenthaler-Canseco. (2023, Oct.). TimeGPT-1. [Online]. Available: https://arxiv.org/abs/2310.03589
[176] C. Qian, M. Liu, E. Reif *et al.*, "LLM adoption in data curation workflows: industry practices and insights," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, Apr. 2025, pp. 1-10.
[177] S. Goyal, P. Maini, Z. C. Lipton *et al.*, "Scaling laws for data filtering – data curation cannot be compute agnostic," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, Jun. 2024, pp. 22702-22711.
[178] M. Tan, M. Merrill, V. Gupta *et al.*, "Are language models actually useful for time series forecasting?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 60162-60191, Dec. 2024.
[179] A. F. Ansari, L. Stella, C. Turkmen *et al.* (2024, Mar.). Chronos: learning the language of time series. [Online]. Available: https://arxiv.org/abs/2403.07815
[180] R. Yang, Y. Li, C. Hu *et al.*, "Exdiffusion: classifier-guidance diffusion model for extreme load scenario generation with extreme value theory," *IEEE Transactions on Smart Grid*, vol. 16, no. 2, pp. 1-12, Mar. 2025.
[181] J. Han, L. Yan, and Z. Li, "A task-based day-ahead load forecasting model for stochastic economic dispatch," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5294-5304, Nov. 2021.
[182] X. Chen, Y. Yang, Y. Liu *et al.*, "Feature-driven economic improvement for network-constrained unit commitment: a closed-loop predict-and-optimize framework," *IEEE Transactions on Power Systems*, vol. 37, no. 4, pp. 3104-3118, Jul. 2022.

Pengfei Zhao received the B.E. degree in electrical engineering and its automation, the M. S. degree in electrical engineering, from the University of Electronic Science and Technology of China, Chengdu, China, in 2019 and 2022, respectively. From 2022 to 2023, he served as an Engineer at the Electric Power Research Institute of State Grid Ningxia Electric Power Co., Ltd., Yingchuan, China. He is currently working toward the Ph.D. degree in the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China. His research interests include forecasting technology and application of artificial intelligence in modern power system.

Weihao Hu received the B. Eng. and M. Sc. degrees from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2007, respectively, both in electrical engineering, and the Ph. D. degree from Aalborg University, Aalborg, Denmark, in 2012. He is currently a Full Professor and the Director of Institute of Smart Power and Energy Systems (ISPES) at the University of Electronics Science and Technology of China, Chengdu, China. His research interests include artificial intelligence in modern power system and renewable power generation.

Di Cao received the Ph. D. degree in control science and engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2021. He is currently an Associate Professor at UESTC. His research interests include optimization of distribution network and application of machine learning in power system.

Zhaochen Dong received the bachelor's degree in economics from Dongbei University of Finance and Economics, Dalian, China. He is currently pursuing the master's degree in electrical engineering at the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include electrical load forecasting and renewable energy forecasting.

Yaqi Zeng is currently pursuing the B.S. degree in communication engineering at the University of Electronic Science and Technology of China, Chengdu, China. His research interests include forecasting technique and application of artificial intelligence in modern power system.

Qi Hunag received the B. S. degree in electrical engineering from Fuzhou University, Fuzhou, China, in 1996, the M.S. degree from Tsinghua University, Beijing, China, in 1999, and the Ph.D. degree from Arizona State University, Tempe, USA, in 2003. He was the Vice President of the Chengdu University of Technology, Chengdu, China, and President of the Southwest University of Science and Technology, Mianyang, China. He is currently a Professor at the University of Electronic Science and Technology of China, Chengdu, China. He also serves as the Director of the Institute of Scientific and Technical Information of China, Beijing, China. His current research interests include power system instrumentation, power system monitoring and control, and informatics for smart electric energy system.

Zhe Chen received the B.Eng. and M.Sc. degrees from the Northeast China Institute of Electric Power Engineering, Jilin, China, and the Ph. D. degree from the University of Durham, Durham, UK. He is a Full Professor with the Department of Energy Technology, Aalborg University, Aalborg, Denmark. His research interests include power electronics, electric machine, wind energy, and modern power system.