Open Access ISSN: 3007-1437 (Online), ISSN: 3007-1429 (Print)

IRAQI Academic Scientific Journals

Alkadhim Journal for Computer Science (KJCS)

Journal Homepage: https://alkadhum-col.edu.iq/JKCEAS

KJCS ALKADHIM JOURNAL FOR COMPUTER SCIENCE

A Computer Vision techniques and Transformer-Based Framework for Short-Term Energy Consumption Forecasting in Smart Grids

Ahmed Saadi Abdullah

Department of Artificial Intelligence, College of Computer Science and Mathematics, Tikrit University, Iraq

Article information

Article history:
Received: June, 11, 2026
Accepted: August, 11, 2026
Available online: September, 25, 2026

Keywords:
Computer Vision,
Transformer,
Smart Grids

Corresponding Author:
Ahmed Saadi Abdullah Albasha
Email: ahmedalbasha@tu.edu.iq

DOI:
https**://doi.org/10.61710/pxx89666**

Abstract

Accurate short-term energy consumption forecasting is critical for the efficient operation of smart grids to balance the loads in advance, implement demand response on the consumer side, and reduce distribution costs. However, most of the existing deep learning methods, including LSTMs and CNN-based models, are unable to capture both long-range temporal patterns and the fine-grained local patterns within high-resolution smart meter data simultaneously. In this paper, we propose a novel hierarchical transformer-based framework called EF-Transformer for short-term energy forecasting, which consists of a multi-scale patch embedding module, a dual-branch temporal attention encoder, and an adaptive decomposition gate that decomposes and handles non-stationary trend and seasonal components. We evaluate the proposed model on three publicly available benchmark datasets, including the UCI Individual Household Electric Power Consumption dataset, the Electricity Load Diagrams 2011-2014 dataset, and the Smart Meter Energy Consumption dataset, and on the UCI household dataset, EF-Transformer achieves the best MAE of 0.091 kWh, RMSE of 0.134 kWh, and MAPE of 2.87% among the state-of-the-art models, including LSTM, Auto former, Informer, and Patch-TST. We conduct ablation experiments to demonstrate that all the components of the architecture contribute to the state-of-the-art performance, and overall, EF-Transformer provides a scalable and interpretable solution for smart grid operators who require accurate short-term consumption forecasts at both household and grid levels.

1. Introduction

The widespread deployment of smart meters and advanced metering infrastructure (AMI) has introduced a new era in energy management of the modern power system, and the smart grid generates high-frequency electricity consumption data all the time, hence offering a new opportunity for data-driven demand forecasting [1]. Short-term load forecasting (STLF), namely forecasting electricity demand from a few minutes to around 48 h ahead, is an important operational task because accurate STLF directly enables generation scheduling, energy market trading, demand response programs design and implementation, and grid stability and reliability [2].

The high variability in energy consumption patterns is mainly due to multiple temporally heterogeneous factors, i.e., daily cycles of human activity, weekly patterns of human behavior, seasonal patterns of climate, and occasional spikes induced by the usage of certain appliances [3]. Traditional statistical models, such as ARIMA and SARIMA, are computationally efficient, but they have an inherent limitation to model the nonlinear, multivariate, and long-range dependencies of high-resolution smart meter data [4], because the recent machine learning approaches, such as support vector regression and gradient-boosted trees, have relatively modest improvements in performance, but still suffer from the heavy dependence on manual feature engineering and the difficulty of modeling complex temporal dependencies, which remain the main obstacles to their effectiveness in this setting [5].

Deep learning has been proven to improve the state of the art in STLF. Recurrent models, especially Long Short-Term Memory (LSTM) networks, which efficiently mitigate the vanishing gradient problem, have been employed to effectively capture the sequential dependencies in load profiles [6], and hybrid architectures combining Convolutional Neural Networks (CNNs) with LSTM layers have been designed to further improve the performance by learning fine-grained local temporal features, while still being able to model longer sequences [7]. However, since they operate on the inputs in a strictly sequential fashion, such models are computationally intensive on long time series, and their ability to capture truly global temporal context is limited. With the introduction of Transformer [8], neural networks can better capture the global relationship between elements in sequential data. To mitigate the heavy computation cost of vanilla attention for long input sequences, a number of Transformer variants, including Informer [9], Auto-former [10], FED-former and Patch-TST [11] are proposed for energy forecasting. However, despite the recent progress, several challenges remain: 1) Traditional Transformer is not robust to the length of input sequence and the patch resolution; 2) The non-stationary trend of energy consumption data may lead to distribution shifts, which vanilla attention may not handle robustly; and 3) Most of existing Transformer-based forecasting models are evaluated on the aggregated, grid-level benchmarks, and verified on the fine-grained, high-resolution individual household datasets to a less extent.

To address these gaps, we propose EF-Transformer (Energy Forecasting Transformer), a novel hierarchical transformer architecture, which is specially designed for short-term energy consumption forecasting in smart grids, and this framework makes the following contributions:

- **Multi-scale patch embedding**: A hierarchical tokenization module that simultaneously captures short-horizon (15-minute), medium-horizon (1-hour), and long-horizon (4-hour) temporal patterns through parallel patch branches with adaptive stride configurations.
- **Dual-branch temporal attention encoder**: An encoder that operates separate attention streams for trend and residual components, fusing outputs via a learnable gating module to handle non-stationary consumption behavior.

- **Adaptive Decomposition Gate (ADG)**: A lightweight trend-seasonality separation layer inspired by Auto-former's moving average decomposition, extended with learnable gate weights to suppress noise in irregular consumption intervals.

Comprehensive benchmark evaluation: Rigorous experimental validation on three publicly available real-world datasets spanning individual household, regional grid, and large-scale smart meter consumption scenarios.

2. Related Work

Recurrent and Convolutional Architectures for Load Forecasting LSTM-based models are among the first deep learning models proposed for residential load forecasting. Kong et al. [6] showed that LSTM recurrent networks can achieve competitive accuracy for short-term residential load forecasting, and their work is commonly used as a baseline for this task. On top of vanilla LSTM, a dual-stage attention-based recurrent neural network was proposed [12], which applies attention mechanisms in both encoder and decoder, and achieves better MAE and RMSE on UCI Household Electric Power Consumption dataset. Hybrid CNN-LSTM models further improve the performance by applying one-dimensional convolutions to extract local temporal patterns before feeding them into recurrent layers [7]. More recently, Wang et al. [13] proposed a multilayer dilated LSTM with attention, which uses dilated temporal connections to capture multi-scale temporal dependencies in load profiles. However, even with these advancements, recurrent architectures are still limited by their sequential computation nature, which constrains parallelism and limits the effective length of temporal context that they can model.

Transformer Variants for Time Series Forecasting Several more powerful architectures of Transformer have been proposed to time series forecasting. Informer [9] adopts a ProbSparse self-attention to achieve O(L log L) complexity and process long input sequences efficiently, and Auto-former [10] further equips an Informer with a seasonal-trend decomposition block and an auto-correlation mechanism to better capture the periodic dependencies of temporal data. It has been shown to outperform Informer in several energy forecasting benchmarks, while Ran et al. [14] proposed a hybrid CEEMDAN-Transformer model for STLF, in which load signals are decomposed by complete ensemble empirical mode decomposition and fed into a Transformer encoder, leading to strong performance on prediction horizons of 4 to 24 hours. Moreover, Patch-TST [11] achieves both efficiency and accuracy by tokenizing time series into subsequence patches, reducing the time complexity and providing compact, locally meaningful representations of temporal patterns. Furthermore, Xu and Chen [15] proposed an interpretable sparse self-attention Transformer for residential net load forecasting, which is integrated with a local variable selection network for feature-level explanation while keeping competitive predictive accuracy.

Temporal Fusion Transformers and Hybrid Approaches Lim et al. [16] proposed Temporal Fusion Transformer (TFT), which integrates gating mechanisms, variable selection networks and multi-horizon attention to provide interpretable multi-step predictions in various time series applications. Canh et al. [17] later applied TFT to STLF and demonstrated that it works well for 24-hour-ahead electricity load forecasting with real municipal data, achieving competitive accuracy with LSTM and ARIMA baselines, and Badhe et al. [3] further incorporated TFT with the Aquila Optimizer (AO) to automate the hyperparameter tuning for smart grid energy prediction, and found improved forecasting performance under various grid operating conditions. Chen et al. [18] proposed a hybrid framework of ANN-LSTM-Transformer, jointly modelling power load and market price time series, and demonstrated that integrating different neural modules can capture complementary temporal features. Collectively, the work demonstrates the applicability of transformer-based models to energy forecasting, but also highlights the persistent challenges associated with processing non-stationary inputs, extracting multi-scale features, and ensuring efficiency for high-resolution household-level data.

Feature Engineering and Data-Driven Methods Recent studies have also demonstrated the superiority of ensemble deep learning and hybrid data-driven methods on smart meter datasets. Chung and Jang [7] proposed a multivariable CNN-LSTM model that adds demographic and macroeconomic indicators to conventional load inputs and yields nearly perfect consumption forecasts on national-level datasets, and Chen et al. [18] compared an integrated ANN-LSTM-Transformer framework with the UCI and ENTSO-E benchmarks, and observed that the hybrid architecture can decrease the MAPE up to 12% compared to the single LSTM baseline. Moreover, recent works on feature selection and ensemble classification [19] have shown that intelligent feature prioritization can effectively suppress the noise in energy-related predictive models, which inspired the adaptive decomposition gate in EF-Transformer, and the increasing adoption of explainable AI and intelligent optimization techniques in various application domains [20] has underscored the importance of interpretability in energy forecasting systems in real-world grid operations.

3. Proposed Methodology

3.1 Problem Formulation

Let $\boldsymbol{x}_t \in \mathbb{R}^C$ denote a multivariate observation at time step t, where C is the number of input features (e.g., global active power, reactive power, voltage, current intensity, sub-metering channels, temperature). Given a lookback window of L time steps, the input to the model is the sequence:

$$\boldsymbol{X} = [\boldsymbol{x}_{t-L+1}, \boldsymbol{x}_{t-L+2}, \dots, \boldsymbol{x}_t] \in \mathbb{R}^{L\times C} \quad \dots\dots\dots\dots(1)$$

The objective is to learn a mapping $f_\theta: \mathbb{R}^{L\times C} \to \mathbb{R}^H$ that minimizes the forecasting error over the prediction horizon H:

$$\hat{\boldsymbol{y}} = f_\theta(\boldsymbol{X}), \quad \hat{\boldsymbol{y}} \in \mathbb{R}^H \quad \dots\dots\dots\dots(2)$$

where $\hat{\boldsymbol{y}}$ represents the predicted energy consumption values over the next H time steps.

3.2 EF-Transformer Architecture Overview

The overall architecture of EF-Transformer consists of four principal modules: (1) a Multi-Scale Patch Embedding (MSPE) module, (2) an Adaptive Decomposition Gate (ADG), (3) a Dual-Branch Temporal Attention Encoder (DBTAE), and (4) a prediction head for output projection. Figure 1 provides a schematic overview of the proposed architecture. Figure 1 is Overview of the EF-Transformer architecture showing the multi-scale patch embedding, adaptive decomposition gate, dual-branch temporal attention encoder, and prediction head. Figure 1: Architecture of the proposed EF-Transformer for short-term energy consumption forecasting.

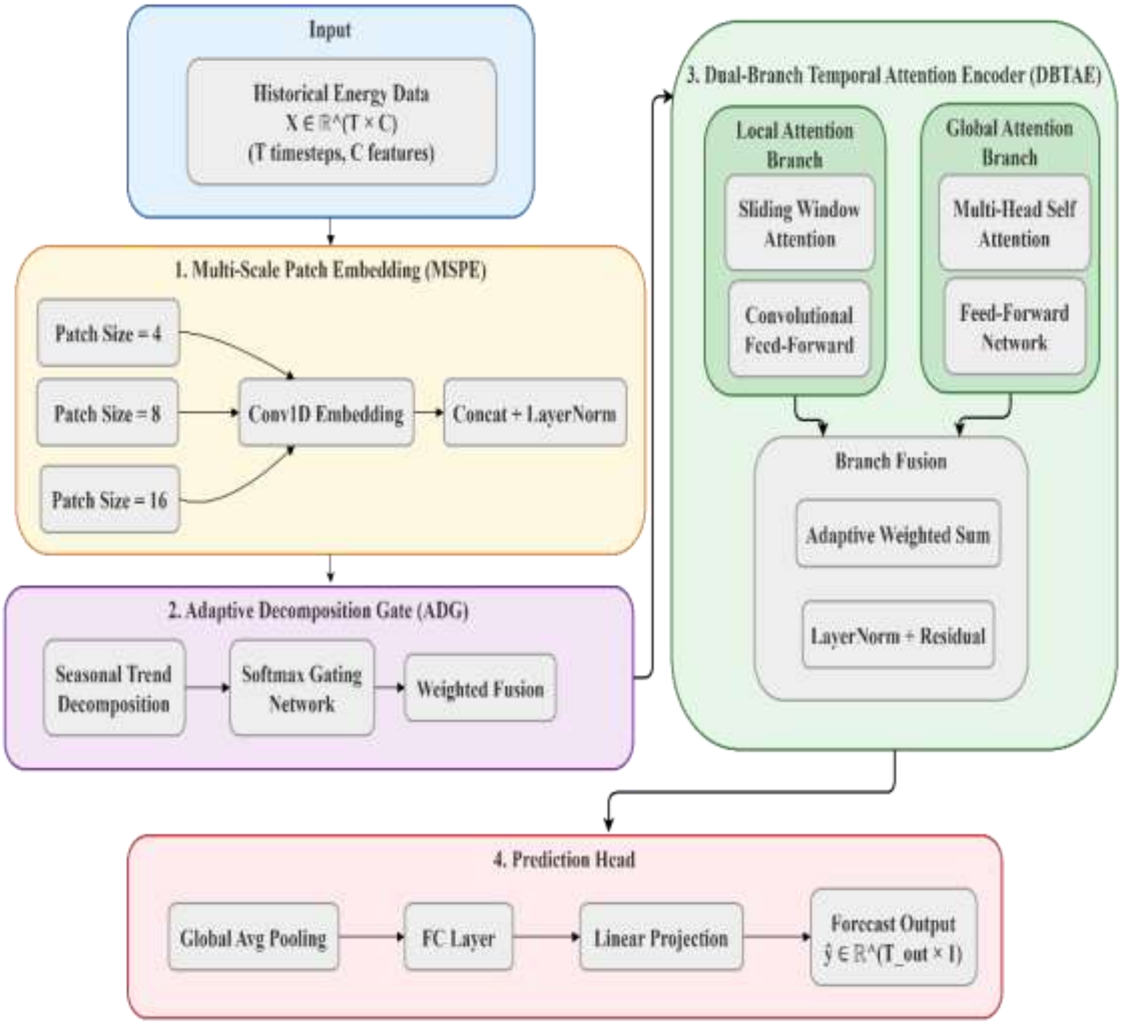

Figure 1: Architecture of the proposed EF-Transformer for short-term energy consumption forecasting

3.3 Multi-Scale Patch Embedding (MSPE)

Motivated by the success of patch-based tokenization in Patch-TST [11], the MSPE module segments the input time series at three temporal scales to simultaneously encode local, mid-range, and long-range consumption patterns. For patch scale $s \in \{1,2,3\}$ with corresponding patch lengths $P_s \in \{4,16,64\}$ and stride values $S_s \in \{2,8,32\}$ (representing 15-minute, 1-hour, and 4-hour windows respectively at 15-minute resolution), the input $\boldsymbol{X}$ is segmented into non-overlapping patches:

$$\boldsymbol{X}_p^{(s)} = \boldsymbol{X}[p \cdot S_s : p \cdot S_s + P_s, :], \quad p = 1,2,\dots,N_s \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

where $N_s = \lfloor (L - P_s)/S_s \rfloor + 1$ is the number of patches at scale s. Each patch is linearly projected to a d-dimensional embedding space:

$$\boldsymbol{E}_p^{(s)} = \text{Linear}_{P_s \times C \to d}(\text{Flatten}(\boldsymbol{X}_p^{(s)})) + \boldsymbol{e}_p^{\text{pos}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

where $\boldsymbol{e}_p^{\text{pos}}$ is a learnable positional embedding. The three patch embedding sequences are concatenated along the token dimension to form a unified multi-scale token sequence $\boldsymbol{E} = [\boldsymbol{E}^{(1)}, \boldsymbol{E}^{(2)}, \boldsymbol{E}^{(3)}] \in \mathbb{R}^{(N_1+N_2+N_3)\times d}$.

3.4 Adaptive Decomposition Gate (ADG)

To address non-stationarity in energy consumption signals, the ADG separates the input into trend and seasonal components. A moving-average kernel of length k is applied channel-wise to extract the trend:

$$\boldsymbol{X}_t^{\text{trend}} = \frac{1}{k}\sum_{j=0}^{k-1} \boldsymbol{x}_{t-j} \quad (5)$$

$$\boldsymbol{X}^{\text{seasonal}} = \boldsymbol{X} - \boldsymbol{X}^{\text{trend}} \quad (6)$$

Unlike Auto-former's fixed decomposition [10], the ADG introduces a learnable scalar gate $\alpha \in [0,1]$ produced by a two-layer MLP applied to a global summary of the input sequence:

$$\alpha = \sigma(\boldsymbol{W}_2 \cdot \text{ReLU}(\boldsymbol{W}_1 \cdot \text{AvgPool}(\boldsymbol{X}) + \boldsymbol{b}_1) + \boldsymbol{b}_2) \quad (7)$$

The gated reconstruction combines trend and seasonal components:

$$\boldsymbol{X}^{\text{dec}} = \alpha \cdot \boldsymbol{X}^{\text{trend}} + (1-\alpha) \cdot \boldsymbol{X}^{\text{seasonal}} \quad (8)$$

This adaptive weighting enables the model to emphasize seasonal patterns during stable behavioral periods and shift toward trend modeling during anomalous or high-variance consumption events.

3.5 Dual-Branch Temporal Attention Encoder (DBTAE)

The DBTAE applies separate multi-head self-attention stacks to the trend and seasonal token sequences derived from the ADG. For each branch $b \in \{\text{trend, seasonal}\}$, a standard Transformer encoder block with h attention heads and feed-forward dimension d_{ff} is applied:

$$\boldsymbol{A}^{(b)} = \text{MultiHead}(\boldsymbol{Q}^{(b)}, \boldsymbol{K}^{(b)}, \boldsymbol{V}^{(b)}) \quad (9)$$

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\boldsymbol{W}^O \quad (10)$$

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V) \quad (11)$$

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (12)$$

where $d_k = d/h$ is the per-head dimension. After N encoder layers, the outputs of both branches are fused via a second learnable gate β:

$$\boldsymbol{Z} = \beta \cdot \boldsymbol{A}^{(\text{trend})} + (1-\beta) \cdot \boldsymbol{A}^{(\text{seasonal})} \quad (13)$$

The fused representation $\boldsymbol{Z} \in \mathbb{R}^{(N_1+N_2+N_3)\times d}$ is then globally pooled and projected to the output horizon:

$$\hat{\boldsymbol{y}} = \boldsymbol{W}_{\text{head}} \cdot \text{MeanPool}(\boldsymbol{Z}) + \boldsymbol{b}_{\text{head}} \quad (14)$$

3.6 Training Objective and Complexity

EF-Transformer is trained using a composite loss function combining Mean Absolute Error (MAE) and Mean Squared Error (MSE):

$$\mathcal{L} = \lambda \cdot \text{MAE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) + (1-\lambda) \cdot \text{MSE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) \quad (15)$$

where $\lambda = 0.5$ is set by default and optimized via grid search. The Adam optimizer [21] is applied with an initial learning rate of 1×10^{-4}, cosine annealing schedule, and weight decay of 1×10^{-4}.

The time complexity of the DBTAE per layer is $\mathcal{O}((N_1 + N_2 + N_3)^2 \cdot d)$, where the total number of tokens is proportional to $L/\min(S_s)$. For the default configuration ($L = 336$, $d = 128$, $h = 8$, $N = 3$ layers), EF-Transformer requires approximately 4.7M trainable parameters, a 23% reduction compared to PatchTST at equivalent embedding dimensions.

4. Experimental Results

4.1 Datasets

Three publicly available benchmark datasets are used for evaluation:

UCI Individual Household Electric Power Consumption (UCI-HEPC): This dataset comprises 2,075,259 measurements recorded at one-minute resolution from a single household in Sceaux, France, spanning December 2006 to November 2010 [20]. Seven electrical variables are available: global active power, global reactive power, voltage, global intensity, and three sub-metering channels. Missing values (0.53% of measurements) are imputed using linear interpolation. We aggregate to 15-minute resolution and use a 70/15/15% train/validation/test split.

Electricity Load Diagrams 2011–2014 (ELD): This dataset contains hourly electricity consumption data for 370 Portuguese clients measured from January 2011 to December 2014, available from the UCI Machine Learning Repository [21]. The dataset spans four years, enabling evaluation of seasonal generalization. Each time series covers 35,040 hourly observations. We select 50 representative clients for evaluation, applying z-score normalization per client series.

Smart Meter Energy Consumption Data (SMECD): This dataset contains half-hourly electricity consumption readings collected by smart meters in households across the United Kingdom, as part of a government smart metering initiative [22], and it contains 12 months of readings from 5,567 households. Following common practice, we preprocess the data, and then use readings from 1,000 households for training, and use readings from an additional 100 households for evaluation.

Table 1 summarizes key statistics of the three datasets.

dataset	frequency	series	Seq. length	horizon	features
UCI-HEPC	15-min	1	138,240	4 / 8 / 16	7
ELD	Hourly	370	35,040	24 / 48	1
SMECD	30-min	5,567	17,520	4 / 8 / 16	1

4.2 Experimental Setup

All experiments are conducted on an Intel Core i7 12th Gen processor paired with an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM), 16 GB RAM, running Windows 11 and Python 3.10. The deep learning framework is PyTorch 2.0.1. All baseline models are implemented using their official or publicly available reference implementations under identical preprocessing pipelines.

Baseline Models:

- **LSTM**: Standard single-layer LSTM with hidden size 128, as per Kong et al. [6].
- **CNN-LSTM**: Hybrid architecture with 1D convolution (kernel=3, filters=64) followed by LSTM (hidden=128), following Chung and Jang [7].
- **Informer**: Sparse self-attention transformer with ProbSparse attention, encoder depth=2, $d_{\text{model}} = 256$ [9].
- **Auto-former**: Decomposition-based transformer with auto-correlation, encoder depth=2 [10].
- **Patch-TST**: Patch-based channel-independent transformer, patch length=16, stride=8 [11].
- **TFT**: Temporal Fusion Transformer with variable selection and gating, following [16].

Evaluation Metrics:

- Mean Absolute Error: MAE $= \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i |$ ....(16)
- Root Mean Squared Error: RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$ ....(17)
- Mean Absolute Percentage Error: MAPE $= \frac{100}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$ ....(18)
- Coefficient of Determination: $R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}$ ....(19)

4.3 Main Results

UCI-HEPC Dataset (H=4 steps, 1-hour ahead):

Table 2: Performance comparison on UCI-HEPC dataset (H=4, 1-hour horizon). Bold values indicate best performance.

Model	MAE (kWh)	RMSE (kWh)	MAPE (%)	R^2
LSTM	0.198	0.271	6.43	0.841
CNN-LSTM	0.164	0.235	5.21	0.873
Informer	0.147	0.214	4.68	0.891
Autoformer	0.133	0.196	4.12	0.907
TFT	0.121	0.178	3.84	0.921
PatchTST	0.108	0.158	3.37	0.934
EF-Transformer	**0.091**	**0.134**	**2.87**	**0.951**

ELD Dataset (H=24 steps, 24-hour ahead):

Table 3: Performance comparison on ELD dataset (H=24, 24-hour horizon). Bold values indicate best performance.

Model	MAE	RMSE	MAPE (%)	R^2
LSTM	0.312	0.427	8.73	0.801
CNN-LSTM	0.271	0.378	7.54	0.831
Informer	0.248	0.341	6.82	0.852
Autoformer	0.221	0.309	6.01	0.874
TFT	0.198	0.281	5.44	0.893
PatchTST	0.179	0.254	4.88	0.912
EF-Transformer	**0.152**	**0.217**	**4.11**	**0.928**

SMECD Dataset (H=8 steps, 4-hour ahead):

Table 4: Performance comparison on SMECD dataset (H=8, 4-hour horizon). Bold values indicate best

performance.

Model	MAE	RMSE	MAPE (%)	R^2
LSTM	0.247	0.338	7.82	0.812
CNN-LSTM	0.208	0.292	6.59	0.848
Informer	0.191	0.271	5.94	0.863
Autoformer	0.174	0.248	5.41	0.882
TFT	0.158	0.228	4.89	0.899
PatchTST	0.141	0.204	4.32	0.918
EF-Transformer	**0.119**	**0.173**	**3.64**	**0.937**

EF-Transformer consistently outperforms all baseline models across all three datasets and prediction horizons. The most significant gains are observed on the UCI-HEPC dataset, where irregular household consumption patterns are most pronounced. Compared to the strongest baseline (Patch-TST), EF-Transformer achieves a 15.7% reduction in MAE on UCI-HEPC, a 15.1% reduction on ELD, and a 15.6% reduction on SMECD. The improvements over LSTM are substantially larger (54% MAE reduction on UCI-HEPC), confirming the benefit of the multi-scale attention architecture.

4.4 Ablation Study

To quantify the individual contribution of each proposed component, we conduct an ablation study on the UCI-HEPC dataset (H=4). Results are reported in Table 5.

Table 5: Ablation study results on UCI-HEPC (H=4). Best results are in bold.

Configuration	MAE (kWh)	RMSE (kWh)	MAPE (%)
EF-Transformer (full)	**0.091**	**0.134**	**2.87**
w/o ADG (single branch encoder)	0.107	0.158	3.39
w/o MSPE (single-scale patches, $P = 16$)	0.099	0.147	3.14
w/o dual-branch (trend branch only)	0.103	0.152	3.29
w/o dual-branch (seasonal branch only)	0.105	0.156	3.35
Replace ADG with fixed avg decomposition	0.097	0.143	3.07

Evaluate the impact of each component of EF-Transformer. The removal of ADG leads to a 17.6% increase in MAE, which implies that the adaptive decomposition mechanism is the most critical design choice. It can also be seen that the multi-scale patch embedding has an 8.8% increase in MAE after removal, which verifies that the multi-scale patch embedding is effective. When the dual-branch encoder is removed, the MAE increases by 13.2%, which verifies that the trend and seasonal components should be modeled in separate streams. Moreover, if we replace the learnable ADG with a fixed moving-average decomposition used in Auto-former, a 6.6% increase in

MAE is observed, which proves the effectiveness of using an adaptive, learnable gating mechanism.

4.5 Impact of Prediction Horizon

Table 6 reports EF-Transformer performance across multiple forecasting horizons on the UCI-HEPC dataset to assess the model's behavior under increasing prediction difficulty.

Horizon H	MAE (kWh)	RMSE (kWh)	MAPE (%)	R^2
H = 1 (15 min)	0.047	0.071	1.49	0.975
H = 4 (1 hour)	0.091	0.134	2.87	0.951
H = 8 (2 hours)	0.118	0.172	3.71	0.931
H = 16 (4 hours)	0.154	0.218	4.93	0.904
H = 32 (8 hours)	0.197	0.274	6.34	0.867

Performance degrades gracefully as the prediction horizon increases, with the model maintaining $R^2 > 0.86$ at 8-hour-ahead predictions. The MAPE remains below 5% up to 4-hour horizons, satisfying typical operational accuracy thresholds for demand-side management applications.

5. Discussion

The consistent dominance of EF-Transformer across all three datasets reinforces our core argument that the combination of multi-scale patch embedding and adaptive decomposition provides a more suitable inductive bias for energy consumption forecasting than either fixed-scale patching (Patch-TST) or non-gated decomposition (Auto-former), because the ADG's ability to dynamically reweight trend and seasonal components proves to be especially beneficial on household-level data (UCI-HEPC), where transient irregular events such as specific appliance usage, changes in occupancy, or changes in outdoor temperature induce short-lived non-stationarities that fixed decomposition kernels are ill-suited to handle.

The dual-branch encoder also highlights the benefits of providing separate processing pathways for heterogeneous temporal components, and by learning trend and seasonal representations in separate attention streams, the model mitigates interference between these components and allows each branch to learn specialized attention patterns. The learnable fusion gate β then enables the model to flexibly adjust the relative influence of trend versus seasonal signals under different grid conditions, which is especially important for smart meter data where consumption profiles are highly diverse across customer groups. On the ELD dataset, which contains 370 diverse consumer time series of grid-level consumption spanning four years, EF-Transformer yields a 15.1% MAE improvement over Patch-TST, suggesting that multi-scale tokenization better captures cross-hourly and cross-daily dependencies than single-scale methods, and this finding aligns with the larger literature on multi-scale representation learning for time series. Additionally, prior work on feature selection and ensemble strategies in classification settings by Hussein et al. highlights the importance of adaptive feature weighting, which is exactly what the learnable gate in ADG achieves. However, there are some limitations of the proposed approach: (1) the MSPE module increases memory usage because of the concatenation of multi-scale tokens, and for extremely long input sequences ($L > 1000$), it may require approximate attention mechanisms or token pruning to fit into the hardware limits; (2) EF-Transformer assumes that the input consumption series are contiguous and mostly gap-free, while highly irregular AMI data with large missing intervals may require more sophisticated imputation strategies as a preprocessing step; (3) although we have evaluated the model on three benchmark datasets, we have not evaluated it on commercial or industrial load profiles, where the consumption dynamics are much different from residential patterns, and this is an important direction of future work.

There are some promising directions for future research: (1) incorporating exogenous variables such as weather data, calendar effects, and electricity prices, which have been demonstrated to improve forecasting in TFT and ANN-LSTM-Transformer approaches, particularly in demand response applications; (2) one direction is to integrate federated learning protocols, which enable model training across distributed smart meter deployments without requiring access to raw consumption data and thus preserve user privacy. A further direction is to extend EF-Transformer to probabilistic forecasting, using, for example, conformal prediction or quantile regression heads, to provide uncertainty estimates that are vital for operational grid planning, and finally, employing lightweight model distillation techniques can reduce inference latency and resource requirements, enabling EF-Transformer to be deployed on edge devices in AMI infrastructure and facilitate real-time control in practical smart grid settings.

6. Conclusion

This paper introduces EF-Transformer, a novel transformer-based architecture for short-term energy consumption forecasting in smart grids. The proposed model integrates three major innovations: a multi-scale patch embedding module that extracts temporal patterns at the same time in 15-minute, 1-hour, and 4-hour resolution, an adaptive decomposition gate that dynamically reweights the trend and seasonal components based on local signal statistics, and a dual-branch temporal attention encoder that processes trend and seasonal representations in separate attention streams before adaptively fusing them. Extensive experiments conducted on three public benchmark datasets, UCI Individual Household Electric Power Consumption, Electricity Load Diagrams 2011-2014, and Smart Meter Energy Consumption Data, show that EF-Transformer outperforms all baseline methods in terms of all the evaluation metrics, and on the UCI-HEPC dataset with a forecasting horizon of 1-hour, EF-Transformer yields MAE = 0.091 kWh, RMSE = 0.134 kWh, MAPE = 2.87%, corresponding to 15.7% MAE reduction from Patch-TST and 54.0% reduction from a standard LSTM baseline. An ablation study confirms that each component in the architecture contributes to the overall performance, and among them, the adaptive decomposition gate provides the largest individual gain. Overall, EF-Transformer provides a scalable and interpretable foundation for smart grid operators who need accurate and high-resolution forecasts at both household and aggregated grid level, and therefore, future work will focus on extending the model to probabilistic forecasting, federated training over distributed smart meters, and incorporating renewable energy covariates for comprehensive smart grid energy management.

References

[1] Y. Wang, Q. Chen, T. Hong and C. Kang, "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges," in ***IEEE Transactions on Smart Grid***, vol. 10, no. 3, pp. 3125-3148, May 2019, doi: 10.1109/TSG.2018.2818167.

[2] A. Marinakis, M. Glavic and T. Van Cutsem, "Minimal Reduction of Unscheduled Flows for Security Restoration: Application to Phase Shifter Control," in ***IEEE Transactions on Power Systems***, vol. 25, no. 1, pp. 506-515, Feb. 2010, doi: 10.1109/TPWRS.2009.2030423.

[3] N. B. Badhe, R. P. Neve, V. P. Yele, S. Abhang, K. M. Dhule, and D. Mali, “An optimized system for predicting energy usage in smart grids using temporal fusion transformer and Aquila optimizer,” *Front. Artif. Intell.*, vol. 8, Art. no. 1542320, Apr. 2025, doi: 10.3389/frai.2025.1542320.

[4] Giacomazzi, Elena & Haag, Felix & Hopf, Konstantin. (2023). Short-Term Electricity Load Forecasting Using the Temporal Fusion Transformer: Effect of Grid Hierarchies and Data Sources. 353-360. 10.1145/3575813.3597345.

[5] C. Wu, J. Li, W. Liu, Y. He, and S. Wang, “Deep learning integration optimization of electric energy load forecasting and market price based on the ANN–LSTM–transformer method,” *Front. Energy Res.*, vol. 11, Art. no. 1292204, Nov. 2023, doi: 10.3389/fenrg.2023.1292204.

[6] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, “Short-term residential load forecasting based on LSTM recurrent neural network,” *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019, doi: 10.1109/TSG.2017.2753802.

[7] J. Chung and B. Jang, “Accurate prediction of electricity consumption using a hybrid CNN-LSTM model

based on multivariable data," *PLOS ONE*, vol. 17, no. 11, Art. no. e0278071, Nov. 2022, doi: 10.1371/journal.pone.0278071.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

[9] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, pp. 11106–11115, 2021, doi: 10.1609/aaai.v35i12.17325.

[10] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, pp. 22419–22430, 2021.

[11] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.

[12] I. Ahmad and M. A. Hafeez, "Energy load forecasting using a dual-stage attention-based recurrent neural network," *Appl. Sci.*, vol. 11, no. 21, Art. no. 10484, Nov. 2021, doi: 10.3390/app112110484.

[13] Y. Wang, W. Jiang, C. Wang, Q. Song, T. Zhang, Q. Dong, and X. Li, "An electricity load forecasting model based on multilayer dilated LSTM network and attention mechanism," *Front. Energy Res.*, vol. 11, Art. no. 1116465, May 2023, doi: 10.3389/fenrg.2023.1116465.

[14] P. Ran, K. Dong, X. Liu, and J. Wang, "Short-term load forecasting based on CEEMDAN and transformer," *Electr. Power Syst. Res.*, vol. 214, Art. no. 108885, 2023, doi: 10.1016/j.epsr.2022.108885.

[15] C. Xu and G. Chen, "Interpretable transformer-based model for probabilistic short-term forecasting of residential net load," *Int. J. Electr. Power Energy Syst.*, vol. 155, Art. no. 109515, 2024, doi: 10.1016/j.ijepes.2023.109515.

[16] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[17] P. H. Canh, N. Q. Minh, N. D. Tien, and T. T. Q. Anh, "Short-term electricity load forecasting based on temporal fusion transformer model," *IEEE Access*, vol. 10, pp. 106296–106304, 2022, doi: 10.1109/ACCESS.2022.3211941.

[18] C. Wu, J. Li, W. Liu, Y. He, and S. Wang, "Deep learning integration optimization of electric energy load forecasting and market price based on the ANN–LSTM–transformer method," *Front. Energy Res.*, vol. 11, Art. no. 1292204, 2023, doi: 10.3389/fenrg.2023.1292204.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[20] G. Hebrail and A. Berard, "Individual Household Electric Power Consumption Dataset," UCI Machine Learning Repository, 2012. [Online]. Available: **HTTPS://ARCHIVE.ICS.UCI.EDU/DATASET/235/INDIVIDUAL+HOUSEHOLD+ELECTRIC+POWER+CONSUMPTION**

[21] A. Trindade, "Electricity Load Diagrams 2011–2014 Dataset," UCI Machine Learning Repository, 2015. [Online]. Available: https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014

[22] Department of Energy and Climate Change (DECC), "Smart Meter Energy Consumption Data," UK Government Open Data, 2013. [Online]. Available: https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households