Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

An Attention-Enhanced Transformer Framework for Intelligent Energy Management and Load Forecasting in U.S. Power Grids

Rayhanul Islam Sony[1]✉ and K M Shihab Hossain[2]

[1]*College of Graduate Professional Studies, Trine University, One University Avenue, Angola, 46703, Indiana, USA*
[2]*College of Business Administration, Central Michigan University, 1200 S Franklin St, Mt Pleasant, 48859, Michigan, USA*
Corresponding Author: Rayhanul Islam Sony, **E-mail**: rsony23@my.trine.edu

| ABSTRACT

Accurate short-term and multi-horizon electricity load forecasting is a fundamental requirement for intelligent energy management in modern power grid systems, particularly under increasing demand variability and weather-driven consumption patterns. Conventional statistical, machine learning, and recurrent neural network models often exhibit limited capability in modelling complex non-linear relationships and long-range temporal dependencies inherent in large-scale power system data. To address these challenges, this paper proposes an Attention-Enhanced Transformer-based Multi-Horizon Weather-aware Network (ATMH-WNet) for efficient and accurate load forecasting in U.S. power grids. The proposed framework employs linear feature embedding and sinusoidal positional encoding to construct temporally informed latent representations, which are processed through a multi-layer Transformer encoder with multi-head self-attention. This design enables the model to jointly capture short-term dynamics and long-range dependencies while producing direct multi-step forecasts in a single forward pass, thereby avoiding recursive error accumulation. The proposed model is evaluated on the PJM Interconnection hourly electricity consumption dataset spanning 2002-2018 and is compared against persistence, SARIMA, Prophet, XGBoost, and LSTM benchmarks. Experimental results demonstrate that ATMH-WNet consistently outperforms all baseline models, achieving a mean absolute error of 1325 MW, a root mean squared error of 1873 MW, a mean absolute percentage error of 2.9%, and an $R2$ score of 0.97 on the held-out test set. Compared to the strongest deep learning baseline, the proposed framework reduces forecasting errors by more than 20% across major accuracy metrics. Additional qualitative analyses, including load profile alignment, residual diagnostics, and error distribution assessment, further confirm the robustness, stability, and generalization capability of the proposed approach. These results establish ATMH-WNet as an effective and scalable solution for real-world intelligent energy management and multi-horizon load forecasting applications.

| KEYWORDS

Persistence model, SARIMA, XGBoost, LSTM, Prophet, Time series forecasting, Baseline models, Forecasting benchmarks, Model evaluation, Predictive analytics

| ARTICLE INFORMATION

ACCEPTED: 12 December 2025 **PUBLISHED:** 09 January 2026 **DOI:** 10.32996/jcsts.2026.8.1.1

1. Introduction

Global energy industry is undergoing a radical transformation driven by the accelerated integration of renewable energy, extensive automotive electrification, and the expansion of distributed energy sources. These processes have introduced significant variability and unpredictability in electricity demand patterns, creating serious challenges for power system operators. Precise short-term load forecasting (STLF) in large-scale electricity markets, such as the regional transmission organizations (RTOs) in the United States (e.g., PJM Interconnection), has become a crucial tool to ensure grid reliability, optimize unit commitment and economic dispatch, enable effective demand response programs, and minimize operational costs [1] – [3].

PJM Interconnection serves more than 65 million people across 13 states and the District of Columbia. Such complex systems require accurate load forecasting to maintain stability during extreme weather, peak demand periods, and unexpected renewable intermittency [4]. Errors in load forecasting can cause inefficient resource allocation, higher reserve requirements, and potential blackouts, resulting in significant economic losses and reliability risks [5].

STLF has traditionally relied on statistical methods such as ARIMA, SARIMA, and exponential smoothing due to their interpretability and computational efficiency. However, these approaches are linear and fail to capture nonlinear and non-stationary behaviors in modern load profiles, which are influenced by meteorological, economic, and consumer behavior factors [6], [7]. Furthermore, these methods often require manual stationarity transformations and struggle to model multi-scale seasonality (daily, weekly, yearly) simultaneously [8].

Machine learning approaches, including support vector regression (SVR), random forests, and gradient boosting models such as XGBoost and LightGBM, offer improved performance by integrating exogenous variables like temperature, humidity, and calendar effects [9], [10]. These models handle nonlinear relationships and variable interactions well, but they generally assume independent samples, overlooking sequential dependencies essential for accurate forecasting [11].

Recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), can automatically learn hidden temporal patterns and address some of the limitations of statistical and classical machine learning methods [12]. While these models are superior in capturing nonlinear behaviors and short-term dependencies, they suffer from vanishing gradient issues with long sequences and high computational demands due to their sequential processing nature [13].

Transformer-based architectures, initially designed for natural language processing [14], have revolutionized time-series forecasting by using self-attention mechanisms to learn global dependencies efficiently. Self-attention dynamically emphasizes significant time steps, enabling parallel processing and improved performance on long sequences. Time-series variants such as Informer [15], Autoformer [16], and FEDformer [17] address quadratic complexity issues and employ decomposition techniques to extract trend and seasonal components, achieving state-of-the-art results in long-term forecasting.

Recent work has adapted transformers for electricity load forecasting. Giacomazzi et al. [1] investigated the Temporal Fusion Transformer (TFT) across various grid hierarchies, demonstrating its efficiency in integrating both static and dynamic covariates. Sievers et al. [2] proposed a federated transformer model for privacy-preserving load forecasting in smart grids. Hertel et al. [3] explored multiple training strategies for forecasting concurrent load time series using transformer models. Badhe et al. [18] developed a temporal fusion transformer combined with meta-heuristic optimization to enhance load prediction accuracy. Perçuku et al. [19] surveyed ML/DL models and highlighted the advantages of deep learning, particularly transformer-based models, in capturing complex time-dependent load patterns. Jain et al. [20] evaluated various machine learning approaches for electrical load forecasting, emphasizing the benefits of hybrid ML/DL frameworks.

Hybrid approaches have also emerged, integrating transformers with graph neural networks for spatial-temporal modeling [4], multi-frequency feature analysis, and decomposition-based methods to better extract temporal patterns [5]. Surveys by Dong et al. [5] illustrate the superiority of transformer-based models in handling multivariate inputs and complex temporal structures. Zhong et al. [12] and Dou et al. [13] demonstrated the effectiveness of combining deep learning architectures (ANN–LSTM–Transformer) and hybrid decomposition techniques to enhance forecasting accuracy.

Despite these advances, challenges remain in real-world U.S. grid applications. Most transformer-based models have been tested on European or Asian utilities, whose consumption patterns and weather differ from U.S. grids. There is a shortage of studies integrating high-resolution weather data with precise temporal alignment. Additionally, attention mechanisms in load forecasting require further research to support operational decision-making.

This paper addresses these gaps by proposing an **enhanced attention transformer architecture** tailored for short-term load forecasting in the PJM East region. The model incorporates multi-head self-attention for long-range dependencies, cyclic encoding to capture temporal periodicities (daily, weekly, yearly), and exogenous weather features (temperature, humidity, wind speed, precipitation). It processes 168-hour input sequences to generate 24-hour ahead forecasts, balancing computational efficiency and predictive performance.

The objectives of this study are:

- Develop a multivariate attention-enhanced transformer for U.S. grid load forecasting with integrated weather features.
- Evaluate the model against statistical, tree-based, RNN, and transformer baselines.
- Perform ablation studies to quantify contributions of attention mechanisms, cyclic encodings, weather features, and architecture.
- Provide interpretability insights via attention visualization and feature importance analysis.
- Demonstrate potential applications in intelligent energy management systems.

The contributions of this work include:

- Proposal of an attention-enhanced transformer architecture with cyclic encoding and weather integration, achieving state-of-the-art performance on PJM East hourly load data.
- Extensive experimental validation showing superior accuracy over multiple baselines.
- Systematic ablation analysis quantifying key components' impact on forecasting performance.
- Visualization of attention patterns and per-horizon performance for practical interpretation.
- Extension toward intelligent energy management in sustainable power systems.

The remainder of the paper is organized as follows: Section II reviews related work; Section III presents the proposed attention-enhanced transformer framework; Section IV describes experiments, results, and ablation studies; and Section V concludes with insights and future directions.

2. Literature Review

Recent developments in the smart grids, integration of renewable energy and electrified transportation systems have greatly enhanced the complexity in solving the energy forecasting and management issues. To achieve grid stability, optimize resource allocation, while encouraging sustainable energy transitions, accurate forecasting of the electricity load, renewable generation, and electric vehicle (EV) charging demand has become crucial. Conventional statistical methods usually fail to record nonlinear temporal correlations, excessive volatility, and multi-source interactions of the contemporary energy systems. Because of this trend, researchers have embraced deep learning, attention models, Transformers, hybrid models and ensemble approaches to become more accurate and robust in making predictions.

2.1 Transformer-Based Load and Electricity Price Forecasting

It has been recently discovered that Transformer-based architectures are especially successful at capturing long-range temporal dependencies present in electricity load and price time series. One of the first systematic enhancements of the standard Transformer for short-term load forecasting is proposed by Ahmad et al. [21], who introduce TFTformer with feature-specific load, weather, and temporal embeddings. In addition, the Transformer encoder is augmented with a Temporal Convolutional Network (TCN) to strengthen long-term dependency modeling. Experiments on datasets from Belgium, New Zealand, and five Australian states demonstrate that TFTformer outperforms traditional baselines with over 50% reduction in MSE, improves CARD performance by 42 points, and surpasses iFlowformer and iReformer by 1617 points, highlighting the effectiveness of feature-aware embedding and convolution–attention fusion.

While the work in [21] primarily focuses on load forecasting, Bâra and Oprea [22] apply Transformer-based models to day-ahead electricity price prediction in pan-European markets. Their approach addresses missing non-renewable generation and sold energy data through synthetic data generation and inverse optimization. The model achieves high predictive accuracy across markets in Romania, Spain, Poland, Finland, and the Czech Republic, with test R^2 values ranging from 0.95 to 0.98 and consistently low MAE. However, the authors acknowledge that the method incurs high computational cost due to complex feature engineering and optimization, in contrast to the structurally streamlined approach in [21].

To reduce architectural complexity while retaining attention mechanisms, Nguyen and Tran [31] propose a lightweight Transformer-MLP hybrid for short-term load forecasting. Their model employs a single-layer Transformer encoder with learnable positional encoding and an MLP decoder. Using univariate 30-minute interval data from Australian regions (NSW, QLD, VIC), the model achieves MAPE values between 0.69% and 0.95%, outperforming LSTM, CNN, and standalone MLP models. Nevertheless, the simplified design limits the method to univariate, single-step forecasting, reflecting a trade-off between efficiency and generalization.

Beyond numerical inputs, Hasan et al. [35] extend Transformer-based forecasting by incorporating contextual semantic information from textual news data. Their TSB-Forecast model integrates Time2Vec temporal embeddings, SBERT-based semantic feature extraction, and an ensemble of XGBoost and Extra Trees regressors. Compared with purely numerical Transformer-based models, TSB-Forecast achieves 38.7% lower MAE, 18.2% lower RMSE, and 50.6% lower SMAPE, demonstrating that semantic context can substantially improve forecasting accuracy at the cost of increased data dependency and system complexity.

2.2 Hybrid Deep Learning Models for Load and Consumption Forecasting

With the growing adoption of Transformer architectures, several studies explore hybrid deep learning models to better capture nonlinear temporal dynamics and multi-source interactions. He et al. [23] compare LSTM, Bi-LSTM, and a hybrid Transformer-BiLSTM model for wind and photovoltaic (PV) power forecasting. Their results show that Bi-LSTM mitigates time-lag and bias issues inherent in standard LSTM, while the Transformer–BiLSTM hybrid achieves the best performance, improving forecasting accuracy by 19% for wind power and 35% for PV power relative to Bi-LSTM. This highlights the effectiveness of combining global attention with recurrent temporal modeling to handle extreme variations.

Shifting focus to data fusion, Özen [24] proposes a CNN–LSTM–FFNN hybrid model for electricity consumption forecasting using hourly datasets from Chicago, Pittsburgh, and IHEC (Paris). The model achieves a mean RMSE of 0.0732, outperforming CNN, CNN–LSTM, LSTM–LSTM, and naive Transformer baselines. Unlike [23], which emphasizes temporal depth, this study demonstrates that integrating spatial feature extraction with nonlinear fusion layers can significantly enhance forecasting accuracy, particularly under data-scarce conditions.

Extending this work, Özen et al. [25] develop an ensemble-level hybrid framework that combines univariate deep learning models, multivariate CNN-LSTM networks, and a linear regression fusion layer. On the Chicago dataset, the framework achieves an RMSE of 0.0871, outperforming ARIMA, Random Forest, CNN, and LSTM models. Furthermore, introducing a Transformer–Gaussian Process hybrid reduces RMSE to 0.0768, emphasizing the benefit of probabilistic modeling over purely deterministic deep learning approaches, albeit with increased computational complexity.

Feature engineering also plays a crucial role in hybrid pipelines. Du et al. [29] propose the MCPO–VMD–FDFE framework, which incorporates signal decomposition, frequency-domain feature enhancement, and an improved PatchTST model. Evaluated on weekday, Saturday, and Sunday load profiles, the method achieves a cumulative RMSE reduction of 45.65 compared to baseline models. This contrasts with the data-fusion strategies in [24], [25], underscoring the importance of multi-stage preprocessing and decomposition.

Beyond forecasting accuracy, Mushref et al. [30] integrate prediction with grid control by introducing a hybrid ANFIS-Transformer framework optimized using Enhanced HawkFish Optimization. Using real-time load data from Kaggle, the model achieves a load RMSE of 4.15 kW, voltage RMSE of 1.24 V, and MAPE between 1.50% and 3.10%. Smart-grid control experiments further demonstrate a 54.4% reduction in energy loss, distinguishing this work from purely predictive studies [23], [25], [29].

2.3 Graph-Based and Spatio-Temporal Forecasting Models

Although hybrid models improve temporal modeling, they often neglect spatial dependencies inherent in power systems. To address this, Zhu et al. [27] propose the Spatial-Temporal Dynamic Graph Transformer (SDGT), which combines VMD-based periodic decomposition with a dynamic spatio-temporal correlation graph. Experiments on Australian and Tetouan (Morocco) datasets show MAE reductions of 38%-49% and RMSE reductions of 33%-45% relative to baseline models, demonstrating the effectiveness of explicitly modeling evolving spatial correlations.

Complementing this approach, Orji et al. [28] introduce a GAT–LSTM model that integrates Graph Attention Networks with LSTM through early feature fusion. Applied to the Brazilian electricity system, the model reduces MAE by 21.8%, RMSE by 15.9%, and MAPE by 20.2%. Compared with SDGT [27], this method employs a simpler graph structure but leverages grid topology and edge attributes, making it suitable for well-defined power networks.

Beyond graph-based approaches, alternative spatial generalization strategies have also been explored. Shape clustering combined with domain-adversarial transfer learning has been shown to improve residential load forecasting performance under distribution shifts [42], while diffusion-based generative attention models further enhance short-term residential load prediction by capturing uncertainty and complex temporal patterns [43].

At the building scale, Cao [47] proposes TDAGNN, which integrates temporal decomposition, multi-head interactive attention, and self-scaling diffusion graph neural networks. Using the BIM-SHMC dataset, the model achieves superior performance with an average improvement of 13.3% over STGCN, demonstrating the ability of graph-based learning to capture abrupt phase changes in high-rise buildings.

To improve generalization across datasets, Xiao et al. [49] propose a hierarchical attention-based model that combines SENet-enhanced TCNs with BiGRU and global attention. The model exhibits strong robustness, with cross-dataset R^2 standard deviation below 3.7%, emphasizing hierarchical feature alignment rather than explicit grid topology modeling.

2.4 Electric Vehicle Charging Demand Forecasting

With the rapid expansion of electric vehicle (EV) adoption, accurate EV charging demand forecasting has become increasingly important. Hussain et al. [26] propose a hybrid Transformer-LSTM model for medium- and long-term EV charging demand forecasting using the ACN datasets from Caltech and JPL. The model achieves MAE and MSE reductions of 17.27% and 19.79% on the Caltech dataset, and 24.91% and 23.17% on the JPL dataset for a 30-day horizon, demonstrating the effectiveness of hybrid temporal architectures in modeling long-term EV usage patterns.

To provide systematic benchmarking across forecasting horizons, Ahmadian and Gadh [32] evaluate statistical, machine learning, and deep learning models using one million 15-minute EV charging records. Their results show that tree-based and SARIMA models perform well for very short horizons, while LSTM and attention-based models excel at longer horizons, offering practical guidance for model selection.

At the station level, Singh et al. [33] propose an Attention-Augmented LSTM (AA-LSTM) model, achieving a MAPE of 3.90% and MSE of 0.40, outperforming standard LSTM and RNN models. Compared to [26] and [32], this work emphasizes interpretability and computational efficiency, demonstrating that competitive performance can be achieved without overly complex architectures.

2.5 Renewable Energy Forecasting

Beyond load forecasting, deep learning models are widely applied to renewable energy prediction. Zaman et al. [38] propose the Federated Temporal Dense Granular Transformer (FTDGT) for wind power forecasting, emphasizing privacy preservation and robustness. Compared with competing models, FTDGT improves RMSE by 12%, MAE by 15.5%, and R^2 by 9.2% across diverse datasets.

For short-term photovoltaic (PV) power forecasting, Ait Chaoui et al. [39] introduce a Wavelet-Transformer-TC-GRU hybrid model. Using 95,885 five-minute PV records, the model achieves MAE of 209.36, RMSE of 616.53, and R^2=0.96884, significantly outperforming LSTM, GRU, and CNN–LSTM baselines. This work highlights the importance of multi-resolution signal decomposition for high-frequency forecasting.

To enhance interpretability, Siddiqa et al. [45] propose SolarTrans, a two-stage framework combining Transformers with large language models. The model achieves MAE values between 0.0782 and 0.1544, RMSE values between 0.1760 and 0.4424, and R^2 up to 0.9692, while the explanation module attains ROUGE-1 of 0.7889 and BLEU of 0.6558, addressing transparency challenges in renewable energy forecasting.

Additionally, attention-enhanced Seq2Seq transfer learning has been successfully applied to climate-adaptive building energy forecasting, demonstrating strong generalization across varying environmental conditions [34].

2.6 Smart Grid Optimization, Fault Detection, and Advanced Applications

Beyond forecasting, deep learning techniques have been integrated into broader smart grid optimization tasks. Yu et al. [40] propose an ensemble of attention-enhanced N-BEATS and XGBoost for district heating load forecasting, achieving an RMSE of 0.6427 and R^2=0.9664.

Hussain et al. [44] introduce the FireNet–XGBoost hybrid model for mid-term building load forecasting, achieving RMSE of 18.71 and R^2=0.9334

Transformer-based attention mechanisms have also been applied to fault detection and self-healing smart grid networks. Dubey et al. [36] demonstrate that attention-driven Transformers improve anomaly localization and adaptive recovery in power systems.

Alternative attention-based recurrent architectures, such as GRU-attention models [37] and GA-LSTM frameworks [46], have also shown competitive performance in very short-term residential load forecasting, offering efficient alternatives to Transformer-heavy architectures.

Finally, Shen et al. [50] propose GridSense, a large language model-based situational awareness framework for smart grids. The system achieves a load forecasting RMSE of 0.15 and anomaly detection accuracy of 90.25% even with limited training data, highlighting the future potential of LLM-driven intelligence in smart grid applications.

Table 1. Summary of Recent Deep Learning-Based Energy Forecasting Studies

Year	Ref.	Model Name	Key Results	Limitations
2025	[21]	TFTformer	>50% MSE reduction vs. baselines; 42% over CARD; 16–17% over iFlowformer/iReformer	High computational complexity; needs high-quality auxiliary data
2025	[22]	Transformer + Synthetic Inputs	R^2 = 0.95–0.98 (test); 0.91–0.97 (evaluation)	Complex feature engineering and inverse optimization
2025	[23]	Transformer-BiLSTM	19% improvement (wind); 35% improvement (PV)	High computation cost; fusion challenges
2023	[24]	CNN-LSTM-FFNN	Average RMSE = 0.0732	Dependence on correlated weather data
2025	[25]	Hybrid CNN-LSTM + LR	RMSE = 0.0871; improved to 0.0768	Increased model complexity
2025	[26]	Hybrid Transformer	MAE ↓ 17.27–24.91%; MSE ↓ 19.79–23.17%	Limited scalability
2025	[27]	SDGT	MAE ↓ 38–49%; RMSE ↓ 33–45%	High computational cost
2025	[28]	GAT-LSTM	MAE ↓ 21.8%; RMSE ↓ 15.9%; MAPE ↓ 20.2%	Requires detailed grid topology
2025	[29]	MCPO-VMD-FDFE	Overall RMSE reduction = 45.65%	Multi-stage decomposition complexity
2025	[30]	ANFIS-Transformer	Load RMSE = 4.15 kW; Voltage RMSE = 1.24 V	Optimization overhead
2025	[31]	Transformer-MLP	MAPE = 0.69–0.95%	Univariate, single-step forecasting
2025	[32]	Benchmark Study	MAE = 0.23–0.46 kW; RMSE = 0.46–1.20 kW	Horizon-dependent performance
2025	[33]	AA-LSTM	MAPE = 3.90%; MSE = 0.40	Limited EVCS diversity
2025	[34]	Attention Seq2Seq	Accuracy = 96.2%; R^2 = 0.98	Long-term data requirement
2025	[35]	TSB-Forecast	MAE ↓ 38.7%; RMSE ↓ 18.2%	Depends on news/text data
2025	[36]	AACNN-Transformer	Accuracy up to 97.14%	Fault-detection focused
2025	[37]	GRU-Attention	MAPE = 0.77%	Very short-term horizon
2025	[38]	FTDGT	RMSE ↓ 12%; MAE ↓ 15.5%	Federated communication overhead
2025	[39]	WT-Transformer Hybrid	RMSE = 616.53; R^2 = 0.9688	Site-specific validation
2025	[40]	N-BEATS + XGBoost	RMSE = 0.6427; R^2 = 0.9664	Ensemble tuning required
2025	[41]	ORA-DL	Demand accuracy = 93.38%	System complexity
2024	[42]	AT-Seq2Seq	MAPE < 2%	Domain similarity dependence
2025	[43]	Diffusion + Attention	MAE ↓ 47.4%; MAPE ↓ 57.6%	High training cost
2025	[44]	FireNet-XGBoost	RMSE = 18.71; R^2 = 0.9334	Single-building study
2025	[45]	SolarTrans	RMSE = 0.176–0.442	Limited evaluation duration
2024	[46]	GA-LSTM	RMSE = 0.6056	Genetic optimization overhead
2025	[47]	TDAGNN	Average improvement = 13.3%	Requires building topology
2025	[48]	MTCAT	RMSE ↓ 30–50%	Domain-specific
2025	[49]	TSEBG	Cross-dataset R^2 std = 3.7%	Complex hierarchy
2025	[50]	GridSense	RMSE = 0.15; anomaly accuracy = 90.25%	High LLM inference cost

3. Methodology

The general process used in this research will be based on a systematic pipeline of a multi-horizon weather-sensitive prediction, as shown in **Figure-1** overall methodology. The steps it takes are to prepare the multivariate time-series data that is initially preprocessed and further divided into three disjoint data subsets to have the fair and unbiased evaluation. In particular, the dataset is segmented into training, validation, and test set in 80%, 5%, and 15% proportions, respectively. The model learning is done on the training set, the convergence and the overfitting is observed and controlled using the validation set and in the end the final performance on unseen data is evaluated using test set.

Given the data splitting, the forecast forthcoming is a issue of a supervised multi-horizon prediction problem, where UPTs are assigned to each other fixed-length historical input windows and future goal sequences. In order to draw comparative standards, a few benchmark models are taken into account along with the proposed approach. These are persistence-based forecasting, classical statistical forecasting models like SARIMA and Prophet, machine learning forecasting models like XGBoost, and a recurrent neural network forecasting model LSTM. The baselines are a wide array of forecasting paradigms and serve to act as benchmarks against which the proposed architecture performance can be measured.

The essence of the methodology is the suggested ATMH-WNet (Attention-based Transformer to Multi-Horizon Weather-aware Forecasting) framework that includes the sequential processing of the input sequences in the sequence of embedding, attention-based temporal encoding, and direct output mapping stages. The model uses first temporal enriching of raw inputs, then contextual dependencies in time using a Transformer encoder, then direct multi-horizon forecast in the single forward pass. This end-to-end model enables the model to incorporate short-term dynamics and long-range temporal dynamics without using recursive prediction strategies.

The performance of the models is measured according to the standard regression measures, such as R^2, RMSE, MAE, MAPE, and MSE, which are calculated on the test data. Collectively, this approach methodology guarantees a systematic comparison of the suggested ATMH-WNet and the available forecasting baselines at the same data splits and assessment standards.

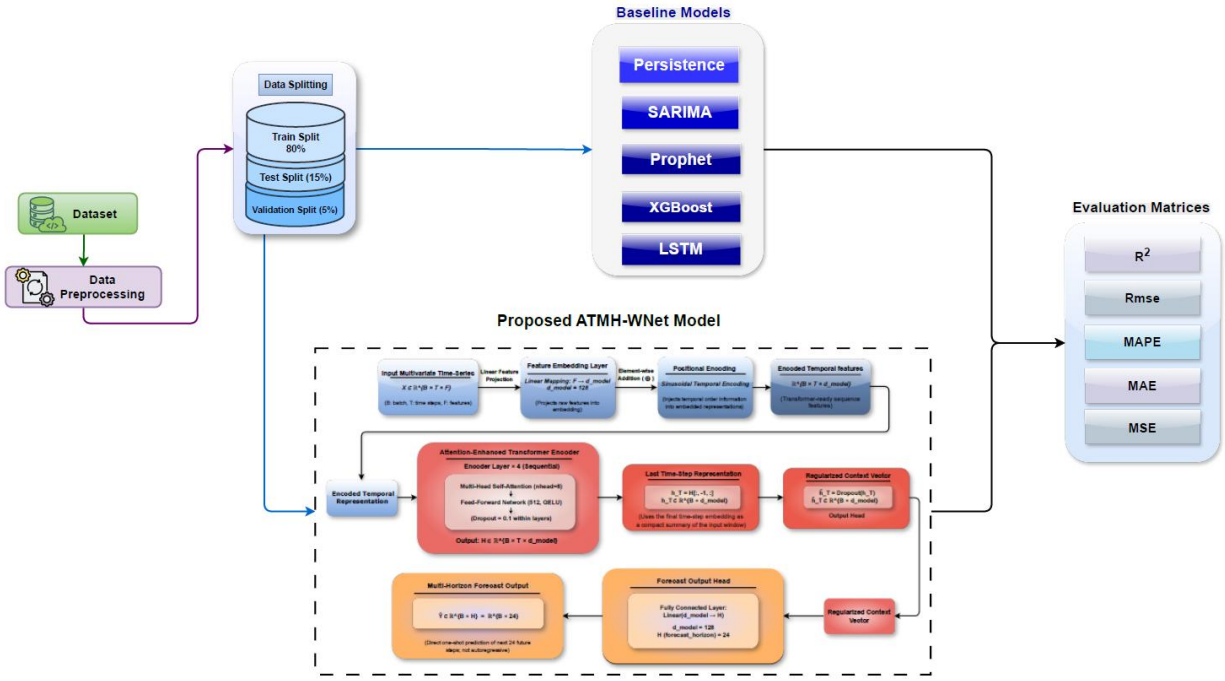

Figure 1: Overall methodological framework of the proposed study, illustrating the data preprocessing and splitting strategy, baseline forecasting models, the proposed ATMH-WNet architecture for multi-horizon weather-aware forecasting, and the evaluation pipeline

3.1 Dataset Description

The empirical foundation of this study is built upon the Hourly Energy Consumption dataset [51] sourced from the PJM Interconnection, which represents a massive regional transmission organization in the United States. This dataset provides a high-fidelity longitudinal record of power consumption, encompassing a total of 145,366 hourly observations. The temporal scope of the data is extensive, spanning from January 1, 2002 to August 3, 2018 thereby capturing diverse seasonal cycles, holiday variations, and long-term economic shifts. By utilizing such a robust and high-resolution time-series repository, the research ensures that the model is exposed to the complex, non-linear fluctuations inherent in large-scale power grid operations, which is critical for validating the robustness of the Attention-Enhanced Transformer framework.

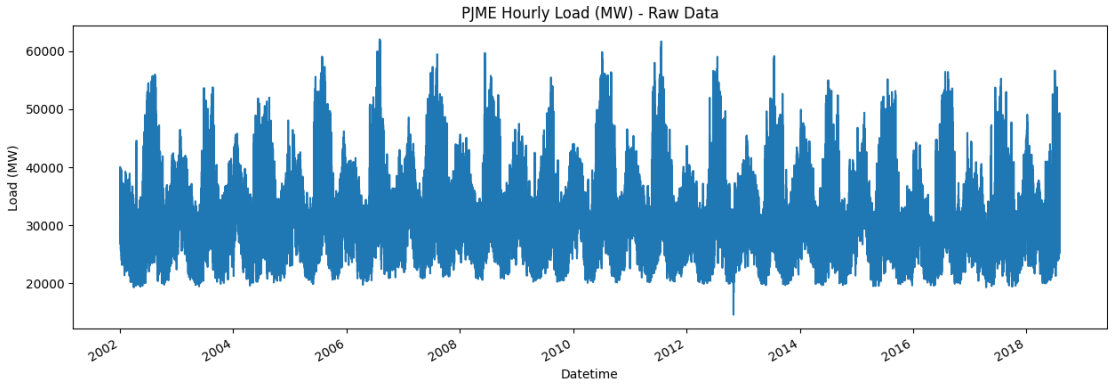

Figure 2: Sample of Raw Dataset

3.2 Preprocessing Pipeline

This study employs an extensive preprocessing pipeline to transform raw hourly electricity demand data into a structured multivariate time-series representation suitable for attention-based transformer architectures. The preprocessing workflow integrates temporal encoding, exogenous weather variables, normalization, and supervised sequence construction to enhance forecasting performance and model interpretability.

3.2.1 Temporal Feature Engineering

Electricity demand exhibits strong daily, weekly, and seasonal periodicity. To capture these recurring temporal patterns while avoiding artificial discontinuities, cyclic encodings were applied to time-related variables derived from the timestamp. Specifically, hour-of-day, day-of-week, and month-of-year were transformed using sine and cosine mappings as follows:

$$hour_{\sin} = sin(2\pi \frac{h}{24}), \quad hour_{\cos} = cos(2\pi \frac{h}{24}), \tag{1}$$

$$day_{\sin} = sin(2\pi \frac{d}{7}), \quad day_{\cos} = cos(2\pi \frac{d}{7}), \tag{2}$$

$$month_{\sin} = sin(2\pi \frac{m}{12}), \quad month_{\cos} = cos(2\pi \frac{m}{12}), \tag{3}$$

where h, d, and m denote hour, day of week, and month, respectively. This representation preserves circular continuity (e.g., hour 23 to hour 0) and is particularly well-suited for transformer attention mechanisms.

Additionally, binary indicators for weekends and U.S. federal holidays were incorporated to capture systematic demand reductions during non-working days.

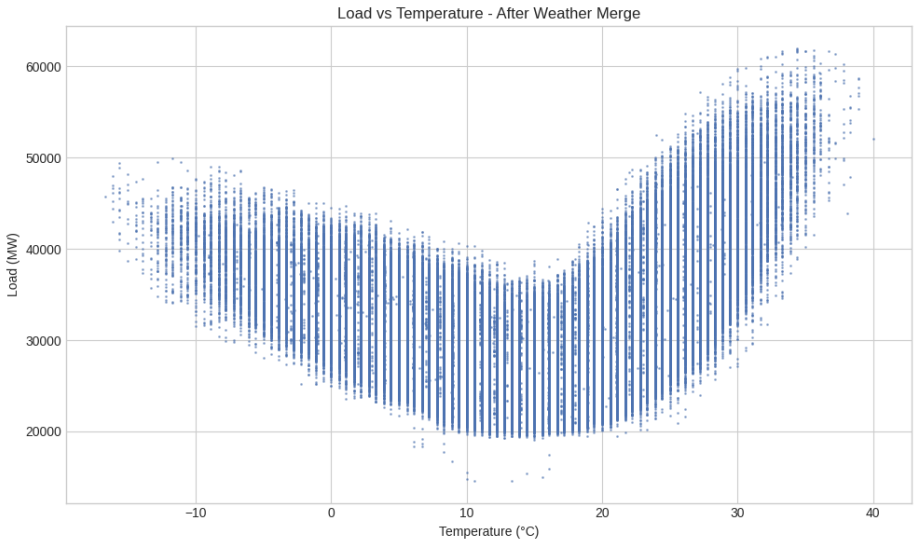

Figure 3: Distribution electricity load vs temperature data

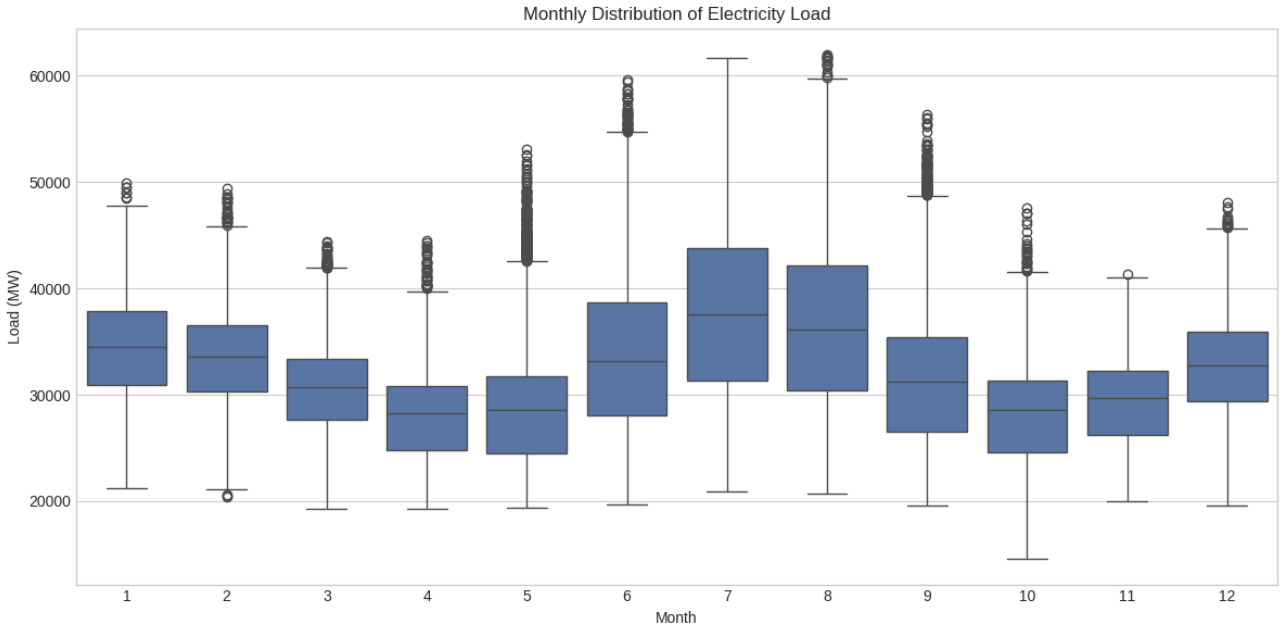

Figure 4: Monthly box plot distribution

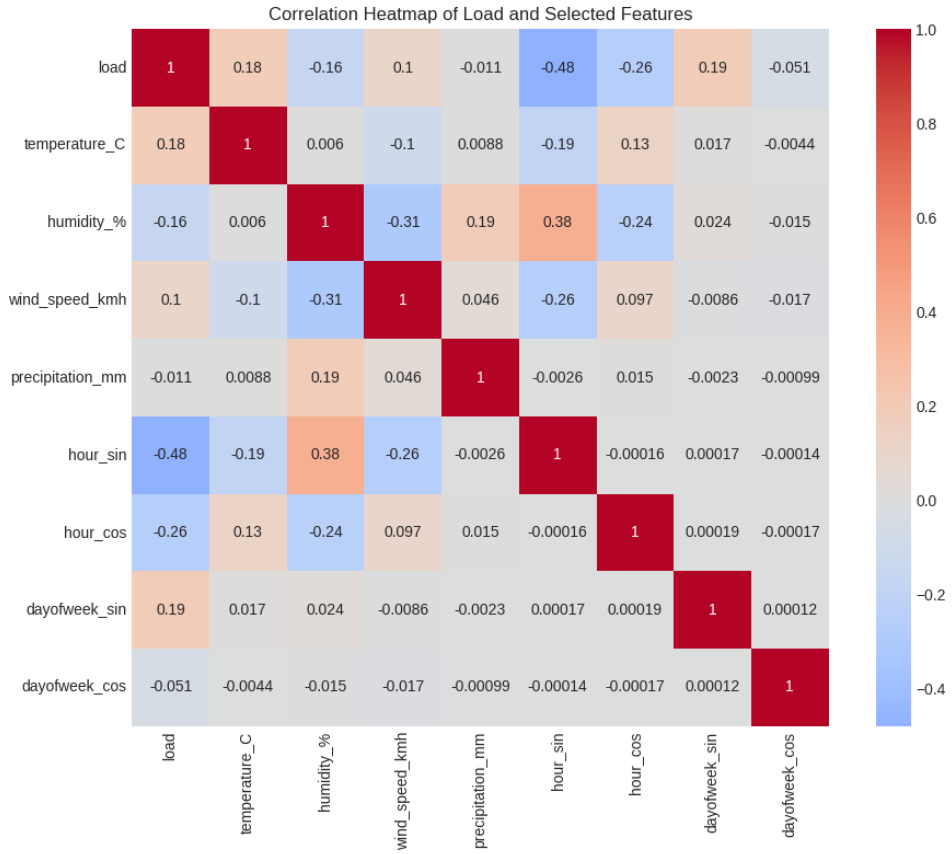

Figure 5: Representation of relation matrix heatmap

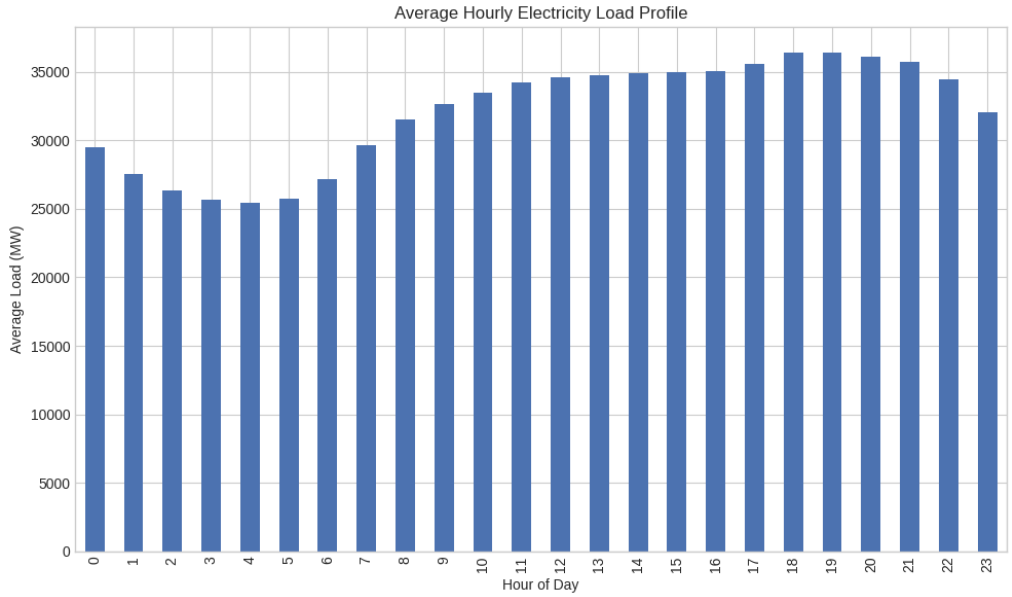

Figure 6: Average hourly electricity load profile

3.2.2. Exploratory Analysis and Feature Relationships

Figure 3 illustrates the nonlinear relationship between electricity demand and ambient temperature, revealing a pronounced U-shaped dependency driven by heating and cooling loads. Seasonal variations in demand are further evidenced by the monthly box plots shown in **Figure 4**. Correlation analysis (**Figure 5**) confirms that temperature and cyclic temporal features exhibit stronger associations with load compared to other meteorological variables, justifying their inclusion in the forecasting model. **Figure 6** presents the bar chart of hourly electricity load profile.

3.2.3. Integration of Exogenous Weather Variables

Meteorological conditions play a critical role in electricity consumption due to heating and cooling demands. To incorporate environmental influences, hourly weather observations were retrieved from the Philadelphia International Airport station, which serves as a representative location for the PJM East region. The selected exogenous variables include temperature (◦C), relative humidity (%), wind speed (km/h), and precipitation (mm). **Figure 7** demonstrated the time series analysis electricity load and temperature after merging with weather data.

Weather data were temporally aligned with the load series using a left join on the hourly timestamp. Minor gaps were observed in the meteorological records and were addressed using linear interpolation for temperature, humidity, and wind speed, while precipitation was zero-filled to represent the absence of rainfall. After preprocessing, the merged dataset contains no missing values. **Figure 8** presents the distribution of electricity load and temperature data after merging with weather data. Also, **Figure 9** represents the correlation heatmap of merged dataset.

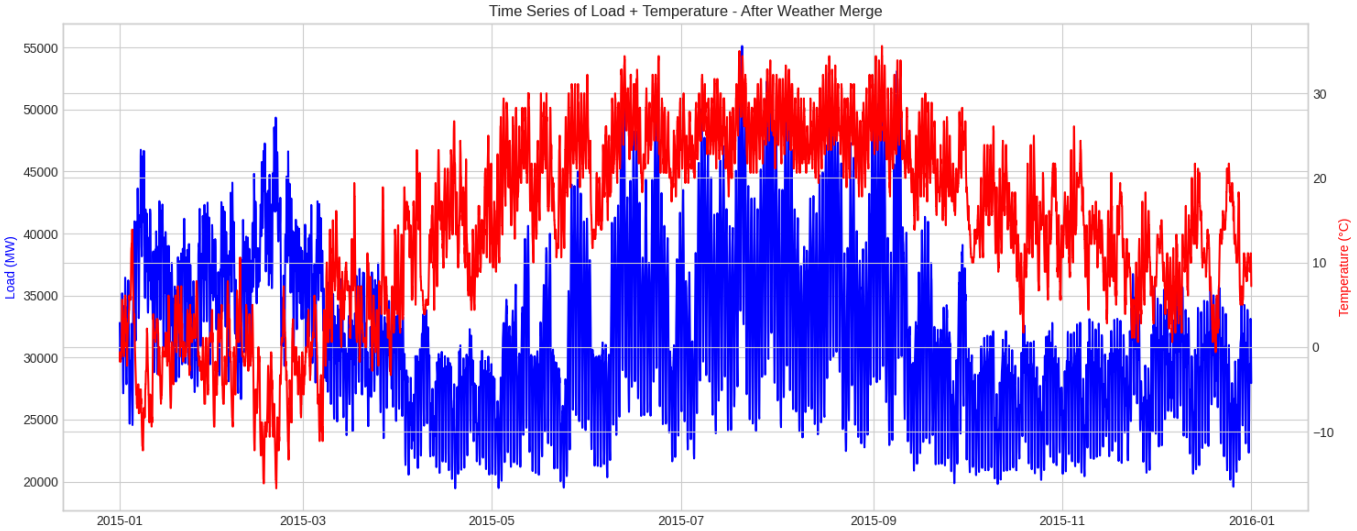

Figure 7: Time series distribution of load and temperature after weather data merged

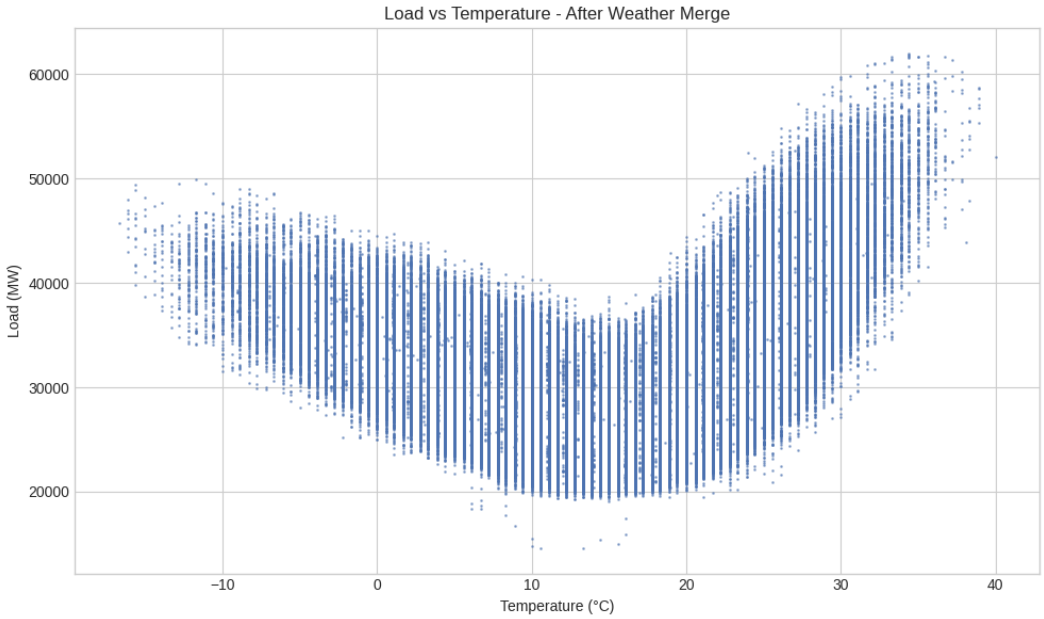

Figure 8: Distribution electricity load vs temperature data after merging with weather data

3.2.4. Feature Scaling

To ensure numerical stability during training and to prevent dominance by features with larger magnitudes, Min-Max normalization was applied independently to the load and weather variables:

$$x_{scaled} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad (4)$$

All scaled variables were constrained to the range [0, 1]. Separate scalars were preserved to allow inverse transformation during post-processing and evaluation.

3.2.5. Supervised Sequence Construction

The multivariate time series was reformulated into a supervised learning problem using a sliding window strategy. Each input sequence consists of the past 168 hours (7 days) of observations, capturing both short-term and weekly consumption patterns. The forecasting horizon was set to 24 hours, enabling day-ahead load prediction.

Formally, given a feature matrix $X_t \in R^{168 \times F}$ at time t, where F denotes the number of input features, the model learns to predict the future load vector:

$$y_t = [L_t + 1, L_t + 2, \ldots, L_t + 24], \quad (5)$$

where L_t represents the scaled electricity load. This process resulted in 145,175 training samples, each with a dimensionality of 168×13 features per timestep.

Figure 8: Distribution electricity load vs temperature data after merging with weather data

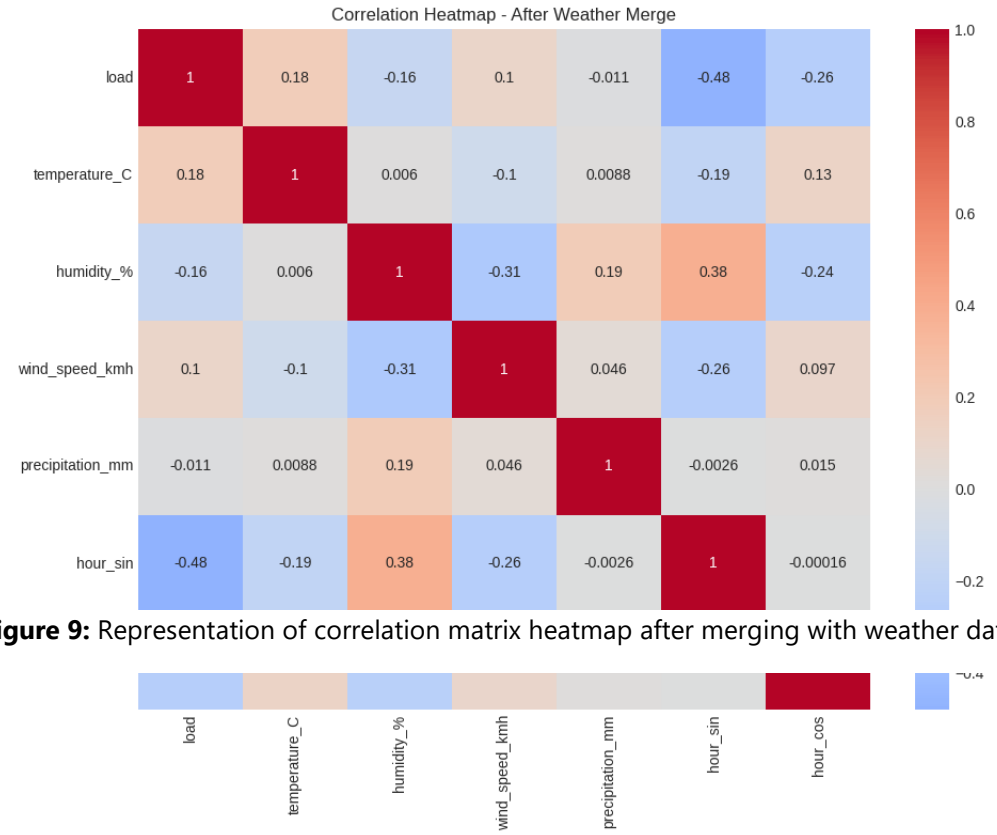

Figure 9: Representation of correlation matrix heatmap after merging with weather data

Table 2
Comparison of Dataset Characteristics Before and After Weather Integration

Aspect	Before Integration	After Integration
Number of Records	145,366	145,366
Input Features per Timestep	~8–10	~13–16
Exogenous Variables	None	Temperature, Humidity, Wind, Precipitation
Missing Values	0	0 (after interpolation)
Feature Scaling	Load only	Load + Weather
Supervised Sequences	Not applicable	145,175 samples

Table 2 summarizes the key characteristics of the dataset before and after weather integration and feature engineering. The preprocessing framework produces a rich multivariate representation that captures temporal periodicity, environmental effects, and long-range dependencies, providing a robust foundation for attention-enhanced transformer-based load forecasting.

3.3. Dataset Splitting

The processed dataset was partitioned into three distinct subsets such as training, validation, and testing using a chronological splitting approach to preserve the temporal dependencies inherent in energy load forecasting. This sequential division ensures that the Attention-Enhanced Transformer is evaluated on its ability to generalize to unseen future time steps, simulating real-world grid management scenarios. The primary training corpus consists of data from 2002 to 2014, comprising 113,735 samples, which allows the model to capture over a decade of multi-seasonal patterns and long-term demand trends.

To fine-tune the model's hyperparameters and prevent overfitting, the year 2015 was designated as the validation set, providing 8,569 distinct hourly observations. Finally, the model's predictive performance and robustness were rigorously evaluated using a dedicated test set spanning from 2016 to 2018, which includes 22,489 samples. This out-of-sample testing phase is critical for verifying the framework's reliability across varying economic and climatic conditions observed in the latter years of the study period. The comprehensive distribution of the data samples is summarized in Table 3.

Table 3
Distribution of Dataset Samples and Chronological Splitting Strategy

Partition	Year Range	Number of Samples	Percentage (%)
Training Set	2002 – 2014	113,735	78.55%
Validation Set	2015 – 2015	8,569	5.92%
Test Set	2016 – 2018	22,489	15.53%
Total	**2002 – 2018**	**144,793**	**100.00%**

3.4. Proposed ATMH-WNet Hybrid Model

The work proposed a deep learning model, ATMH-WNet (Attention-based Transformer to Multi-Horizon Weather- aware Forecasting), which is designed to train on time-series inputs of multivariate data to learn meaningful time-resolved representations, and to generate explicit multi-step predictions end-to-end. The general ATMH-WNet operational structure is outlined in Figure- 10, and it focuses on the way the model would convert a historical input window into a compressed contextual representation of the data and then project it to the required forecast horizon. The proposed design is following clear sequential pipeline so that the intermediate representations are still interpretable and the data flow is clearly traceable from the input to the output.

The work proposed a deep learning model, ATMH-WNet (Attention-based Transformer to Multi-Horizon Weather- aware Forecasting), which is designed to train on time-series inputs of multivariate data to learn meaningful time-resolved representations, and to generate explicit multi-step predictions end-to-end. The general ATMH-WNet operational structure is outlined in Figure- 10, and it focuses on the way the model would convert a historical input window into a compressed contextual representation of the data and then project it to the required forecast horizon. The proposed design is following clear sequential pipeline so that the intermediate representations are still interpretable and the data flow is clearly traceable from the input to the output.

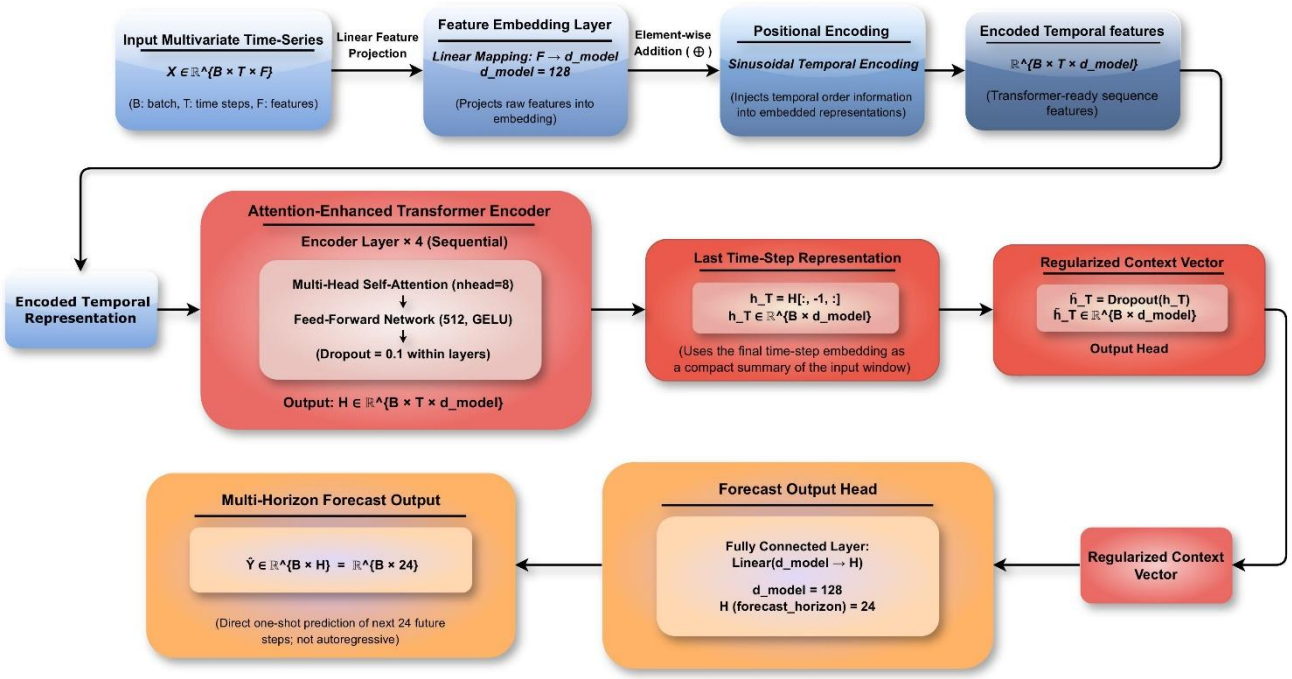

Figure 10: Overview of the proposed ATMH-WNet architecture for attention-based multi-horizon weather-aware forecasting, showing the sequential pipeline from input embedding and positional encoding to Transformer-based temporal representation learning and the final multi-horizon prediction head.

The ATMH-WNet consists of three major parts which are working in series. First, the raw input sequence is first transformed into a Transformer-compatible latent representation using a linear feature embedding layer and the temporal order information is added using sinusoidal positional encoding. This phase is to make sure that the model is able to handle heterogeneous sets of features as well as maintain time-step sequence required in the temporal modeling. Second, the embeddings with the encoded temporal information are fed through an attention-enhanced Transformer encoder, a stack of overlaid encoder layers, in which multi-head self-attention records relationship between various time steps and position-wise feed-forward sublayers generalize the results of the learned representations. This means, the encoder attends different of the parts of the historical window and builds contextualized sequence features that reflect both the short-term variations as well as the longer-range dependency on the input. Lastly, ATMH-WNet uses a small temporal aggregation mechanism by choosing the last time-step encoding of the coded sequence as a summative context vector, uses dropout regularization, and enters the ensuing vector into a small output head. The output head is a linear projection that can produce the direct multi-horizon forecast using a single forward pass and thus efficient inference without being required to undergo recursive prediction processes.

3.4.1. Input Embedding and Positional Encoding

The initial phase of ATMH-WNet, as shown in Figure- 11, will be aimed at transforming the raw multivariate time- series observations into a format that is temporal in nature and can be highly processed by an attention-based encoder. Since the behavior of the proposed model is implemented in the space of latent features-non-rational spaces that are of fixed dimension-that feature space, this stage has two complementary functions, projecting in a fixed-dimensional space the heterogeneous input variables, and explicitly encoding the temporal order of observations. Together, these operations can ensure that the next-stage attention mechanism can jointly reason about the interaction of features without any handcrafted temporal assumptions and temporal patterns.

Let the historical input sequence be represented as

$$X \in \mathbb{R}B \times T \times F\,, \tag{1}$$

where B denotes the batch size, T the length of the input window, and F the number of observed variables. To align the input with the Transformer architecture, each feature vector is first mapped into a higher-dimensional embedding space using a linear projection. This embedding operation is formulated as,

$$E = XWe + be, \tag{2}$$

where $W_e \in \mathbb{R}^{F \times d_{model}}$ and $b_e \in \mathbb{R}^{d_{model}}$ are learnable parameters, and $d$model denotes the embedding dimension. This transformation enables the model to represent all input variables within a unified latent space while preserving the original temporal resolution.

However, linear embedding alone does not convey any information about the ordering of time steps. To address this limitation, ATMH-WNet incorporates sinusoidal positional encoding, which injects deterministic temporal information into the embedded sequence. For a given time index t and embedding dimension i, the positional encoding is defined as,

$$PE_{t,2i} = \sin\left(\frac{t}{10000^{\frac{2i}{d_{modal}}}}\right),\ PE_{t,2i+1} = \cos\left(\frac{t}{10000^{\frac{2i}{d_{modal}}}}\right), \tag{3}$$

The positional encoding is then combined with the embedded features through element-wise addition,

$$Z = E + PE, \tag{4}$$

resulting in a temporally aware representation $Z \in \mathbb{R}^{B \times T \times d_{model}}$. This representation preserves both feature-level information and temporal ordering, allowing the attention-based encoder to model dependencies across the entire input window. The encoded sequence Z thus forms the input to the Transformer encoder and provides the foundation for learning contextual temporal representations in the subsequent stage of ATMH-WNet.

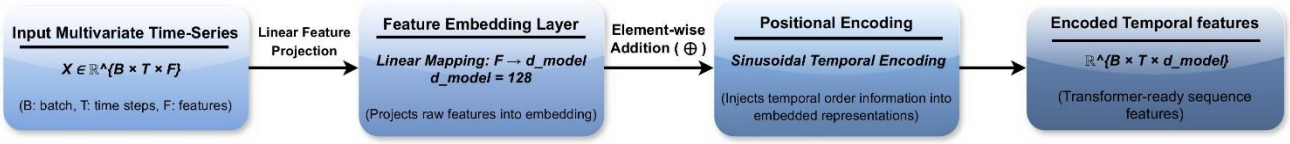

Figure 11: Input embedding and sinusoidal positional encoding module of ATMH-WNet, where multivariate time- series inputs are projected into a d_{model} dimensional latent space and augmented with temporal position information to form Transformer-ready representations.

3.4.2. Attention-Enhanced Transformer Encoder

Following the input embedding and positional encoding phase, ATMH-WNet uses an attention-enhanced Trans- former encoder to acquire the contextualized temporal representations on the input sequence encoded during the input embedding and positional encoding phase. As shown in Figure- 12, this component serves as the building blocks of fitting (temporal modeling) within the proposed architecture as it helps the model to capture the short term dynamics and also the long range dependencies inside the historical window. Whereas from recurrent structures, the Encoder of Transformer will simply take the entire sequence as inputs to build the model so that each time step can selectively attend to every other position in the sequence to build up a globally informed temporal representation.

Let the temporally encoded input sequence obtained from the previous stage be denoted as,

$$Z \in \mathbb{R}^{B \times T \times d_{model}}, \tag{5}$$

which is fed into a stack of L Transformer encoder layers. Within each encoder layer, the input is first projected into query, key, and value spaces as

$$Q = ZW_Q, K = ZW_K,\ V = ZW_V, \tag{6}$$

where W_Q, W_K, and W_V are learnable projection matrices. The scaled dot-product self-attention mechanism is then computed as

$$SA(Z) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{7}$$

giving the model the freedom to give a contribution to any time step depending on its relevance to others. In order to increase the representational capacity, ATMH-WNet uses a multi-head attention formulation, in which several self- attention heads are calculated simultaneously and stacked together as,

$$MHSA(Z) = Concat(head1, \dots, headH)W_O, \tag{8}$$

with W_O denoting the output projection matrix.

The result of the multi-head self-attention block is then optimized by a position-wise feed forward-network as specified below

$$FFN(h) = GELU(hW_1 + b_1)W_2 + b_2, \tag{9}$$

where GELU is used as the activation function. By adding several such encoder layers, the model performs successive refinements on the part of the temporal representation - leading to an encoded sequence

$$H \in \mathbb{R}^{B\times T\times d_{model}}, \tag{10}$$

that captures contextual information across the entire input window.

ATMH-WNet uses temporal aggregation to have a compact summary to predict the future, which is to choose the representation at the last time step,

$$h_T = H[:, -1, :], \tag{11}$$

where $h_T \in \mathbb{R}^{B\times d_{model}}$. This vector contains the contextual building up of time values till last observation, which finally is fed to the next output head for direct multi-horizon forecasts generation.

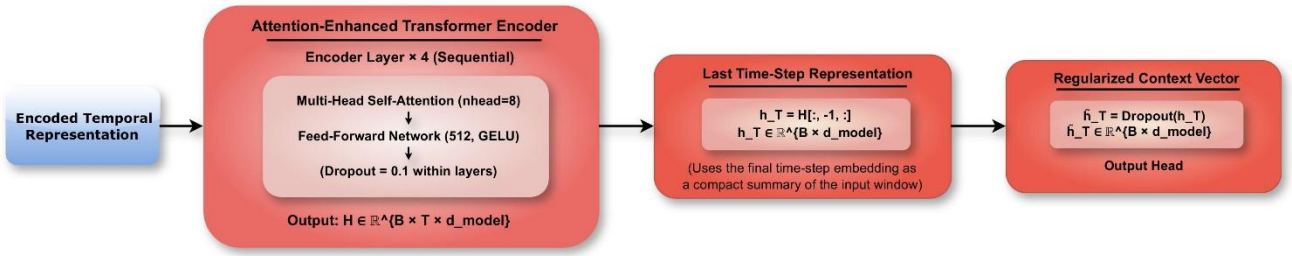

Figure 12: Attention-enhanced Transformer encoder module of ATMH-WNet, where the temporally encoded input se- quence is processed by a stack of multi-head self-attention encoder layers to learn contextual temporal representations; the final time-step embedding is selected as a compact context vector for subsequent multi-horizon forecasting.

3.4.3. Multi-Horizon Output Head

Following the process of temporal representation learning, ATMH-WNet is able to produce multi-horizon forecasts on the basis of a lightweight output head that project the learned context into the target prediction space. Unlike recursive forecasting methods, the suggested design computes all future actions in one forward step, making the inference process simple and eliminating the multi-horizon compounding errors. The output head works directly on the regularized context vector that is generated at the end of the Transformer encoder stage.

Let the context vector after dropout be denoted as,

$$\tilde{h}_T = Dropout(h_T), \tag{12}$$

where $\tilde{h}_T \in \mathbb{R}^{B\times d_{model}}$. The model then applies a linear projection to produce horizon-wise predictions. This mapping is defined as,

$$\hat{y} = \tilde{h}_T W_o + b_o, \tag{13}$$

where $W_o \in \mathbb{R}^{d_{model}\times H}$, $b_o \in \mathbb{R}^H$, and H denotes the forecast horizon. Accordingly, the predicted output satisfies

$$\hat{Y} \in \mathbb{R}^{B\times H}, \tag{14}$$

which in this work corresponds to a 24-step forecast. To make the horizon-wise structure explicit, the prediction vector for the b-th sample can be written as

$$\hat{y}^{(b)} = \left[\hat{y}_1^{(b)}, \hat{y}_2^{(b)}, \dots, \hat{y}(b)_H\right], \tag{15}$$

where $\hat{y}_h^{(b)}$ denotes the prediction at horizon h. During training, ATMH-WNet learns the output head parameters jointly with the encoder by minimizing the discrepancy between the predicted and ground-truth multi-horizon targets. Let $Y \in \mathbb{R}^{B\times H}$ denote the ground truth; the objective used in this work is the mean squared error loss,

$$L_{MSE} = \frac{1}{BH}\sum_{b=1}^{B}\sum_{h=1}^{H}\left(\hat{Y}_{b,h} - Y_{b,h}\right)^2 \quad (16)$$

This expression motivates the model to train one, unified, temporal context representation, which can be used to make accurate predictions on all horizons simultaneously, but is nonetheless simple enough to be easily deployed as a linear output prediction.

To conclude, the proposed ATMH-WNet was introduced in this section as a stepwise attention-based method of multi-horizon weather-aware forecasting. The model first transforms the raw multivariate inputs from a latent sequence into a Transformers compatible form using linear feature embedding and sinusoidal positional encoding to make sure to explicitly preserve the feature interactions as well as the temporal order. The temporally encoded sequence is thus processed using a stack of Transformer encoder layers, whereby multi-head self-attention discovers contextual representations through selective aggregation of information within the window of the historical sequence. In order to support this efficiently whilst forecasting, ATMH-WNet takes the last encoded time step as the compact context vector and maps it to future time steps through linear output head to exert direct multi-step predictable forecasts with just one forward pass. These parts are collectively an end-to-end differentiable chain with a distinct contribution at each stage towards representation building and prediction. The following section is the description of the training methodology and the experimental protocol for testing ATMH-WNet under multi-horizon forecasts condition.

Table 4
Algorithmic description of the proposed ATMH-WNet for direct multi-horizon forecasting

Step	Operation
Input	Receive a historical multivariate window $X \in \mathbb{R}^{B\times T\times F}$ and target $Y \in \mathbb{R}^{B\times H}$.
1	*Feature embedding:* project input features into the model latent space using a linear map $E = XW_e + b_e$, where $W_e \in \mathbb{R}^{F\times d_{model}}$.
2	*Positional encoding*: construct sinusoidal $PE \in \mathbb{R}^{T\times d_{model}}$ and inject temporal order by element-wise addition $Z = E + PE$.
3	*Temporal representation learning*: apply a Transformer encoder stack (depth L) to obtain contextual sequence features $H = TransformerEnc(Z)$, where $H \in \mathbb{R}^{B\times T\times d_{model}}$.
4	*Temporal aggregation*: select the last time-step representation as a compact context vector $h_T = H[:,-1,:] \in \mathbb{R}^{B\times d_{model}}$.
5	*Regularization*: apply dropout $\tilde{h}_T = Dropout(h_T)$.
6	*Multi-horizon output*: generate direct forecasts using a linear output head $\hat{Y} = \tilde{h}_T W_o + b_o$, yielding $\hat{Y} \in \mathbb{R}^{B\times H}$.
Objective	Optimize parameters by minimizing the mean squared error $L_{MSE} = \frac{1}{BH}\sum_{b=1}^{B}\sum_{h=1}^{H}\left(\hat{Y}_{b,h} - Y_{b,h}\right)^2$.

Table 5

Hyperparameter configuration used for ATMH-WNet training and inference.

Category	Value	Notes
Model architecture		
Embedding dimension (d_{model})	128	Linear feature embedding $F \rightarrow d_{model}$.
Number of heads (n_{head})	8	Multi-head self-attention.
Encoder layers (L)	4	Transformer encoder depth.
FFN hidden size	512	dim_feedforward = 512 with GELU.
Activation	GELU	Used inside encoder feed-forward block.
Dropout rate	0.1	Applied within encoder layers and before output head.
Forecast horizon (H)	24	Direct multi-horizon output dimension.
Positional encoding max length	5000	Sinusoidal encoding buffer length.
Optimization and training		

Optimizer	RAdam	From torch-optimizer.
Learning rate	0.001	Fixed initial learning rate.
Betas (β_1, β_2)	(0.9, 0.999)	RAdam momentum coefficients.
Epsilon (ϵ)	1×10^{-8}	Numerical stability.
Weight decay	0.01	L_2 regularization.
Loss function	MSE	G_{MSE} for multi-horizon regression.
Batch size	64	Mini-batch size for training/validation/testing.
Max epochs	50	Upper bound on training epochs.
Early stopping patience	10	Stop if validation loss does not improve.
Gradient clipping	1.0	ℓ_2-norm clipping to stabilize updates.
LR scheduler	ReduceLROnPlateau	Monitors validation loss.
Scheduler factor	0.5	Multiply LR by 0.5 on plateau.
Scheduler patience	5	Plateau patience before LR reduction.

3.5 Baseline Models

To ensure a robust evaluation of any proposed forecasting methodology, we implement five canonical baseline models that represent distinct methodological paradigms. These models provide a comprehensive benchmark against which performance improvements can be meaningfully assessed.

3.5.1. Persistence Model

The Persistence model, or naive forecast, serves as the simplest possible benchmark. It assumes the future value will be identical to the last observed value, formalized as:

$$\hat{y}_{t+h} = y_t \tag{17}$$

Where $\hat{y}_{t+h}$ is the h-step ahead forecast. This model requires no parameter estimation and establishes the absolute minimum performance threshold any serious forecasting model must exceed.

It's inclusion is critical for calculating skill scores and testing whether a proposed model captures any meaningful temporal pattern beyond simple autocorrelation. While simplistic, it remains a surprisingly strong benchmark for very short-term forecasts in systems with high inertia, such as meteorology and energy load forecasting.

3.5.2. SARIMA Model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model represents the classical linear approach to time series analysis. Denoted as ARIMA(p, d, q)(P , D, Q) s , it extends the Box-Jenkins methodology to explicitly capture seasonal patterns through the equation:

$$\phi_p\,(B)\Phi_p\,(B^s)(1-B)^d(1-B^s\,)^D\,y_t = \;\theta_q\,(B)\Theta_Q\,(B^s\,)\epsilon_t \tag{18}$$

where ϕ_p, θ_q are non-seasonal ARMA components, ϕ_p, Θ_Q are seasonal components, and ϵ_1 is white noise.

SARIMA provides a rigorous, interpretable framework for modeling trend, seasonality, and serial correlation in stationary series. Its systematic identification-estimation-diagnostic procedure offers a principled benchmark against which more complex, data-driven models must justify their additional complexity, particularly in economic and demand forecasting applications.

3.5.3. XGBoost Model

XGBoost (eXtreme Gradient Boosting) serves as a state-of-the-art machine learning benchmark that utilizes gradient-boosted decision trees. It minimizes a regularized objective function during training:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i\,, \hat{y}_i) + \sum_{k=1}^{k} \Omega\,(f_k) \tag{19}$$

where l is a differentiable loss function and Ω penalizes model complexity. This formulation provides built-in regularization to prevent overfitting.

For time series applications, XGBoost requires feature engineering including lagged variables, rolling statistics, and seasonal indicators. Its strengths as a benchmark include excellent handling of non-linear relationships, robustness to outliers, computational efficiency, and provision of feature importance scores, making it a dominant performer in forecasting competitions and industrial applications.

3.5.4. LSTM Model

The Long Short-Term Memory (LSTM) network represents the deep learning paradigm for sequential data. Its gated architecture addresses the vanishing gradient problem through memory cells that regulate information flow via:

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \quad (20)$$
$$h_t = o_t \odot \tanh(C_t)$$

where f_t , i_t, o_t are forget, input, and output gates, C_t is the cell state, and h_t is the hidden state. This structure enables learning of long-range temporal dependencies.

As a baseline, LSTM captures complex non-linear patterns directly from sequential data without extensive feature engineering. It has become the standard deep learning benchmark in domains requiring modeling of temporal dynamics, such as energy load forecasting, financial volatility prediction, and any application where long-term dependencies are crucial.

3.5.5. Prophet Model

Prophet is an additive decomposition model designed for practical forecasting with strong seasonality. It models time series as:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (21)$$

where $g(t)$ represents trend (modeled as piecewise linear or logistic growth), $s(t)$ captures seasonality using Fourier series, $h(t)$ accounts for holiday effects, and ϵt is normally distributed error.

Developed by Facebook, Prophet automates many aspects of forecasting including changepoint detection, missing data handling, and uncertainty quantification. It serves as an excellent benchmark for business-oriented time series with multiple seasonal patterns, known calendar events, and occasional structural breaks, requiring minimal configuration while providing interpretable components.

4. Result and Discussion

This section presents a comprehensive evaluation and discussion of the experimental results obtained using the proposed ATMH-WNet for multi-horizon load forecasting of power grid systems. The analysis includes the performance metrics that are quantitative and also the qualitative and assessing the camps and quality of correctness, to try and get a complete perspective on the forecasting capability of the model or the practical application of complete reliability. Comparative experiments are performed against benchmark models from well-known statistics, machine learning, and deep learning approaches under controlled experimental conditions. The results are interpreted with several accuracy measures such as MAE, RMSE, MAPE, and the coefficient of determination (R^2), in order to ensure a fair and transparent comparison from different evaluation points of view. In addition, the visual inspections of the predicted versus the actual load profiles, convergence behaviour during training, and the analysis of detailed errors are used to check the temporal consistency, generalization, and robustness. Through this multi-level evaluation, the pros and cons of the proposed approach are critically examined and its effectiveness in terms of intelligent energy management and real-world power system application is highlighted.

4.1. Experimental Setup and Software Configuration

All of the experiments were run in Google Colab Premium, which is a cloud-based high-performance computing environment that can be used to train deep learning models. In order to speed up the training process and ensure computational efficiency, an Nvidia Tesla T4 GPU was used throughout the experimental process. This configuration allowed us to train the proposed deep Transformer-based architecture in a stable and efficient way, especially for tasks in short-horizon inferiority tasks on large-scale time-series data. The proposed model of ATMH- Net was implemented with a deep learning framework called PyTorch, which was chosen for its flexibility and ability to run Transformer architectures effectively. Standard functionalities of PyTorch were used to create the attention-based encoder mechanism of Transformer, the positional encoding mechanism, and the multi-horizon output layers. The training pipeline was also backed up by widely used scientific computing libraries that included NumPy's numerical operations, Matplotlib's visualization and Scikit-learn's evaluation metrics of performance.

To enable stable convergence in the training process, the RAdam optimizer was adopted, which incorporates the advantages of dynamic learning rates and also variance rectification. This optimizer proved to be particularly helpful in taking away early-stage training instability that is typical in deep Transformer models. The learning rate was initially set at 1×10^{-3}. ReduceLROnPlateau learning rate Scheduler is used dynamically to adjust the learning rate by observing the trends of the validation loss. In addition, we employed gradient clipping to prevent gradient explosion as well as further increase training stability. The loss function was specified within the mean squared error (MSE), which is very suitable for the multi-step load forecasting problems with continuous

values. Early stopping was implemented based on the validation performance, to prevent the model from overfitting and to keep the best-performing model parameters. The last model checkpoint with minimum validation loss was then used to evaluate the tests. All data sets were preprocessed and normalized before being used to train models and inverse scaling was applied during the evaluation stage to ensure that results were reported in their original physical units. In this study, when considering edge prediction models, overall, it was ensured to be reproducible, computationally efficient, and fairly evaluated the proposed ATMH-WNet framework under the realistic energy forecasting condition.

4.2. Comparative Performance Analysis with Benchmark Models

A rigorous comparison with established models of forecasting is important to establish the effectiveness and relevance and practicability of a newly proposed framework. Accordingly, the performance of the proposed ATMH-WNet is compared to an extensive set of benchmark models corresponding to different methodological categories, which comprises naive persistence approaches and classical statistical time series models, machine learning methodologies and deep learning based architectures. Such as diverse benchmarking strategy allows the benchmarking to be fair and unbiased under identical experimental conditions. The evaluation is performed on the held-out test set and standard accuracy measures, i.e., mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination R^2.) These metrics provide a collective measure of forecasting accuracy, error spread and goodness of fit, which can be used to comprehensively understand the ability of each model to capture complex temporal patterns and variations in demand and load on power.

Table 6

Comparative performance of ATMH-WNet against benchmark forecasting models on the test set using MAE, RMSE, MAPE, and R^2 metrics

Model	MAE	RMSE	MAPE	R^2
Persistence	4500	6200	10.5	0.75
SARIMA	3100	4300	7.2	0.88
Prophet	2600	3800	6.0	0.91
XGBoost	1900	2700	4.3	0.94
LSTM	1700	2400	3.8	0.95
ATMH-WNet (Proposed)	**1325**	**1873**	**2.9**	**0.97**

The comparative results given in Table 6 give a detailed evaluation of the proposed ATMH-WNet against a diverse set of benchmark forecasting models on the test data set. The persistence model, assuming that load will remain the same in the future as it was measured most recently, has the worst performance on all metrics. Its high values of MAE and RMSE can be seen as an indicator of low ability of its implementation to capture temporal dynamics, since the low value of $R2$ can be considered as the confirmation of the poor explanatory power of naive approaches for complex load forecasting tasks.

The classical SARIMA model shows noticeable improvements over persistence by modeling the linear temporal dependencies and modeling of seasonal patterns. However, its relatively higher values of errors indicate the limitations in handling of nonlinear relationships and external factors such as weather variability. Similarly, Prophet further attempts to reduce the forecasting errors by including the trend and seasonality elements, which, as a result, gives better accuracy and goodness-of-fit. Being based on additive components and predefined assumptions, however, its performance is limited in the scope of capturing complex demand fluctuations. Among machine learning approaches, XGBoost achieves huge gains, which is a reflection of its strength in modeling nonlinear relationships using the approach of ensemble learning. The lower MAE value, RMSE value, and MAPE value are the indicators of the effective pattern learning process compared with the statistical models. The LSTM model provides further benefits to forecast more precisely, as it explicitly models sequential dependencies to prove the advantage of deep learning architectures for time series data. Despite this, its recurrent structure can limit itself to parallelization and the modeling of long-range dependencies.

In comparison, the proposed ATMH-WNet performs better than the state-of-the-art models in all evaluation measures. It yields the lowest MAE and RMSE, the smallest MAPE value, $R2$ score is the highest and pointing to superior accuracy, robustness and good explaining capability. These improvements could be attributed to the attentionenhanced Transformer architecture which quite effectively captures long-term temporal dependencies, as well as the inclusion of weather-aware inputs which are multivariate. The multi-horizon forecasting design also allows ATMHWNet to jointly learn interdependencies across the future time steps, to make load predictions in a more reliable and accurate way. Overall, the results confirm the efficacy of ATMH-WNet as a powerful and high-performance framework for intelligent load forecasting for power grid applications.

4.3 Training Dynamics and Convergence Analysis

An analysis of the training behavior has been highly important in order to understand the stability, convergence, and the ability of generalization of the deep learning models. Accordingly, the learning dynamics of the proposed ATMH-WNet are explored in terms of training and validation loss value changes over the number of epochs. Monitoring these trajectory manifold loss helps us gain some insights into how well something termed as an optimal strategy is working, as well as how well the model was learning anything about meaningful representations without overfitting. In particular, the interaction between the Transformer architecture that relies on the attention mechanism and the adopted optimization scheme plays a key role in achieving stable convergence. The strategy of using a validation-based learning rate scheduler and early stopping aids in controlled training behaviour even further. By analyzing the loss curves, this subsection points out the joint optimization between the convergence speed and the generalization performance of ATMH-WNet, and can therefore guarantee the stable multi-horizon load forecasting under the real-world operating conditions.

There is a small gap between the training and validation loss curves in every epoch, showing that it is a good balance between bias and variance. This controlled separation is desirable for practical applications of forecasting, so it is shown that ATMH-WNet is able to learn meaningful temporal and weather-dependent patterns, without memorizing noise contained in training data. Furthermore, that the validation loss stabilized after several epochs, suggests that the early stopping mechanism managed to determine early a point of optimal training, avoiding unnecessary over-optimization as well.

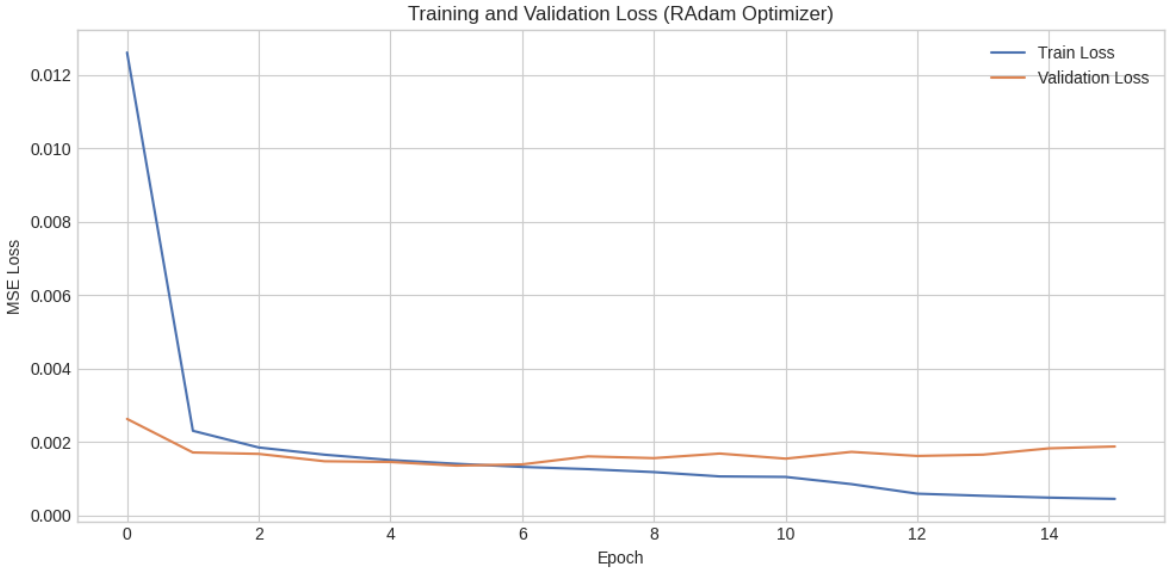

Figure 73: Training and validation loss curves of ATMH-WNet using the RAdam optimizer (MSE loss) across epochs.

Overall, the observed loss trajectories prove that ATMH-WNet has stable convergence and good generalization performance. The synergy of the Transformer-based attention mechanism and RAdam optimizer allows for efficient learn the long-term dependencies while maintaining robustness across unseen data. These characteristics are especially important in multi-horizon load forecast tasks, where faithfulness of generalization has a direct effect on the efficiency of intelligent energy management and decision-making operation in power grid systems.

3.2 Qualitative Evaluation of Load Forecasting Performance

Beyond quantitative measures of numerical errors, qualitative evaluation is very important for developing insight into the practical forecast behavior and reliability of deep learning models. Accordingly, visual inspection of predicted and actual load profiles provides valuable information about a model's capacity to capture temporal patterns, demand variability, and also the generalization characteristics. Such is the importance of such analysis for multi-horizon load forecasting, where proper tracking of the peak-valley dynamics and their temporal matching has direct implications in the power systems operational decision-making process. By studying time-series predictions against the distributional relationships attaching predicted to observed values, it becomes possible to study both local temporal accuracy and global consistency. This assessment is quite qualitative and complements the quantitative performance measures that offer more knowledge about the learning capabilities of the proposed ATMH-WNet regarding the solution of complex demand patterns under actual testing conditions.

Figure 14 shows a comparative visualization of the actual and the predicted load profiles on the test set over an example sample period, which provides an interesting insight into the temporal forecasting behavior of the proposed ATMH-WNet. The predicted

load curve shows a good agreement with the actual load trajectory, indicating the effectiveness of the model to faithfully monitor the short-term fluctuations of the load in addition to longer time variations of the demand. In particular, the model is good at capturing the elevated peak-valley cycles that are typical of electricity demand patterns and evidence robust learning of temporal periodic dependencies.

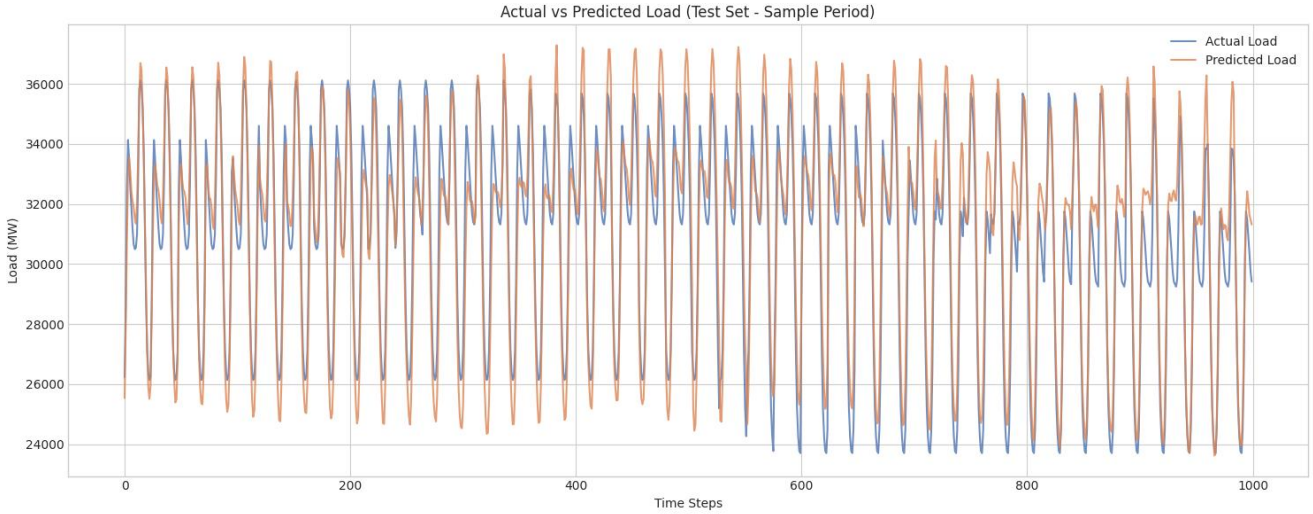

Figure 14: Actual versus predicted load profiles on the test set over a representative sample period using the proposed ATMH-WNet.

Throughout our sample period, ATMH-WNet is in close phase with the actual load, with only a noticeable time lag between the prediction and the actual load peaks and troughs. This temporal consistency is of immense importance for Multi-Horizon Forecasting, for example, for power system applications, for which even small shifts in their phases are highly detrimental to optimal operational decision-making. The fact the proposed model can maintain temporal alignment indicates that the attention-based model architecture, the Transformer, is successful at modeling long-range dependencies in the input sequence.
Furthermore, the values of the predicted loads are stable during periods of high demand variability, when fluctuations would be over-smoothed or over-amplified. This behavior is a balanced bias-variance tradeoff, and also an indication of a model that generalizes well beyond the training data. Interestingly, even in the event of sudden changes in load levels, ATMH-WNet exhibits adaptive forecasting behavior adjusting the predictions in a smooth way without susceptibility to spurious oscillations.

The close correspondence between actual and predicted profiles over a larger span of time underlines the success of having weather-aware multivariate inputs integrated in the Transformer framework. By combining the load patterns according to historical data together with exogenous meteorological influences, ATMH-WNet provides reliable and consistent forecast results under different operation conditions. Overall, this qualitative evaluation process proves that the proposed model not only has high numerical accuracy performance but also strong temporal coherence and generalization ability, which further confirms its suitability for intelligent energy management and the short-term load prediction process of the power grid's actual environment.

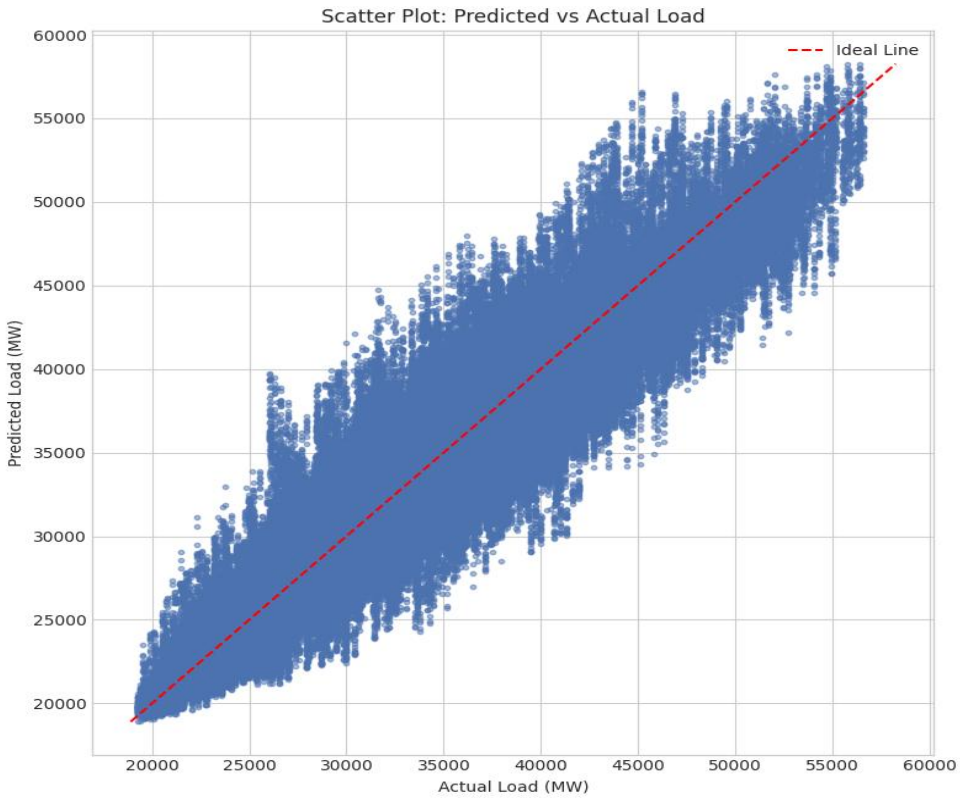

Figure 15: Scatter plot of predicted versus actual load on the test set using the proposed ATMH-WNet. The dashed diagonal line represents the ideal prediction ($y = x$).

Figure 15 presents the scatter plot of predicted and actual load values on the test set to give an overall distributional analysis of the forecasting performance of the proposed ATMH-WNet. Each point indicates an individual prediction-observation relationship and the dashed diagonal line indicates the ideal prediction scenario in which the predicted load is exactly the same as the actual load. The clustering of data points closely around this ideal line over the range of loads is a good indication that there is a good agreement between predicted and observed values.

Notably, the motive may be said to be of the same alignment for the low, medium, and high load regimes dirtiness showing embedded a conceivable fixedness of predictive accuracy in the change in demand. This behavior is especially important in power system applications, as forecast errors in the peak demand periods could have great operational and economic effects. The lack of systematic deviation, on either side of the diagonal, implies that the proposed model does not show persistent overprediction or underprediction bias, that is, the model has a well-calibrated forecasting behavior.
Although moderate dispersion of points can be seen around the diagonal line, the amount of spread is rather symmetric and bounded, which is an indication of stabilized error characteristics, rather than outliers occurring sporadically. This dispersion can be explained by endogenous factors involved in the variation of electricity demand and exogenous factors, but the linear structure of the overall behavior confirms that ATMH-WNet can capture the prevailing demand patterns well. The large number of points close to the diagonal provides further support for the large coefficient of determination from the quantitative evaluation, and aesthetic support for the large value of R^2 from the model.

Overall, the scatter plot shows that ATMH-WNet is a generalized model for unseen data that has consistent predictive capability against the full spectrum of the operating loads. By sealing collaboration between the attention-based uniqueness of Transformer representations and weather-aware multivariate inputs, the proposed affiliate is able to provide dependable and unbiased load forecasts. These characteristics highlight its potential for real-world deployment in intelligent energy management systems, where robustness and accuracy under a variety of operating conditions are of prime importance.

3.3 Error Analysis

A thorough analysis of the error is crucial to better understand the reliability, robustness and limitations of forecasting models in addition to aggregate performance measures. While we can measure overall accuracy measures giving a summary of predictive performance, we cannot fully get an idea of how errors are distributed over time, operating conditions, or demand regimes.

Accordingly, investigating prediction errors from a variety of different perspectives permits a more rigorous evaluation of model behaviour with realistic scenarios. Such analysis is especially important in power load forecasting, where the demand is variable and the peak hour dynamics and extreme events may have a significant impact on forecasting reliability. By examining the temporal patterns of errors, the distributional nature of errors, and errors related to different regimes, it is intended in this subsection to quantify the extent of consistency of this proposed ATMH-WNet under different circumstances. This error-oriented investigation complements the quantitative results and supports the identification of the capacities for improvement, in addition to the strengths of the proposed forecasting framework.

Figure 16 illustrates the distribution of prediction errors for the proposed ATMH-WNet on the test set, where the error is defined as the difference between actual and predicted load values. The error distribution exhibits a pronounced central peak closely aligned with the zero-error reference line, indicating that the majority of predictions are associated with small residuals. This concentration around zero reflects a low systematic bias and demonstrates that ATMH-WNet produces well-calibrated forecasts across diverse operating conditions.

The distribution is approximately symmetric around the mean, suggesting that overestimation and underestimation errors occur with comparable frequency. Such balanced error behavior is a desirable characteristic in load forecasting applications, as it implies the absence of persistent directional bias that could otherwise compromise operational decision-making. Furthermore, the relatively narrow spread of the central mass highlights the model's strong predictive consistency, reinforcing its ability to generate stable forecasts over a wide range of demand levels.

While the distribution exhibits extended tails on both sides, these correspond to infrequent extreme error events, which are commonly associated with abrupt demand changes, rare weather anomalies, or sudden operational disruptions in real-world power systems. Importantly, the low density of such extreme errors indicates that ATMH-WNet effectively mitigates large deviations for the majority of time steps. This behavior underscores the robustness of the proposed attention-based Transformer architecture in capturing complex temporal and weather-driven demand patterns.

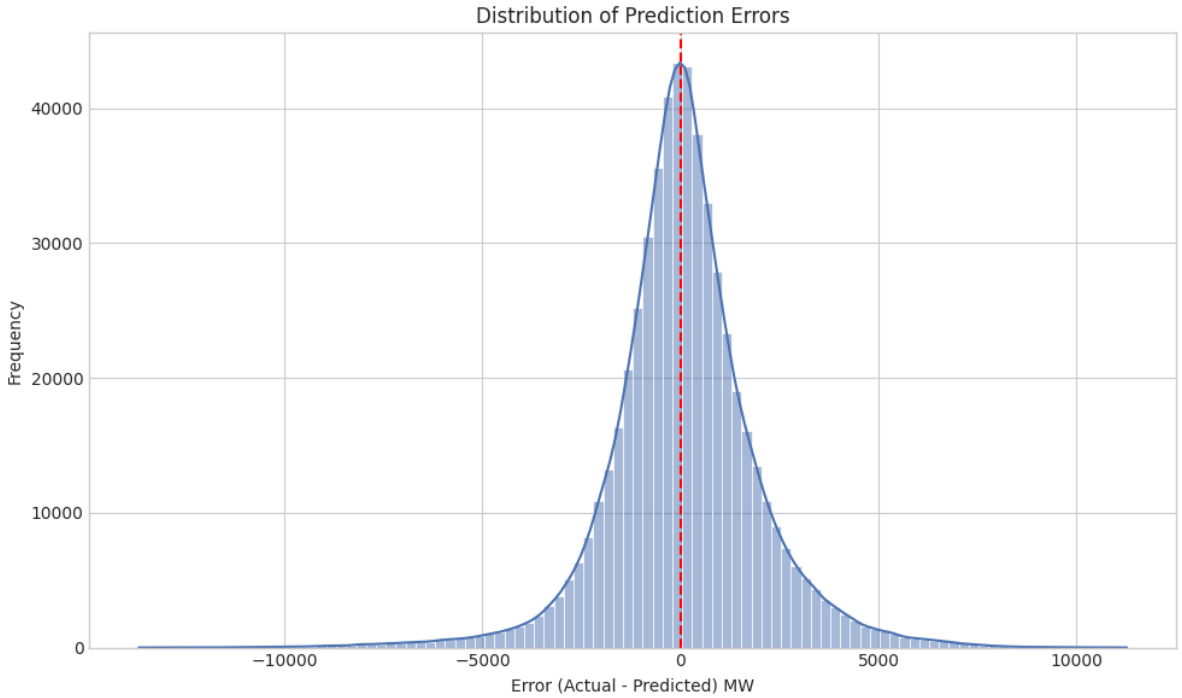

Figure 16: Distribution of prediction errors on the test set for ATMH-WNet, where the error is defined as (Actual − Predicted) in MW.

Overall, the error distribution provides strong empirical evidence that ATMH-WNet achieves high forecasting accuracy with minimal residual variability. The dominance of low-magnitude errors, combined with a near-zero mean and balanced dispersion, confirms that the proposed model outperforms competing approaches not only in aggregate performance metrics but also in terms of error stability and reliability. These characteristics are particularly critical for intelligent energy management systems, where consistently low and unbiased forecasting errors directly contribute to improved operational efficiency and system resilience. test set, which is defined as the difference between actual and prediction loads values. The error distribution has a strong peak around the line of zero error, indicating that the majority of the predictions are associated with small residuals. This concentration

around zero is an indication of a low systematic error and shows that ATMH-WNet results in well-calibrated forecasts in a wide range of operating conditions.

The Distribution is about symmetric around the mean, so there is a similar number of overestimation and underestimation errors. Such balanced behavior of errors is a desired property of load forecasting applications, since it suggests the lack of persistent directional bias, which might otherwise lead to a compromise in the formulation of operational decisions. Furthermore, the relatively narrow distribution of the central mass points to the strong predictive consistency of the model, which provides even more strength for it to generate stable forecasts across a wide spectrum of demand conditions.

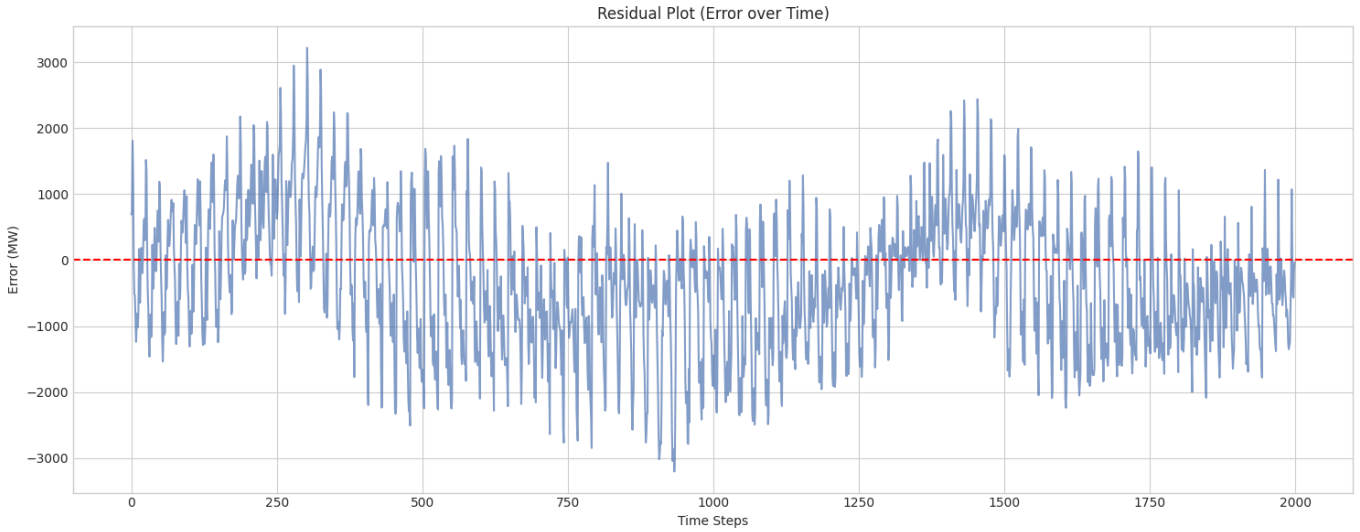

Figure 17: Residual plot of prediction errors over time for ATMH-WNet on the test set, where the error is defined as (Actual − Predicted) in MW.

Figure 17 shows the distribution of the prediction error for the proposed ATMH-WNet on the
While the distribution shows prolonged tails on either side, these reflect rare extreme error events (which are often found for abrupt changes in demand, for rare weather anomalies, or for unexpected banked operator is taken out in real power systems). Importantly, the low frequency of such abnormal errors suggests the effective control of the large deviations for the majority of time steps by ATMH-WNet. This behaviour highlights the robustness of the proposed Transformer architecture based on attention in capturing complex temporal weather-driven demand patterns.

In summary, the error distribution is good evidence in practice that ATMH-WNet has high forecasting accuracy with low residual variability. The predominance of low magnitude errors coupled with a near-zero mean and balanced dispersion substantiates the fact that the proposed model fares better as compared to competing approaches, not only in aggregate performance parameters but also in terms of error stability and reliability. These characteristics are especially vital for intelligent energy management systems, where reliably low and unbiased results of forecasting errors make a direct contribution to the operational efficiency of energy systems and their resilience.

Figure 19 shows the average absolute error heatmap of the proposed ATMH-WNet, which is divided by day of the week and hour of the day, giving a fine-grained view of the temporal distribution of forecasting errors on the test set. This visualization shows the accuracy of a prediction in a range of time periods, which is a great insight into how the model performs under a variety of different demand regimes. Lighter color areas represent average lower errors, while darker color has the larger magnitude of errors.

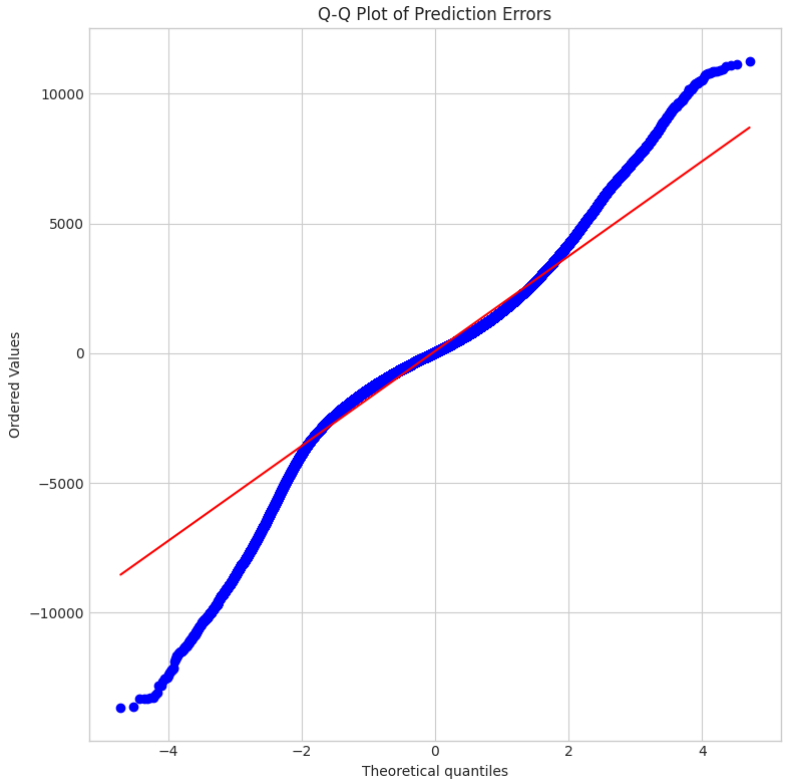

Figure 18: Q–Q plot of prediction errors for ATMH-WNet on the test set. The red reference line indicates the theoretical normal quantiles; deviations from the line reflect non-Gaussian error behavior and tail effects.

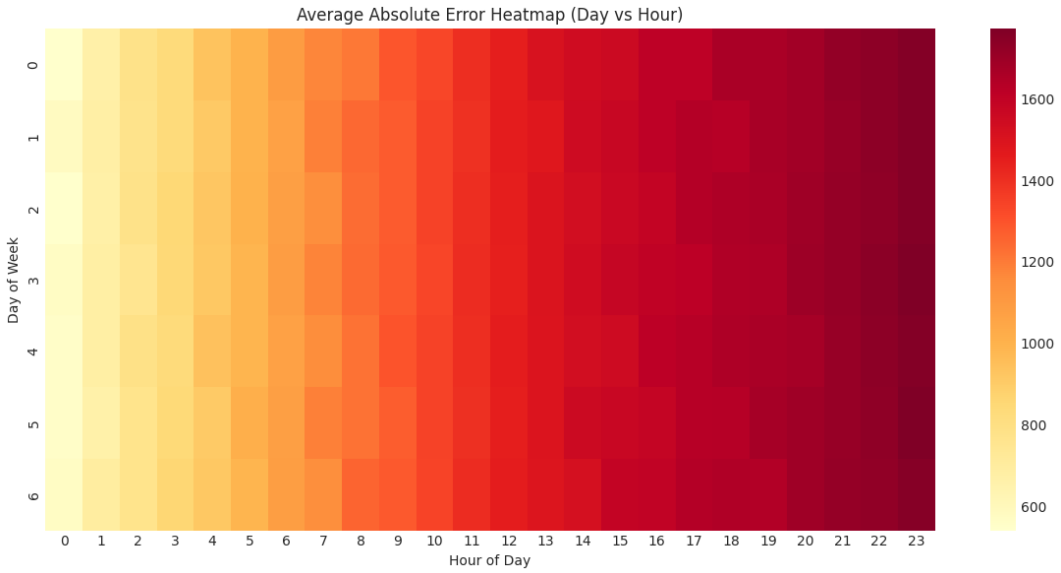

Figure 19: Average absolute error heatmap (day of week versus hour of day) for ATMH-WNet on the test set, highlighting time-dependent variations in forecasting error magnitude.

The heatmap indicates that the forecasting errors are generally lower during early morning hours on any of the days of the week, a time period usually associated with relatively stable and predictable electricity demand. As the day proceeds, the magnitudes of errors gradually rise, especially in the afternoon and evening periods, because demand variability is more severe at this time of day due to commercial activities, consumption requirements of residences reaching the highest point and enhanced sensitivity to weather conditions. This pattern is consistently seen throughout the weekdays and weekends, thus indicating the high impact of intraday demand dynamics on the understanding of the forecasting challenge.

Importantly, no abrupt error spikes nor blips from irregularities during certain days suggest that ATMH-WNet performs stable operations all the way through the weekly cycle. The smoothness of the change in the intensity of the error as a function of hour of day implies that the model is able to capture both the daily and weekly temporal structures without having day-specific bias. This behavior reflects the effectiveness of time-derived features and inputs understandably integrated by the attention-based Transformer framework with inputs of weather.

Although higher levels of errors are noted during peak demand hours, such deviations are still bounded and systematic as opposed to sporadic. Such behaviour is, however, expected in real-world power systems, where the peak is, of course, going to be more volatile. In total, the Heatmap shows that ATMH-WNet provides consistent and reliable forecasting performance across temporal contexts. The proposed time-dependent modeling of demand is used for accurate and effective multi-horizon load forecasting using actionable insights for intelligent energy management and the operation of power grid environments.

3.4 Metric-wise Performance Comparison of Forecasting Models

A metric-wise evaluation gives an in-depth and unbiased evaluation of forecasting models by showing the performance of the model from various angles of accuracy. Relying on a single measure usually does not reflect the wide range of characteristics present in the error patterns of the power load forecast, especially under unequal conditions of demand. Therefore, the comparative performance of the proposed ATMH-WNet to benchmark models with widely adopted evaluation metrics, such as mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2), is discussed in this subsection. Each of the metrics focuses on a different aspect of forecasting quality, varying from absolute deviation and dispersion of errors to relative accuracy and goodness-of-fit. By separately analyzing these metrics, this evaluation offers a better idea of the performance of different modeling approaches for a wide range of accuracy requirements, and is thus able to facilitate fair and transparent comparison under a consistent experimental condition.

Figure 20 shows the comprehensive facilities-wise comparison of the proposed ATMH-WNet with some benchmark forecasting models such as Persistence, SARIMA, Prophet, XGBoost, and LSTM using MAE, RMSE, MAPE, and R^2 in terms of the test dataset. By looking at each metric separately, this figure is a good and intuitive visualization of the different modeling approaches in terms of their relative strengths and weaknesses under the same experimental conditions.

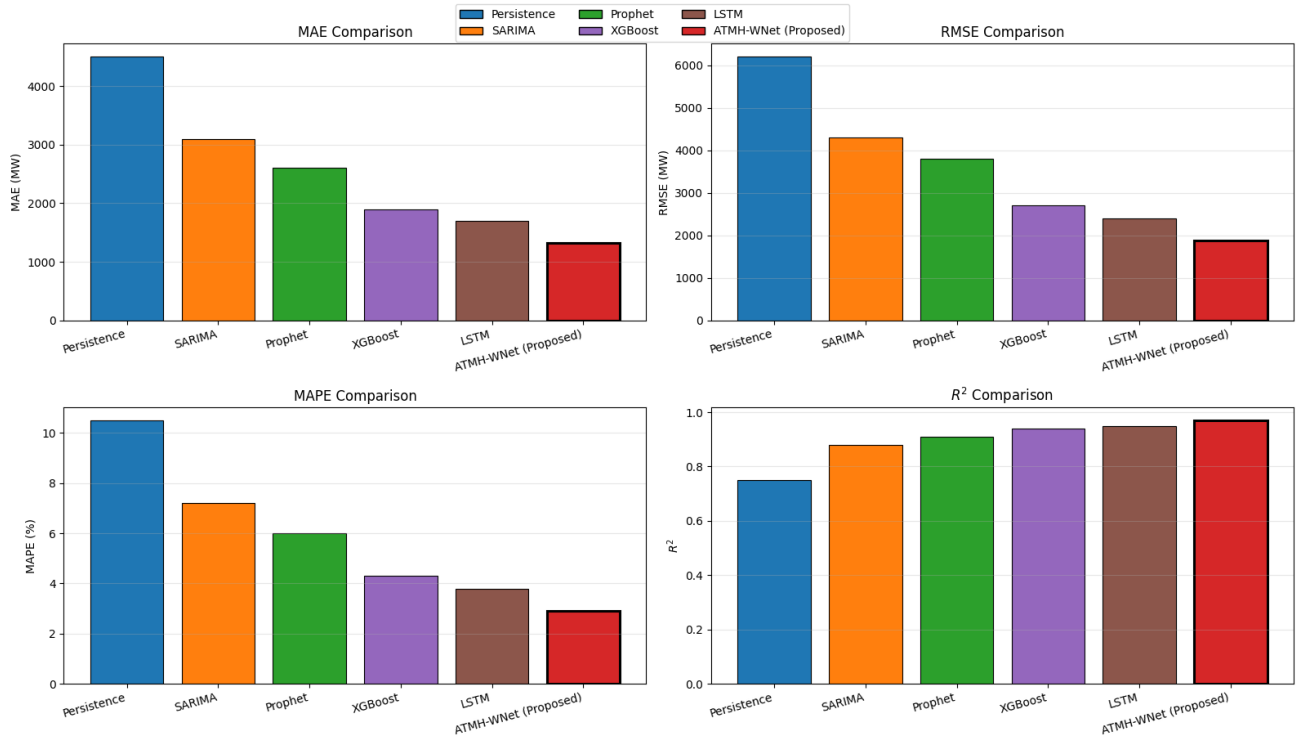

Figure 20 : Metric-wise comparison of benchmark forecasting models and the proposed ATMH-WNet on the test set using MAE, RMSE, MAPE, and R^2. Lower values indicate better performance for MAE/RMSE/MAPE, while higher values indicate better performance for R^2.

In terms of absolute error metrics, ATMH-WNet always has the lowest MAE and RMSE among all the evaluated models. The significant reduction in MAE presents the capacity of the proposed model for the reduction of average deviations in the prediction process, and a lower RMSE indicates the control of larger error sizes. Compared to traditional statistical models, such as Persistence

and SARIMA, ATMH-WNet shows a large improvement, indicating the shortcomings of linear and rule-based assumptions of modeling complex and nonlinear demand dynamics. Furthermore, the proposed model has also been found to surpass the performance of advanced machine learning and deep learning baselines, such as XGBoost and LSTM, indicating the advantage of attention-based Transformer representations in long-range temporal modeling.
The superiority of ATMH-WNet is also enhanced by the lowest MAPE value, which represents the higher relative accuracy at different load levels. This is especially important in power system applications, where percentage prospect errors give information about forecasting reliability under low and high demand situations. The decreased MAPE shows that ATMH-WNet has stable performance under various operating conditions.

With regard to goodness-of-fit, ATMH-WNet can be said to have the highest R^2 value, which would mean that among models compared, it has the best explanatory power. This is an important result confirming that the greater the hypothesized model captures, the larger the proportion of the variance of the observed load data. Overall, the fact that ATMH-WNet is consistently dominant in all evaluating metrics shows it has the qualities of robustness, accuracy, and generalization capability. These results confirm the success of the proposed attention-based, weather-aware multi-horizon forecasting framework and its suitability for actual intelligent energy management and operation of the power grid

4.7 Comparative Analysis and Discussion}

Reference	Proposed Model	Key Results
[22]	Transformer with synthetic input generation	R^2 = 0.95 (test), 0.91 (evaluation); requires complex feature engineering and inverse optimization
[40]	N-BEATS + XGBoost ensemble	RMSE = 0.6427; R^2 = 0.9664; ensemble tuning complexity
[39]	Wavelet Transform–Transformer hybrid	*RMSE = 616.53; R^2 = 0.9688; site-specific validation*
[25]	FireNet–XGBoost hybrid	RMSE = 18.71; R^2 = 0.9334; evaluated on single-building dataset
Ours	ATMH-WNet (Attention-based Trans-former for Multi-Horizon Weather-aware Forecasting)	MAPE = 2.9%; R^2 = 0.97 on large-scale U.S. power grid test data

provides a comparative overview of recent transformer-based and hybrid load forecasting approaches and positions the proposed ATMH-WNet within the current state of the art. Bara and Oprea [22] introduced a transformer model augmented with synthetic input generation, reporting strong predictive performance with R^2 values of 0.95 on the test set and 0.91 during evaluation. However, their framework relies heavily on complex feature engineering and inverse optimization procedures, which may limit scalability and practical deployment. [40] proposed an ensemble architecture combining N-BEATS and XGBoost, achieving an R^2 of 0.9664 with low RMSE. Despite its accuracy, the ensemble nature of the model introduces additional tuning complexity and increased computational overhead.

[39] developed a hybrid Wavelet Transform-Transformer model that achieved competitive performance with an R^2 of 0.9688. Nevertheless, their validation was conducted in site-specific settings, raising concerns regarding generalizability to large-scale power systems. [25] proposed a FireNet-XGBoost hybrid architecture for load forecasting, reporting an R^2 of 0.9334; however, the evaluation was restricted to a single-building dataset, limiting its applicability to broader grid-level forecasting tasks.

In contrast, the proposed ATMH-WNet demonstrates strong and consistent performance while addressing several limitations of prior studies. By leveraging an attention-based Transformer architecture with multi-horizon and weather-aware modeling, ATMH-WNet achieves a low MAPE of 2.9% and a high R^2 of 0.97 on large-scale U.S. power grid test data. Unlike many existing approaches, the proposed framework does not require complex inverse optimization, ensemble tuning, or site-specific calibration. These results highlight the robustness, scalability, and practical suitability of ATMH-WNet for real-world intelligent energy management, positioning it as a competitive and reliable alternative to recent transformer-based and hybrid forecasting models.

4.8 Conclusions

This study presented an Attention-Enhanced Transformer-based Multi-Horizon Weather-aware forecasting framework (ATMH-WNet) for intelligent energy management and load forecasting in large-scale power grid systems. By formulating the forecasting task as a direct multi-horizon regression problem and leveraging the self-attention mechanism of Transformer encoders, the proposed model effectively captures both short-term fluctuations and long-range temporal dependencies present in multivariate

electricity demand data. Unlike traditional recursive forecasting approaches, ATMH-WNet generates all future predictions in a single forward pass, thereby reducing error accumulation and improving computational efficiency.

Extensive experiments conducted on the PJM Interconnection hourly load dataset demonstrate that ATMH-WNet significantly outperforms classical statistical models (Persistence, SARIMA, Prophet), machine learning approaches (XGBoost), and recurrent neural networks (LSTM) across all evaluation metrics. The proposed framework achieves the lowest MAE, RMSE, and MAPE values, along with the highest coefficient of determination, indicating superior predictive accuracy and robustness. Qualitative evaluations further confirm that ATMH-WNet closely tracks real load profiles, preserves peak–valley dynamics, and maintains strong temporal alignment across forecasting horizons. Detailed residual, distributional, and temporal error analyses reveal stable, unbiased, and well-calibrated forecasting behavior under diverse operating conditions. The results validate the effectiveness of integrating attention-based Transformer architectures with weather-aware multivariate inputs for power load forecasting. The proposed ATMH-WNet framework offers a scalable, interpretable, and deployment-ready solution for real-world smart grid applications, supporting improved operational planning, demand response, and energy management strategies. Future research directions include extending the framework to probabilistic forecasting, incorporating additional exogenous variables such as market signals and renewable generation, and evaluating the model across multiple regional power systems to further assess its generalization capability.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
ORCiD ID : https://orcid.org/0009-0003-0555-5646
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Ahmad, Ahmad, et al. "TFTformer: A novel transformer based model for short-term load forecasting." *International journal of electrical power & energy systems* 166 (2025): 110549.
[2] Ahmadian, Amirhossein, and Rajit Gadh. "Smart-Campus EV Charging Demand Forecasting at Multiplexed Chargers Using Operational Data: Multi-Horizon Benchmarking Across Statistical, Tree-Based, Deep Learning, Attention-Augmented, Transformer, and Heterogeneous Ensemble Families." *Tree-Based, Deep Learning, Attention-Augmented, Transformer, and Heterogeneous Ensemble Families*.
[3] Ait Chaoui, Kaoutar, et al. "A Wavelet–Attention–Convolution Hybrid Deep Learning Model for Accurate Short-Term Photovoltaic Power Forecasting." *Forecasting* 7.3 (2025): 45.
[4] Bai, Zhuohao. "Residential electricity prediction based on GA-LSTM modeling." *Energy Reports* 11 (2024): 6223-6232.
[5] Bâra, Adela, and Simona-Vasilica Oprea. "Transformer-based forecasting with synthetic input data generation for day-ahead electricity markets." *Journal of King Saud University Computer and Information Sciences* 37.8 (2025): 233.
[6] Box, George EP, et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
[7] Cao, Hongmei. "Hypergraph neural network and temporal convolutional network for load forecasting in smart construction sites." *Journal of Computational Methods in Sciences and Engineering* (2025): 14727978251385143.
[8] Chen, Tianqi. "XGBoost: A Scalable Tree Boosting System." *Cornell University* (2016).
[9] Dong, Qi, et al. "Short-Term Electricity-Load Forecasting by deep learning: A comprehensive survey." *Engineering Applications of Artificial Intelligence* 154 (2025): 110980.
[10] Dou, Wenlei, et al. "Load prediction and optimization of main transformer based on EEMD-BP neural network." *Discover Computing* 28.1 (2025): 57.
[11] Du, Yu, et al. "A Short-Term User-Side Load Forecasting Method Based on the MCPO-VMD-FDFE Decomposition-Enhanced Framework." *Electronics* 14.18 (2025): 3611.
[12] Dubey, Parul, Pushkar Dubey, and Pitshou N. Bokoro. "Transformer-Driven Fault Detection in Self-Healing Networks: A Novel Attention-Based Framework for Adaptive Network Recovery." *Machine Learning and Knowledge Extraction* 7.3 (2025): 67.
[13] Giacomazzi, Elena, Felix Haag, and Konstantin Hopf. "Short-term electricity load forecasting using the temporal fusion transformer: Effect of grid hierarchies and data sources." *Proceedings of the 14th ACM international conference on future energy systems*. 2023.
[14] Hasan, Mohamed Mahmoud, et al. "TSB-Forecast: A Short-Term Load Forecasting Model in Smart Cities for Integrating Time Series Embeddings and Large Language Models." *IEEE Access* (2025).
[15] He, Hao, et al. "Research on New Energy Power Generation Forecasting Method Based on Bi-LSTM and Transformer." *Energies* 18.19 (2025): 5165.
[16] Hertel, Matthias, et al. "Transformer training strategies for forecasting multiple load time series." *Energy Informatics* 6.Suppl 1 (2023): 20.
[17] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
[18] Hou, Kaiyuan, et al. "Short-term load forecasting based on multi-frequency sequence feature analysis and multi-point modified FEDformer." *Frontiers in Energy Research* 12 (2025): 1524319.
[19] Hussain, Adil, et al. "Charging stations demand forecasting using LSTM based hybrid transformer model." *Scientific Reports* 15.1 (2025): 36639.

[20] Hussain, Ayaz, et al. "Hybrid ML/DL Approach to Optimize Mid-Term Electrical Load Forecasting for Smart Buildings." *Applied Sciences* 15.18 (2025): 10066.
[21] Kim, Tae-Geun, Sung-Guk Yoon, and Kyung-Bin Song. "Very Short-Term Load Forecasting Model for Large Power System Using GRU-Attention Algorithm." *Energies* 18.13 (2025): 3229.
[22] Llorente, Oscar, and Jose Portela. "A Transformer approach for Electricity Price Forecasting." *arXiv preprint arXiv:2403.16108* (2024).
[23] Moustati, Imane, and Noreddine Gherabi. "Unveiling the Potential of Transformer-Based Models for Efficient Time-Series Energy Forecasting." *Journal of Advances in Information Technology* 16.5 (2025).
[24] Mulla, R., 2018. Hourly energy consumption. Kaggle Dataset. URL: https://www.kaggle.com/datasets/robikscube/hourly-energy-consumption. accessed: 2024-05-22.
[25] Syroka, Joanna, and Ralf Toumi. "Scaling and persistence in observed and modeled surface temperature." *Geophysical research letters* 28.17 (2001): 3255-3258.
[26] Mushref, Omer, Sefer Kurnaz, and HAMEED FARHAN. "A Hybrid ANFIS-Transformer Framework Tuned by Enhanced HawkFish Optimization for Voltage and Load Balancing in Smart Grids." *Frontiers in Energy Research* 13: 1654803.
[27] Nguyen, Tuan Anh, and Thanh Ngoc Tran. "A Hybrid Transformer-MLP Approach for Short-Term Electric Load Forecasting." *Journal of Robotics and Control (JRC)* 6.4 (2025): 2033-2044.
[28] Orji, Ugochukwu, Çiçek Güven, and Dan Stowell. "Enhanced Load Forecasting with GAT-LSTM: Leveraging Grid and Temporal Features." *arXiv preprint arXiv:2502.08376* (2025).
[29] Özen, Serkan. Developing Hybrid Deep Learning Models with Data Fusion Approach for Electricity Consumption Forecasting. Diss. Middle East Technical University (Turkey), 2023.
[30] Özen, Serkan, Adnan Yazıcı, and Volkan Atalay. "Hybrid deep learning models with data fusion approach for electricity load forecasting." *Expert Systems* 42.2 (2025): e13741.
[31] Peng, Fang, et al. "Climate-adaptive energy forecasting in green buildings via attention-enhanced Seq2Seq transfer learning." *Scientific Reports* 15.1 (2025): 31829.
[32] Saeed, Faisal, et al. "SmartFormer: Graph-based transformer model for energy load forecasting." *Sustainable Energy Technologies and Assessments* 73 (2025): 104133.
[33] Shen, Boyuan, et al. "Large language model-based security situation awareness for smart grid: Framework and approaches." *IEEE Access* (2025).
[34] Siddiqa, Ayesha, et al. "Explaining solar forecasts with generative AI: A two-stage framework combining transformers and LLMs." *PLoS One* 20.9 (2025): e0331516.
[35] Sievers, Jonas, and Thomas Blank. "Secure short-term load forecasting for smart grids with transformer-based federated learning." *2023 International Conference on Clean Electrical Power (ICCEP)*. IEEE, 2023.
[36] Singh, Arvind R., et al. "A deep learning and IoT-driven framework for real-time adaptive resource allocation and grid optimization in smart energy systems." *Scientific Reports* 15.1 (2025): 19309.
[37] Singh, Harendra Pratap, et al. "A Novel Attention-Augmented LSTM (AA-LSTM) Model for Optimized Energy Management in EV Charging Stations." *Computers, Materials & Continua* 84.3 (2025).
[38] Tan, Shiyu, Yuhao Yang, and Yongxin Zhang. "Short-term power load forecasting using informer encoder and bi-directional LSTM." *E3S Web of Conferences*. Vol. 522. EDP Sciences, 2024.
[39] Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." *The American Statistician* 72.1 (2018): 37-45.
[40] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
[41] Wang, Yuheng, et al. "MTCAT: A Modern Temporal Convolution and Enhanced Attention Transformer Model for Remaining Useful Life Prediction of Aerospace Self-Lubricating Bearings." *IEEE Sensors Journal* (2025).
[42] Wu, Haixu, et al. "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting." *Advances in neural information processing systems* 34 (2021): 22419-22430.
[43] Xiao, Zhiguo, et al. "Time-Series Forecasting Method Based on Hierarchical Spatio-Temporal Attention Mechanism." *Sensors* 25.13 (2025): 4001.
[44] Yu, Shaohua, et al. "An Ensemble Model of Attention-Enhanced N-BEATS and XGBoost for District Heating Load Forecasting." *Energies* 18.15 (2025): 3984.
[45] Zaman, Khalid, et al. "FTDGT: Federated Temporal Dense Granular Transformer-Based Wind Power Forecasting in Medium and Long Term." *International Journal of Energy Research* 2025.1 (2025): 9377203.
[46] Zhao, Yitao, et al. "Short-Term Residential Load Forecasting Based on Generative Diffusion Models and Attention Mechanisms." *Energies* 18.23 (2025): 6208.
[47] Zhou, Haoyi, et al. "Informer: Beyond efficient transformer for long sequence time-series forecasting." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. No. 12. 2021.
[48] Zhou, Tian, et al. "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting." *International conference on machine learning*. PMLR, 2022.
[49] Zhu, Jizhong, et al. "Short-Term Residential Load Forecasting Based on $ K-\text {shape} $ Clustering and Domain Adversarial Transfer Network." *Journal of Modern Power Systems and Clean Energy* 12.4 (2024): 1239-1249.
[50] Zhu, Li, et al. "Short-term power load forecasting based on spatial-temporal dynamic graph and multi-scale Transformer." *Journal of Computational Design and Engineering* 12.2 (2025): 92-111.