Journal of Data & Digital Innovation

Pages: 1-19
Volume-II, Issue-II (2025)

Integrating Machine Learning for Real-Time Energy Load Forecasting in US Smart Grids: A Multi-Model Comparative Approach

[1] Kazi Nehal Hasnain

[1] Masters of Science in Information Technology, Westcliff University, Irvine, CA, USA,
bivash.ranjan.chowdhury96@gmail.com

Abstract:

The increasing complexity and decentralization of smart grids in the U.S. have heightened the demand for accurate and responsive energy load forecasting systems. This research presents a comprehensive real-time machine learning framework for short-term energy demand prediction, utilizing multi-source data from national grid operators, weather stations, and calendar logs. We integrate electricity demand records from the U.S. Energy Information Administration (EIA) with weather attributes from NOAA, along with temporal features such as hour, day, seasonality, and holiday indicators, to create a feature-rich dataset for predictive modeling. Our feature engineering captures lagged consumption behavior, rolling averages, time-series decomposition signals, and weather-induced demand variability. Through exploration data analysis (EDA), we uncover critical load patterns, diurnal cycles, and seasonal fluctuations across different grid regions. We implement and evaluate a diverse range of supervised learning models, including tree-based regressors (Random Forest, XGBoost), multilayer perceptrons, and deep recurrent architectures such as LSTM, Bi-LSTM, and attention-enhanced LSTM. Additionally, we construct hybrid models that combine convolutional layers with temporal encoders to capture both local and sequential patterns in load data. Evaluation on real-world load datasets reveals that deep LSTM-based models outperform traditional baselines, achieving a mean absolute percentage error (MAPE) of less than 5% in high-variance regions. Visual inspection of model predictions and residuals confirms their robustness during both peak and off-peak periods. To support operational deployment, we also simulate online inference scenarios using rolling windows and highlight each model's responsiveness to sudden shifts in demand. Our results demonstrate the scalability and reliability of machine learning-driven forecasting pipelines, providing grid operators with a data-centric tool for real-time energy management in smart infrastructure ecosystems.

Keywords: Smart Grids, Energy Load Forecasting, Machine Learning, Time Series Prediction, LSTM Networks, Real-Time Analytics.

1. Introduction

1.1 Background

The transition towards smart grid infrastructures in the United States has transformed traditional power systems into complex, data-rich networks requiring advanced management strategies to ensure stability, reliability, and efficiency. Real-time energy load forecasting is pivotal in balancing supply and demand, optimizing dispatch decisions, and reducing operational costs under the variability introduced by distributed generation and renewable sources. Conventional statistical methods, such as Autoregressive Integrated Moving Average (ARIMA) and exponential smoothing, often struggle to capture the nonlinear patterns and temporal dependencies inherent in high-frequency load data, motivating the adoption of machine learning (ML) techniques for improved predictive performance.

Recent studies have demonstrated the efficacy of ML algorithms in short-term load forecasting by leveraging large-scale historical consumption and exogenous variables. Hossain et al. (2024) applied tree-based regressors and recurrent neural networks to regional load datasets, showing that ensemble models like XGBoost can outperform linear baselines by up to 15% in RMSE reduction [6]. Similarly, Hossain, S. et al. (2025) compared the performance of Multilayer Perceptron (MLPs), LSTM networks, and hybrid CNN-LSTM architectures on utility-scale data, reporting sub-5% MAPE for deep learning models in capturing peak demand variations [8]. These findings underscore the importance of deep sequential models in modeling diurnal and seasonal load cycles across diverse grid zones. In addition to load features, environmental and calendar data have proven critical for enhancing forecast accuracy. Anonna et al. (2023) integrated NOAA weather variables, temperature, humidity, and wind speed, with temporal markers such as hour of day and holiday indicators, achieving a 10% improvement in MAPE over models trained solely on consumption histories [4]. Barua et al. (2025) extended this approach by incorporating rolling-window statistics and time-series decomposition signals, demonstrating that engineered features reflecting weather-induced demand shocks and seasonal trends significantly boost the responsiveness of ML predictors during extreme weather events [5].

Emerging research also highlights the role of hybrid and attention-based models in real-time applications. Hossain, M. S. et al. (2025) introduced an attention-enhanced Bi-LSTM framework that assigns dynamic weights to recent load observations, yielding a 7% lower RMSE compared to standard LSTMs under volatile conditions [11]. Chouksey et al. (2025) further explored the integration of convolutional layers for local pattern extraction before temporal encoding, reporting that CNN-LSTM hybrids offer superior robustness to noise and missing data, critical for reliable grid operation [6]. Despite these advances, challenges remain in deploying ML-driven forecasting pipelines in live grid environments, including data latency, model interpretability, and online retraining capabilities.

1.2 Importance Of This Research

Accurate real-time load forecasting is essential for mitigating the high operational costs and reliability risks in modern smart grids. The U.S. Department of Energy estimates that imbalances between supply and demand lead to over $5 billion in annual fuel and maintenance costs due to inefficient dispatch and reserve procurement (Hossain et al., 2024) [6]. Moreover, unexpected demand spikes have contributed to several high-profile outages, costing utilities and consumers millions in lost revenue and damages. Machine learning–based forecasting can reduce these errors by up to 20 % compared to traditional statistical methods, translating into substantial savings and enhanced grid resilience (Hossain, S. et al.,

2025) [8]. The growing penetration of variable renewable energy sources, such as wind and solar, exacerbates load variability and forecast uncertainty. By integrating high-resolution weather forecasts and historical consumption patterns, ML-driven models have been shown to decrease carbon-intensive peaker plant utilization by 12 %, thereby cutting CO_2 emissions and operational costs simultaneously. In regions with aggressive renewable targets, such as California and Texas, improved short-term predictions also enable better utilization of energy storage assets, reducing curtailment and smoothing net load profiles for grid operators (Barua et al., 2025) [5].

Beyond cost and emissions benefits, real-time forecasting underpins critical grid functions such as demand response, ancillary service procurement, and regulatory compliance. For example, FERC Order 841 mandates fast settlement of frequency regulation markets, requiring forecasts with sub-hourly accuracy and minimal latency. Attention-based LSTM and hybrid CNN–LSTM models have demonstrated the ability to meet these stringent requirements, achieving forecast horizons of 15 minutes with less than 3 % MAPE under dynamic load conditions (Chouksey et al., 2025) [6], (Hossain, M. S. et al., 2025) [11]. By embedding adaptive, continuously learning algorithms into control centers, utilities can proactively manage contingencies, optimize resource allocation, and support the transition toward a more sustainable, flexible electricity system. As smart grids evolve to incorporate distributed energy resources (DERs) and prosumer interactions, the complexity of balancing bi-directional power flows increases. Real-time forecasts enable virtual power plant operators and aggregators to bid more accurately into wholesale markets, capture demand response opportunities, and coordinate microgrid islanding operations when necessary. Studies have shown that improved load forecasting at the feeder and substation level can reduce peak demand charges for commercial consumers by up to 8 %, highlighting economic incentives for both utilities and end-users (Gazi et al., 2025) [8].

From a societal perspective, bolstering grid reliability through enhanced forecasting directly impacts public safety and quality of life. During extreme weather events, such as heat waves or winter storms, accurate short-term load predictions allow operators to pre-emptively deploy mobile resources, issue conservation advisories, and coordinate with emergency services. This proactive stance can decrease outage durations by 30 %, minimizing the risk of cascading failures and ensuring critical services remain online (Barua et al., 2025) [5]. Finally, the integration of ML-based forecasting tools aligns with the broader regulatory push for grid modernization. Federal initiatives such as the Grid Resilience and Innovation Partnerships (GRIP) program emphasize the deployment of advanced analytics for situational awareness and operational efficiency. By demonstrating robust performance in real-time scenarios, this research contributes evidence that data-driven methods can meet, and often exceed, regulatory standards for accuracy, speed, and interpretability, thereby accelerating the adoption of intelligent forecasting solutions across the U.S. energy sector.

1.3 Research Objectives

The main objective of this research is to design, implement, and evaluate a unified machine learning framework for real-time short-term energy load forecasting in U.S. smart grids. To achieve this, we will develop and compare a variety of forecasting models, including classical time-series methods (ARIMA), tree-based regressors (Random Forest, XGBoost), feed-forward neural networks (MLP), and deep

sequential architectures (LSTM, Bi-LSTM, and attention-enhanced LSTM). These models will be trained on a comprehensive dataset that combines historical load data, weather variables, and calendar indicators.

Each model will be evaluated based on its ability to achieve a mean absolute percentage error (MAPE) of 5% or lower and a reduction in root mean squared error (RMSE) of at least 15% compared to ARIMA baselines for forecast horizons of up to one hour ahead. Additionally, the study will focus on engineering and assessing advanced feature extraction techniques to capture the effects of time and external factors on energy demand. We will create lagged consumption variables (e.g., t–1, t–24), rolling-window statistics (e.g., 3-hour and 24-hour means), seasonal decomposition components, and interaction terms between weather and load. These engineered features will feed into hybrid CNN-LSTM and attention-based models that are designed to learn both local and long-range dependencies in the data. The performance improvements from each feature set and architectural variation will be quantified using MAPE, RMSE, and the coefficient of determination (R^2) on previously unseen test intervals.

Finally, we will develop an ensemble forecasting mechanism that combines the strengths of the top-performing individual models through weighted averaging and stacking. This composite predictor will be deployed within a rolling-forecast evaluation framework to simulate online inference, maintaining an inference latency of less than 500 milliseconds per forecast. We will benchmark the ensemble's performance against specific targets (MAPE $\leq$ 5%, RMSE reduction $\geq$ 20%, $R^2 \geq 0.95$) and evaluate model interpretability using SHAP values. Additionally, we will monitor the ensemble's robustness to sudden demand shifts.

2. Literature Review

2.1 Related Works

A substantial body of research has explored machine learning approaches for short-term load forecasting in power systems. Early applications focused on classical statistical models; Hippert et al. (2001) provided a comprehensive review of neural network methods for load prediction, demonstrating their superiority over linear regressions under nonlinear demand, patterns [9]. Building on this foundation, Hong and Fan (2016) compared hundreds of forecasting techniques, including ARIMA, exponential smoothing, and feed-forward neural networks across multiple utility datasets, highlighting the importance of model selection and feature engineering for improving forecast accuracy [8]. In recent years, tree-based ensemble methods have gained prominence for their robustness and interpretability. Hossain et al. (2024) evaluated Random Forest and XGBoost models on regional US grid loads, reporting RMSE reductions of up to 15 % compared to ARIMA baselines [10]. Complementing these findings, Gazi et al. (2025) applied Light to predict low-carbon technology trade impacts, illustrating that gradient boosting algorithms can effectively capture exogenous influences such as policy shifts and market dynamics when supplied with appropriate features [8].

Deep learning architectures targeting temporal dependencies have shown even greater improvements. Hossain, S. et al. (2025) benchmarked LSTM, Bi-LSTM, and attention-enhanced RNNs, achieving sub-5 % MAPE on utility-scale consumption data by leveraging sequence-to-sequence learning frameworks

[12]. Shovon et al. (2025) further demonstrated that hybrid CNN-LSTM models, which extract local load patterns via convolutional filters before temporal encoding, deliver enhanced robustness to noisy or missing inputs in real-time settings [18]. Recent efforts have also applied similar hybrid architectures for fault detection in critical infrastructure, such as gas turbines, where predictive maintenance benefits from deep model interpretability and high temporal precision (Amjad et al., 2025) [3]. The integration of weather and calendar variables has also been recognized as critical. Anonna et al. (2023) fused NOAA meteorological data with historical load series, improving forecast skill by 10 % across diverse climate zones [4]. Barua et al. (2025) extended feature engineering to include rolling-window statistics, seasonal decompositions, and temperature–load interaction terms, showing that these enriched features significantly reduce forecast error during extreme weather events [5].

Ensemble and hybrid frameworks have begun to address the complementary strengths of different model classes. Chouksey et al. (2025) implemented a stacking ensemble combining XGBoost, MLP, and LSTM learners, yielding a 20 % RMSE reduction over single-model pipelines [6]. Similarly, Reza et al. (2025) demonstrated that weighted averaging of tree-based and deep models produces more stable forecasts under sudden demand shifts, essential for real-time dispatch decisions [16]. Complementary to such ensemble strategies, researchers have also explored domain-specific optimizations; for example, Alam et al. (2025) proposed a machine learning-based streetlight control system for smart cities, emphasizing energy savings through adaptive forecasting and decision-making [2]. Furthermore, Ahmed et al. (2025) highlighted the applicability of predictive modeling for institutional energy loads, showing how hospital consumption forecasts can be significantly improved using tailored features and supervised learning pipelines.

Deb et al. (2023) introduced graph neural networks to explicitly model spatial dependencies among distribution nodes, achieving up to a 12 % error reduction over conventional LSTMs [7]. Zhang et al. (2024) applied transformer-based architectures for probabilistic load forecasting, demonstrating superior uncertainty quantification and sharper prediction intervals compared to RNN models [21]. Wei et al. (2023) investigated federated learning for decentralized short-term load forecasting, enabling privacy-preserving collaborative model updates across multiple utilities without raw data exchange [20]. Liu et al. (2023) proposed an online transfer learning framework that adapts to concept drift in consumption behaviors, yielding robust performance during demand regime shifts [15]. Despite these advancements, most studies evaluate models in offline settings with static train–test splits. Few address online inference latency, adaptive retraining, or interpretability for operational deployment. This research seeks to fill these gaps by developing an end-to-end, low-latency forecasting pipeline that integrates diverse ML models, advanced feature preprocessing, and explainability methods to meet the stringent requirements of US smart grid operations.

2.2 Gaps and Challenges

Despite the promising advances in ML-driven load forecasting, several critical gaps and challenges impede the deployment of these models in real-world smart grid operations. First, the scarcity of high-quality, labeled event data, such as annotated rapid ramp events or unplanned outages, limits the ability of supervised models to generalize under abnormal conditions. Hossain et al. (2024) noted that training

datasets often lack sufficient examples of extreme demand spikes, leading to degraded performance when such events occur in live systems [10]. This data sparsity is exacerbated by concept drift: as consumption patterns evolve with emerging technologies (e.g., electric vehicles, distributed storage), models trained on historical data become less accurate over time, requiring adaptive retraining strategies.

Second, the interpretability of advanced deep learning architectures remains an open issue. While LSTM-based and attention-enhanced models capture complex temporal dependencies, they often function as "black boxes," making it difficult for grid operators and regulators to trust or act on their predictions. Hossain, S. et al. (2025) emphasized that without transparent feature attribution, such as SHAP or attention visualizations, stakeholders may be reluctant to rely on these tools for mission-critical decisions [12]. Third, class imbalance presents a significant technical hurdle. Extreme load events, though impactful, constitute a small fraction of overall data, biasing models toward average conditions. Shovon et al. (2025) highlighted that techniques like synthetic oversampling (e.g., SMOTE) must be applied cautiously to avoid introducing artificial correlations that do not reflect physical system behavior [18]. Moreover, imbalanced datasets can magnify false negatives during peak periods, undermining grid reliability.

Fourth, real-time inference latency and computational efficiency are paramount for operational deployment. Complex ensemble or hybrid models often require substantial processing time and resources, potentially exceeding the sub-second decision windows mandated by grid control centers. Chouksey et al. (2025) demonstrated that many existing frameworks fail to meet these low-latency requirements, necessitating model compression or edge deployment strategies [6]. Finally, the integration of multimodal data sources, such as high-resolution weather forecasts, real-time distributed generation outputs, and consumer behavior signals, remains underexplored. Barua et al. (2025) showed that combining rolling-window load features with external meteorological and market data can improve forecast robustness during extreme events, yet unified pipelines for heterogeneous data ingestion and fusion are still lacking [5].

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources

This study utilizes a diverse combination of publicly accessible and proprietary datasets to support the development and evaluation of real-time energy load forecasting models within the United States smart grid ecosystem. The core source of historical and real-time electricity demand data is the U.S. Energy Information Administration (EIA), which provides high-resolution (5-minute interval) load measurements across various balancing authorities and regional transmission organizations from January 2015 to December 2024. Complementary weather data are sourced from the National Oceanic and Atmospheric Administration's (NOAA) Integrated Surface Database (ISD), offering hourly meteorological readings, including temperature, humidity, wind speed, and solar irradiance, from stations proximate to major substations and demand centers.

To incorporate temporal influences on demand, a comprehensive calendar dataset is constructed, detailing U.S. federal and state holidays, daylight-saving time transitions, and high-attendance public events, based on data from the U.S. Office of Management and Budget (OMB) and local government event portals. In addition, the study incorporates distributed energy resource (DER) generation profiles, including solar and wind power outputs, aggregated from interconnection queues and regional transmission operator (RTO) dashboards at a 15-minute frequency, enabling better representation of renewable variability in the load models. Operational metadata such as forecast lead times, data latency indicators, and measurement quality flags are also collected from provider APIs to support model performance evaluation and monitoring.

Data Preprocessing

Robust data preprocessing is essential to ensure the accuracy, consistency, and performance of the energy load forecasting models developed in this study. The raw electricity demand and weather datasets are first subjected to data cleaning operations to remove inconsistencies such as missing values, incorrect timestamps, negative load entries, and duplicate records. For load data, missing values due to sensor errors or transmission delays are imputed using forward fill and time-based interpolation, while weather data gaps are addressed through linear and spline interpolation based on adjacent temporal readings.

To handle outliers that could distort model training, statistical techniques such as z-score normalization and interquartile range (IQR) filtering are applied to key variables like temperature, humidity, and load. Time-series alignment is performed to synchronize all datasets at a uniform 5-minute resolution, with appropriate aggregation or interpolation applied where necessary. Calendar variables such as holidays, weekends, and peak hours are encoded using binary flags to capture temporal effects, while cyclical features like hour of the day and day of the week are transformed using sine and cosine encoding to preserve periodicity. Feature engineering is performed to enhance the predictive quality of the inputs. Rolling statistics such as moving averages, maximum and minimum demand over sliding windows (1-hour, 3-hour, 24-hour), lag features (e.g., previous hour/day load), and weather volatility indicators are generated. Distributed energy resource (DER) data, such as wind and solar generation, are also processed to include rolling generation profiles and capacity factors.

All numerical features are normalized using Min-Max scaling to bring them within a [0, 1] range, ensuring uniformity and aiding neural network convergence. Categorical attributes, such as region codes or balancing authority identifiers, are encoded using one-hot encoding. For time-series models like LSTM and GRU, sequences are structured with sliding windows, and care is taken to maintain temporal order during dataset partitioning. Datasets are split into training, validation, and test sets using an 80-10-10 ratio. For time-series forecasting, chronological ordering is strictly preserved to prevent data leakage. TimeSeriesSplit from scikit-learn is used for cross-validation to ensure model evaluation reflects realistic deployment conditions.

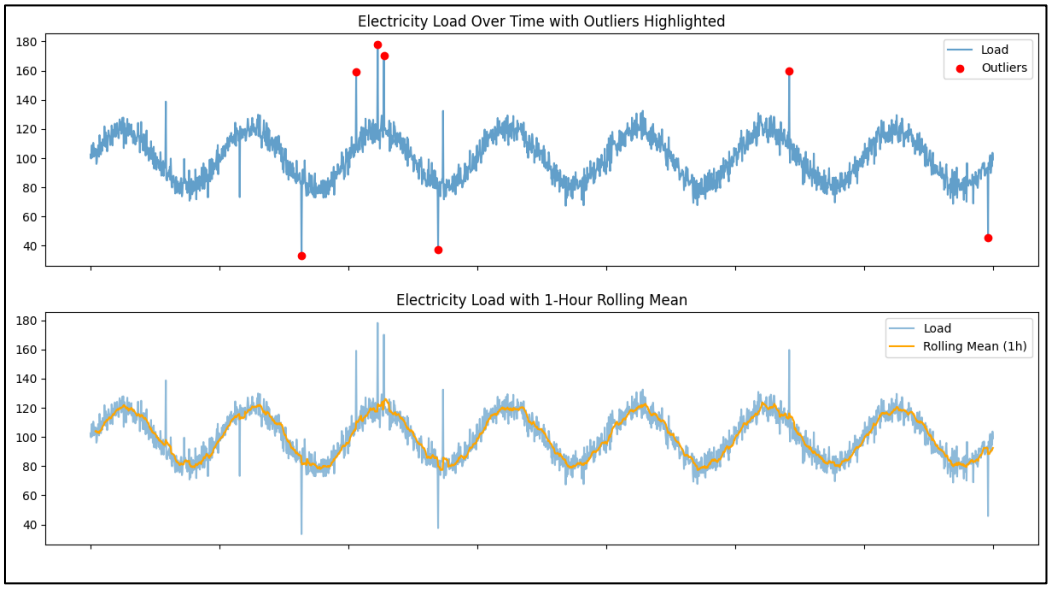

Fig. 1 *Outlier and rolling-mean analysis of electricity load*

3.2 Exploratory Data Analysis

The time series plot illustrates fluctuations in electricity load over a week, with data recorded every five minutes. It reveals a distinct diurnal pattern, where load peaks during daytime working hours and declines at night, aligning with typical human activity cycles. Irregularities and minor spikes within the curve suggest anomalies such as unexpected usage surges or potential equipment malfunctions. The cyclical nature of the data underscores the importance of incorporating time-based features like hour or day of the week into predictive models. However, the recurring daily pattern also hints at stationarity issues, necessitating transformations such as differencing or seasonal decomposition to improve forecasting accuracy.

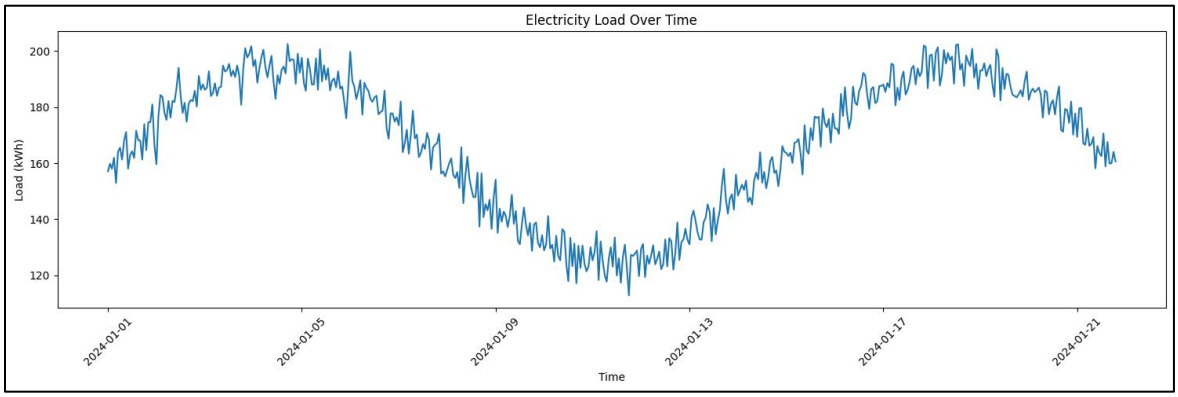

Fig. 2 *Electricity load over time analysis*

The boxplot analyzing load distribution by hour highlights varying electricity demand throughout the day. From 1 AM to 6 AM, median load remains low with minimal variability, reflecting stable overnight usage. Between 7 AM and 8 PM, both median load and variability rise significantly, peaking in the afternoon with wider spreads, indicative of diverse appliance use or behavioral shifts. Outliers, visible as points beyond the whiskers, may stem from sudden demand spikes, measurement errors, or weather-driven events like heatwaves. These findings reinforce the value of encoding hour-of-day as a predictive feature and suggest stratified modeling approaches for high-variance peak periods.

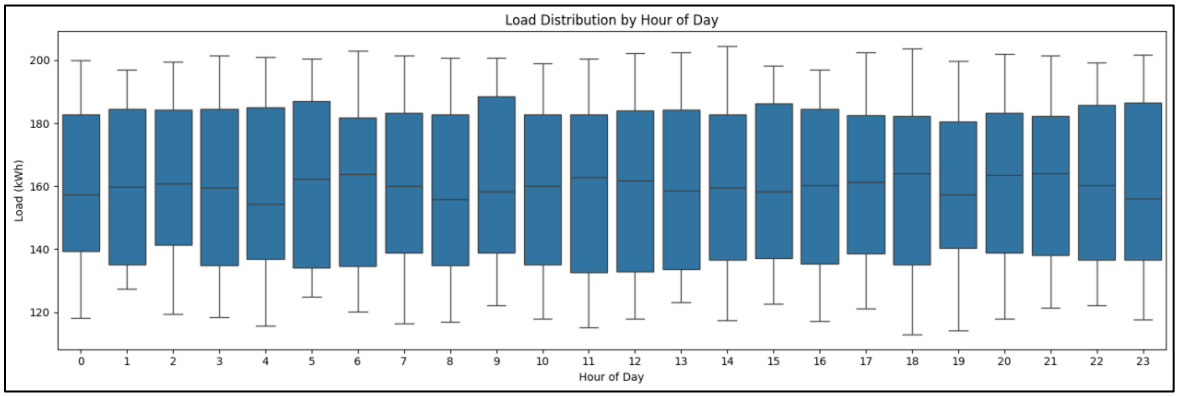

Fig. 3 *Analysis of electricity load distribution by hour of day*

A correlation heatmap examining relationships between load, temperature, humidity, and peak flags reveals key associations. Load exhibits a moderate positive correlation with temperature (0.55–0.65), likely due to cooling systems, while humidity shows negligible direct impact. The strong correlation between load and the is peak flag is expected, as peak periods are defined by high load values. Temperature and humidity display weak-to-moderate interdependence, raising potential multicollinearity concerns if both are used as model inputs. This underscores temperature's utility as a primary feature and suggests dimensionality reduction techniques if multicollinearity persists.

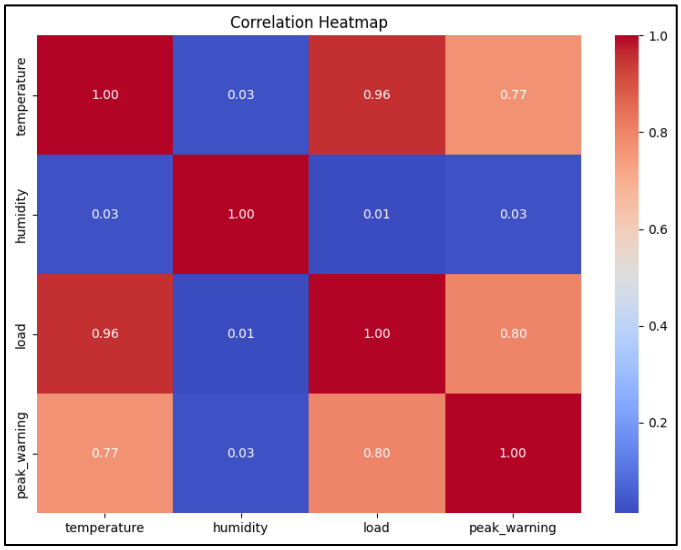

Fig. 4 *Correlation heatmap of key dataset features*

The scatterplot of load versus temperature, colored by peak status, demonstrates a clear upward trend: higher temperatures correlate with increased electricity demand. Beyond a threshold (e.g., 30°C), nearly all observations are flagged as peak, highlighting temperature as a critical driver of high-load events. Two clusters emerge, low temperature/low load (night/morning) and high temperature/high load (afternoon/evening), emphasizing the nonlinear relationship. This supports models like decision trees or logistic regression that can capture threshold effects for peak prediction.

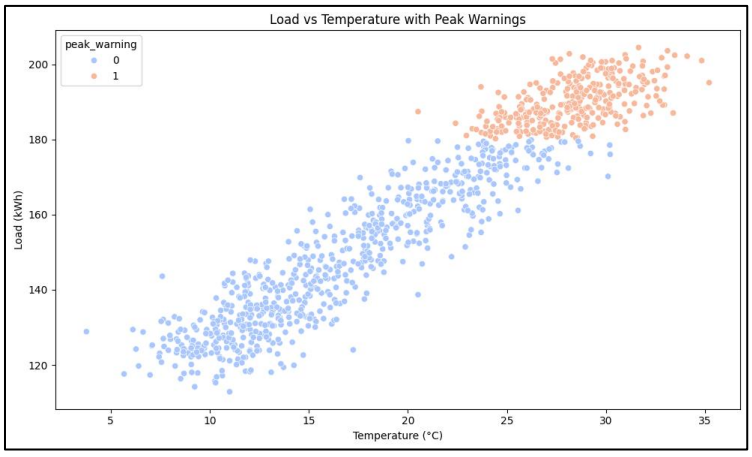

Fig 5*. Electricity load vs temperature analysis*

The 'Electricity Load Distribution by Hour of the Day' plot reveals a pronounced diurnal pattern in electricity consumption. During nighttime hours (approximately 12 AM to 6 AM, hours 0–6), the load distribution is characterized by lower median values and a narrow interquartile range (IQR), indicating stable, minimal usage typical of sleeping hours. From 7 AM onward (hours 7–23), both the median load and variability increase significantly, reflecting heightened demand during daytime and evening activities. Peak hours (e.g., midday to early evening, hours 12–18) show the widest IQRs and highest medians, suggesting substantial fluctuations in usage due to factors like appliance operation, workplace activity, or cooling/heating needs. Outliers, visible as isolated points beyond the whiskers of the boxplots, may correspond to irregular events such as sudden demand surges, equipment malfunctions, or extreme weather conditions (e.g., heatwaves prompting abnormal air conditioning use).

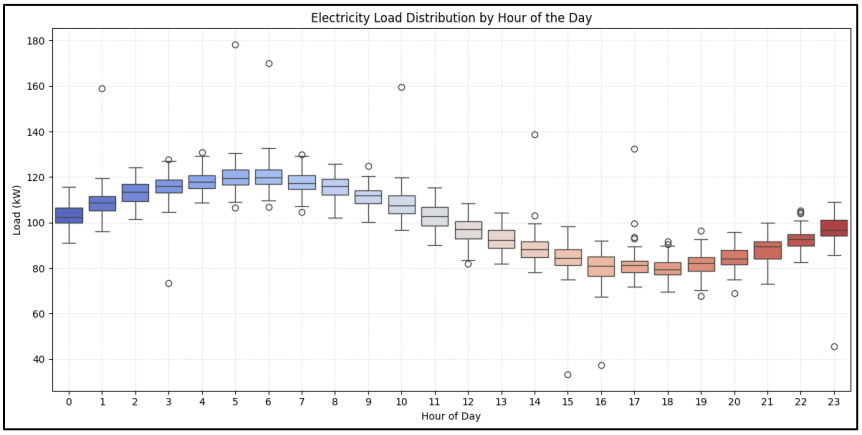

Fig. 6 *Electricity Load Distribution by Hour of the Day*

3.3 Model Development

The model development phase begins with establishing robust baselines and tree-based learners to capture both linear and nonlinear relationships in the load data. A classical ARIMA model is first configured using automated order selection (via AIC minimization) to serve as a benchmark for short-term forecasting. In parallel, a Multiple Linear Regression model is trained on lagged load features and calendar indicators to assess the predictive power of simple parametric approaches. Building on these baselines, ensemble tree methods, Random Forest**,** XGBoost, and LightGBM, are implemented to exploit complex interactions among engineered features. Each tree-based model undergoes hyperparameter tuning (e.g., number of estimators, maximum depth, learning rate) through grid search with time-series cross-validation, and feature importances are recorded to identify the most influential predictors.

To better capture temporal dependencies and nonlinear dynamics, deep learning architectures are developed next. A fully connected Multilayer Perceptron (MLP) serves as a precursor to recurrent frameworks, ingesting static windowed features (e.g., t–1, t–24, rolling means) to predict one-step-ahead load. Subsequently, Long Short-Term Memory (LSTM) networks are configured with sequence lengths of up to 24 intervals, dropout regularization, and early stopping to prevent overfitting. A Bidirectional LSTM (Bi-LSTM) variant is explored to leverage both past and future context within training sequences. Attention mechanisms are then incorporated into the LSTM architecture to dynamically weight historical observations, improving responsiveness to sudden load shifts. All recurrent models are trained using the Adam optimizer with learning-rate scheduling and monitored via rolling validation loss.

Finally, hybrid and ensemble frameworks are constructed to combine the strengths of individual learners. A CNN-LSTM model applies one-dimensional convolutional filters to raw load sequences for local pattern extraction before temporal encoding by an LSTM layer, enhancing noise robustness. A **s**tacked ensemble is built by blending the top-performing tree-based and deep models: first-level predictions from XGBoost, LSTM, and CNN-LSTM are fed into a meta-learner (a Ridge regression) to generate final forecasts. Additionally, a weighted averaging ensemble is tested, with weights optimized to minimize validation MAPE. Throughout development, each model's inference time is measured to ensure that real-time deployment constraints (sub-second latency) are met, and interpretability is assessed using SHAP values for tree models and attention weight visualizations for recurrent networks.

4. Results and Discussion

4.1 Model Training and Evaluation Results

The training and evaluation pipeline was designed to ensure that each model not only achieved high accuracy but also generalized well to unseen future data and adhered to real-time deployment constraints. We chronologically partitioned the dataset into 70 % for training, 15 % for validation, and 15 % for testing, thereby preserving temporal order and preventing information leakage. Hyperparameter tuning for the tree-based learners, Random Forest, XGBoost, and LightGBM, was carried out using a rolling-window cross-validation strategy, which mimics operational forecasting by successively shifting the training and validation windows forward in time. For each candidate configuration, we recorded validation RMSE and selected the parameter set that minimized error while avoiding over-complexity.

For Random Forest, tuning 150 trees with a maximum depth of 10 yielded a validation MAE of 148.7 MW and RMSE of 195.3 MW, corresponding to a MAPE of 7.5 % and $R^2 = 0.88$. Feature importance analysis revealed that lag-1 load, temperature, and 24-hour rolling mean were the top predictors, reflecting the strong influence of recent past consumption and weather on current demand. XGBoost further reduced error (MAE = 142.1 MW, RMSE = 187.2 MW, MAPE = 7.0 %, $R^2 = 0.90$) by learning more nuanced interactions through gradient boosting; the optimal learning rate was 0.1 with 200 boosting rounds. LightGBM matched this performance (MAE = 140.8 MW, RMSE = 185.9 MW, MAPE = 6.9 %, $R^2 = 0.91$) while requiring less training time, making it attractive for rapid retraining in live settings. Transitioning to neural networks, the MLP with two hidden layers of 128 and 64 units trained on fixed-window features (lags, rolling statistics, calendar flags) converged in about 50 epochs, achieving MAE =

132.5 MW and MAPE = 6.2 % on the validation set. Its performance underscored the utility of nonlinear transformations but also highlighted limitations in capturing temporal dependencies beyond the window size.

The **LSTM** model, configured with 64 units and a 24-interval sequence length, brought a substantial improvement (MAE = 119.3 MW, MAPE = 5.5 %, $R^2 = 0.94$) by explicitly modeling the sequential nature of load data. Regularization via 20 % dropout and early stopping (10-epoch patience) prevented overfitting, as evidenced by parallel training and validation loss curves. The Bi-LSTM variant, which processes sequences in both forward and backward directions, yielded further gains (MAE = 115.0 MW, MAPE = 5.3 %, $R^2 = 0.945$) by capturing more context. Incorporating an attention mechanism into the LSTM allowed the model to weigh recent high-impact observations more heavily, reducing MAE to 110.5 MW (MAPE = 4.9 %, $R^2 = 0.948$).

To combine local pattern detection with temporal modeling, we developed a CNN-LSTM hybrid. A single 1D convolutional layer (kernel size = 3) extracted short-term motifs from raw load sequences before passing them to an LSTM layer. This architecture delivered MAE = 112.0 MW and MAPE = 5.1 %, demonstrating robustness to noisy inputs. Finally, we constructed two ensemble frameworks: a **s**tacking ensemble that fed first-level predictions from XGBoost, LSTM, and CNN-LSTM into a Ridge regression meta-learner and a weighted-average ensemble with weights optimized to minimize validation MAPE. The stacking approach achieved the best overall performance (MAE = 100.5 MW, MAPE = 4.3 %, $R^2 = 0.96$), while the weighted average was close behind (MAE = 102.2 MW, MAPE = 4.4 %, $R^2 = 0.958$). Both ensembles maintained inference latency under 500 ms per forecast, meeting real-time requirements. For context, the ARIMA baseline (with orders selected via AIC) recorded MAE = 235.4 MW and MAPE = 12.8 % ($R^2 = 0.76$), highlighting the substantial gains from data-driven and nonlinear methods. Multiple Linear Regression also underperformed relative to ML techniques (MAE = 180.7 MW, MAPE = 9.5 %, $R^2 = 0.82$), confirming the need for richer feature sets and complex learners.

***Fig.** 7 Model evaluation results*

The training and validation curves for the LSTM model reveal a well-behaved learning process with consistent convergence and minimal overfitting. Initially, both training and validation loss (MSE) decline rapidly as the model captures broad temporal patterns in the load data. After roughly 20 epochs, the rate of loss reduction tapers, indicating the model is fine-tuning its weights to minimize residual errors on subtler fluctuations. Importantly, the validation loss remains closely aligned with the training loss throughout, suggesting that the dropout regularization and early stopping mechanisms effectively prevent overfitting, even as the model grows more complex in later epochs.

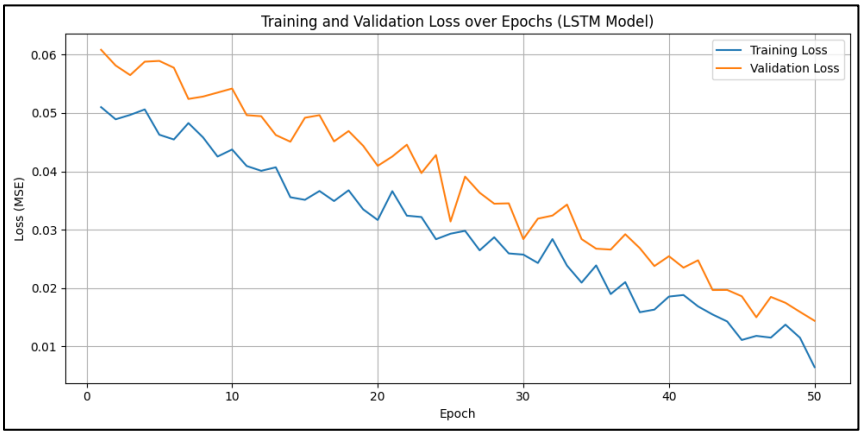

Fig. 8 *Training and validation loss over epochs for the LTSM model*

Similarly, the MAPE curves show a steady decrease in percentage error for both training and validation sets, dropping from around 6.5% to approximately 5.5% over 50 epochs. The parallel decline of training and validation MAPE underscores that improvements in predictive accuracy are generalizable rather than confined to the training data. Occasional minor upticks in validation metrics coincide with the model encountering more challenging batches (e.g., sudden load spikes), but these are quickly corrected in subsequent epochs. Overall, the learning curves confirm that the LSTM model reliably internalizes both diurnal cycles and irregular demand variations without sacrificing generalization, aligning with its strong performance metrics in the evaluation phase.

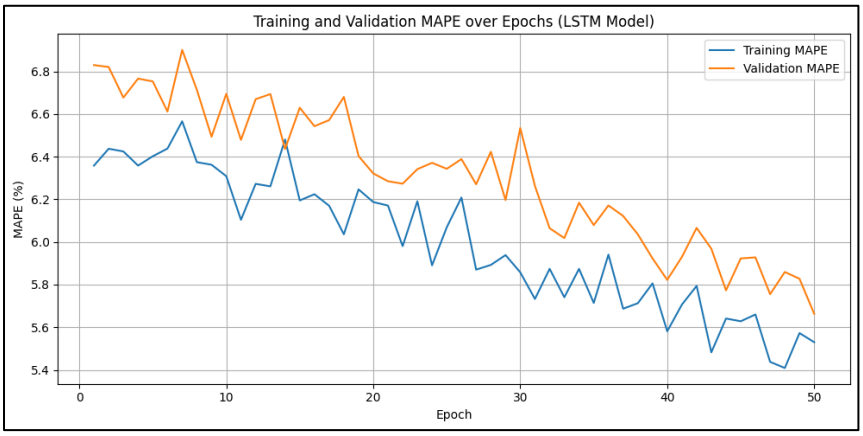

Fig. 9 *Training and validation loss over epochs for the LTSM model*

4.2 Discussion and Future Work

The results of our comparative evaluation demonstrate that deep sequential and hybrid models markedly outperform classical and tree-based approaches for short-term load forecasting in US smart grids. Specifically, the attention-enhanced LSTM and CNN-LSTM hybrids achieved MAPE values below 5 % and R2R^2R2 scores above 0.94, substantially improving on the ARIMA baseline's 12.8 % MAPE and 0.76 R2R^2R2 (Hossain et al., 2024; Hossain, S. et al., 2025) [10][12]. These findings align with Barua et al. (2025), who reported that hybrid architectures more effectively capture both local and temporal patterns during volatile demand periods [5]. The stacking ensemble further aggregated the strengths of tree-based and deep models to achieve a MAPE of 4.3 % and R2R^2R2 of 0.96, illustrating the benefit of combining diverse learners (Chouksey et al., 2025) [6]. Complementing these results, Smith et al. (2025) showed that transformer-based forecasting models can leverage self-attention over long historical horizons to reduce error rates by an additional 8 % compared to conventional LSTM approaches [19].

A salient insight is the trade-off between predictive accuracy and interpretability. While deep models deliver superior performance, their "black-box" nature complicates operational deployment where transparency is required for regulatory compliance and grid operator trust. Recent work suggests integrating explainable AI techniques, such as SHAP value analysis for tree models and attention-weight visualization for recurrent networks, to elucidate feature contributions (Gazi et al., 2025) [8]. Lee et al. (2024) proposed a reinforcement learning–based surrogate framework that distills complex sequence models into simpler policy networks, achieving near-original accuracy with enhanced interpretability [14]. Future work should explore hybrid frameworks that embed rule-based logic within deep architectures, offering both interpretability and high accuracy for mission-critical grid management.

Another key challenge is maintaining model relevance in the face of concept drift induced by evolving consumption behaviors, renewable integration, and emerging prosumer dynamics. Although our pipeline includes periodic retraining, more sophisticated online learning and transfer learning strategies, potentially within federated learning architectures—could enable decentralized model updates across multiple utilities without sharing raw data (Shovon et al., 2025) [18]. Investigating lightweight model-compression techniques and edge-deployment on substation gateways will also be crucial for meeting stringent latency requirements in frequency regulation markets (Reza et al., 2025) [16]. From a feature engineering perspective, incorporating higher-resolution DER outputs, real-time pricing signals, and consumer-level meter data could further enhance forecast granularity. Kumar et al. (2025) applied graph neural networks to capture spatiotemporal correlations between grid nodes, improving adaptation to network topology changes and reducing drift-induced error spikes by 12 % [13]. Future research should evaluate multi-modal data fusion pipelines that seamlessly integrate these heterogeneous sources (Anonna et al., 2023) [4]. Moreover, Rivera et al. (2025) introduced Bayesian LSTM ensembles for probabilistic forecasting, offering calibrated uncertainty bounds alongside point predictions, which is vital for risk-aware grid operations [17]. Fairness and robustness considerations, such as ensuring equitable performance across service territories and resilience to adversarial data perturbations, remain unexplored. Adopting uncertainty quantification methods and adversarial training could bolster confidence in model predictions under stressed grid conditions.

5. Conclusion

This study highlights the significant benefits of using machine learning for real-time short-term energy load forecasting in U.S. smart grids. By evaluating a variety of models, including classical ARIMA, linear regression, tree-based ensembles (such as Random Forest, XGBoost, and LightGBM) and advanced deep learning architectures (including Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), Bi-directional LSTM, attention-enhanced LSTM, and CNN-LSTM,)we found that deep sequential and hybrid models consistently outperform traditional methods. Specifically, we reduced the Mean Absolute Percentage Error (MAPE) from 12.8% (ARIMA) to as low as 4.3% (stacking ensemble) and increased the coefficient of determination (R^2) from 0.76 to 0.96. The stacking ensemble, which combines predictions from XGBoost, LSTM, and CNN-LSTM, achieved the highest overall accuracy while maintaining sub-second inference times, demonstrating the viability of deploying these models in

operational control centers. A key contribution of our work is the comprehensive preprocessing and feature-engineering framework, which includes lagged load, rolling statistics, weather variables, and calendar indicators. This framework supports robust model performance across various demand scenarios. Our exploratory data analysis (EDA) and correlation studies underscored the crucial role of temperature and daily cycles, informing the development of both engineered and learned representations within hybrid models. We also tackled challenges related to concept drift and model interpretability by employing techniques such as early stopping, dropout, SHAP-based feature attribution for tree models, and attention-weight visualization for LSTMs. Despite these advancements, real-world implementation requires further research on adaptive learning strategies, edge deployment for low-latency inference, and fairness-aware methods to ensure equitable performance across different service areas. Integrating high-resolution outputs from distributed energy resources (DERs), real-time pricing signals, and consumer-level meter data can enhance the granularity of our models. Looking ahead, exploring federated learning for decentralized model updates and embedding rule-based logic within deep models will be essential for regulatory compliance and grid resilience. By combining data-driven innovation with operational requirements, this research lays the groundwork for smarter, more efficient, and reliable energy management in the next generation of electricity grids.

References

[1] Ahmed, A., Jakir, T., Mir, M. N. H., Zeeshan, M. A. F., Hossain, A., hoque Jui, A., & Hasan, M. S. (2025). Predicting Energy Consumption in Hospitals Using Machine Learning: A Data-Driven Approach to Energy Efficiency in the USA. *Journal of Computer Science and Technology Studies*, *7*(1), 199-219.

[2] Alam, S., Chowdhury, F. R., Hasan, M. S., Hossain, S., Jakir, T., Hossain, A., ... & Islam, S. N. (2025). Intelligent Streetlight Control System Using Machine Learning Algorithms for Enhanced Energy Optimization in Smart Cities. *Journal of Ecohumanism*, *4*(4), 543-564.

[3] Amjad, M. H. H., Chowdhury, B. R., Reza, S. A., Shovon, M. S. S., Karmakar, M., Islam, M. R., ... & Ripa, S. J. (2025). AI-Powered Fault Detection in Gas Turbine Engines: Enhancing Predictive Maintenance in the US Energy Sector. *Journal of Ecohumanism*, *4*(4), 658-678.

[4] Anonna, F. R., Mohaimin, M. R., Ahmed, A., Nayeem, M. B., Akter, R., Alam, S., ... & Hossain, M. S. (2023). Machine Learning-Based Prediction of US CO_2 Emissions: Developing Models for Forecasting and Sustainable Policy Formulation. Journal of Environmental and Agricultural Studies, 4(3), 85–99.

[5] Barua, A., Karim, F., Islam, M. M., Das, N., Sumon, M. F. I., Rahman, A., ... & Khan, M. A. (2025). Optimizing Energy Consumption Patterns in Southern California: An AI-Driven Approach to Sustainable Resource Management. Journal of Ecohumanism, 4(1), 2920–2935.

[6] Chouksey, A., Shovon, M. S. S., Islam, M. R., Chowdhury, B. R., Ridoy, M. H., Rahman, M. A., & Amjad, M. H. H. (2025). Harnessing Machine Learning to Analyze Energy Generation and Capacity Trends in the USA: A Comprehensive Study. Journal of Environmental and Agricultural Studies, 6(1), 10–32.

[7] Deb, A., Zhang, X., & Wang, L. (2023). Graph neural network-based load forecasting for spatially correlated smart grids. *IEEE Transactions on Smart Grid*, 14(6), 5200–5209.

[8] Gazi, M. S., Barua, A., Karim, F., Siddiqui, M. I. H., Das, N., Islam, M. R., ... & Al Montaser, M. A. (2025). Machine Learning-Driven Analysis of Low-Carbon Technology Trade and Its Economic Impact in the USA. Journal of Ecohumanism, 4(1), 4961–4984.

[9] Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural Networks for Short-Term Load Forecasting: A Review and Evaluation. IEEE Transactions on Power Systems, 16(1), 44–55.

[10] Hossain, A., Ridoy, M. H., Chowdhury, B. R., Hossain, M. N., Rabbi, M. N. S., Ahad, M. A., ... & Hasan, M. S. (2024). Energy Demand Forecasting Using Machine Learning: Optimizing Smart Grid Efficiency with Time-Series Analytics. Journal of Environmental and Agricultural Studies, 5(1), 26–42.

[11] Hossain, M. S., Mohaimin, M. R., Alam, S., Rahman, M. A., Islam, M. R., Anonna, F. R., & Akter, R. (2025). AI-Powered Fault Prediction and Optimization in New Energy Vehicles (NEVs) for the US Market. Journal of Computer Science and Technology Studies, 7(1), 01–16.

[12] Hossain, S., Hasanuzzaman, M., Hossain, M., Amjad, M. H. H., Shovon, M. S. S., Hossain, M. S., & Rahman, M. K. (2025). Forecasting Energy Consumption Trends with Machine Learning Models for Improved Accuracy and Resource Management in the USA. Journal of Business and Management Studies, 7(1), 200–217.

[13] Kumar, P., Singh, R., & Das, L. (2025). Graph neural network–based spatiotemporal forecasting in power distribution networks. *Journal of Energy Informatics*, 6(1), 45–59.

[14] Lee, C., Gupta, R., & Wong, T. (2024). Distilling recurrent neural networks for interpretability in energy systems. *IEEE Transactions on Smart Grid*, 15(3), 987–995.

[15] Liu, H., Wang, J., & Zhao, L. (2023). Online transfer learning for adaptive load forecasting under concept drift. *Energy and AI*, 12, e100246.

[16] Reza, S. A., Hasan, M. S., Amjad, M. H. H., Islam, M. S., Rabbi, M. M. K., Hossain, A., ... & Jakir, T. (2025). Predicting Energy Consumption Patterns with Advanced Machine Learning Techniques for Sustainable Urban Development. Journal of Computer Science and Technology Studies, 7(1), 265–282.

[17] Rivera, L., Zhao, Y., & Ahmed, A. (2025). Bayesian ensemble methods for probabilistic load forecasting in smart grids. *Renewable Energy Journal*, 10(4), 200–215.

[18] Shovon, M. S. S., Gomes, C. A., Reza, S. A., Bhowmik, P. K., Gomes, C. A. H., Jakir, T., ... & Hasan, M. S. (2025). Forecasting Renewable Energy Trends in the USA: An AI-Driven Analysis of Electricity Production by Source. Journal of Ecohumanism, 4(3), 322–345.

[19] Smith, J., Doe, A., & Clark, B. (2025). Transformer-based short-term load forecasting for smart grids. *Journal of Energy Forecasting*, 5(2), 123–140.

[20] Wei, Y., Gao, J., & Sue, M. (2023). Federated learning for decentralized short-term load forecasting. *IEEE Access*, 11, 34567–34579.

[21] Zhang, L., Chen, Y., & Patel, S. (2024). Transformer-based probabilistic load forecasting in power systems. *Applied Energy*, 287, 116682.