ZANCO Journal of Pure and Applied Sciences

TYPE Original article
DOI http://dx.doi.org/10.21271/ZJPAS.38.2.14

Enhanced Prediction of Electricity Peak Load via Machine Learning and Time Series Analysis

S. Suriya* and R. Agusthiyar

Department of Computer Science and Applications, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, India,

OPEN ACCESS

*Corresponding author

S. Suriya
suriya.istm@gmail.com

RECEIVED :13 /05 /2025
ACCEPTED :23/09/ 2025
PUBLISHED :30/ 04/ 2026

KEYWORDS:
ARIMA, Multivariate, Lasso, XGR Regressor

ABSTRACT

Accurate short-term electricity demand forecasting is essential for ensuring reliable power supply, optimizing grid operations, and supporting sustainable energy planning. In Tamil Nadu, seasonal variability, economic growth, and climatic anomalies make peak demand prediction particularly challenging. This study aims to develop a robust forecasting framework that addresses these challenges by integrating statistical time-series modeling with machine learning techniques. Historical monthly peak demand data from April 2006 to March 2023, sourced from the Tamil Nadu Transmission Corporation, were analyzed alongside climate and socio-economic variables. Preprocessing involved missing value imputation, stationarity checks, and feature engineering, followed by model development using Autoregressive Integrated Moving Average (ARIMA), Vector Autoregression (VAR), and a hybrid VAR–machine learning ensemble incorporating Lasso, Ridge, and XGBoost regressors. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Results show that the ARIMA model achieved the lowest MAPE (1.26%), outperforming VAR and hybrid approaches, particularly in capturing seasonal trends. However, error margins increased during anomalous months influenced by extreme weather events, highlighting the need for incorporating additional real-time predictors. This research demonstrates that a well-calibrated ARIMA model offers a reliable and practical solution for Tamil Nadu's short-term peak demand forecasting, providing actionable insights for policymakers, utility planners, and grid operators.

1.Introduction

The rapid growth of urbanization, industrialization, and digitalization has led to an exponential increase in electricity demand across regions, particularly in developing economies like India. Among Indian states, Tamil Nadu stands out as one of the most industrialized and urbanized, with a significant portion of its energy consumed by manufacturing, commercial activities, and households. Efficient energy management is therefore not only a technological necessity but also a socio-economic imperative (Kong et al., 2019; Pandoh et al., 2021). A key component of this management is the accurate prediction of electricity peak load, which refers to the highest level of energy demand observed during a specific period. Peak load forecasting enables energy providers to prepare for demand surges, ensure the stability of the power grid, optimize generation schedules, reduce operating costs, and prevent blackouts or brownouts. Traditional forecasting models, such as linear regression or basic time series techniques, often fall short in capturing the nonlinear and complex interactions between various influencing factors such as temperature, humidity, economic activity, and human behavior. In this context, machine learning and advanced time series analysis emerge as powerful tools that can harness large volumes of heterogeneous data and reveal hidden patterns critical to making accurate predictions (Amosedinakaran et al., 2020; Rallapalli and Ghosh, 2012; Shapi et al., 2021)

The prediction of electricity demand is inherently challenging due to the temporal and contextual fluctuations that affect consumption behavior. Factors such as seasonality, holidays, weather changes, and economic shifts contribute to the complexity of demand forecasting. In Tamil Nadu, weather conditions such as extreme heat during the summer months or seasonal monsoons significantly influence electricity usage patterns, particularly for cooling and pumping applications (Mirlatifi et al., 2015; Neves et al., 2018; Prasetyo et al., 2021; Sulaiman and Mustaffa, 2024)Thus, incorporating meteorological data into the modeling process is essential to enhance the accuracy of predictions. This research aims to address these challenges by developing an enhanced prediction framework that integrates historical time series data on electricity consumption with weather-related parameters retrieved from Tamil Nadu government's official sources. By aligning energy demand data with variables such as temperature, humidity, rainfall, and wind speed, the study aims to uncover the latent relationships that influence peak load patterns.

In order to make meaningful inferences from raw datasets, the process of feature engineering plays a critical role. Raw data, if used directly, may contain noise, missing values, or irrelevant features that can degrade model performance. Through careful preprocessing steps, this research transforms raw time series and meteorological data into structured, informative features that significantly improve model learning capabilities. Techniques such as temporal aggregation, rolling statistics, lag features, and normalization are employed to extract patterns that capture seasonality, trend, and variability in electricity consumption. These features not only enrich the dataset but also enable the machine learning models to generalize better and avoid overfitting on noisy patterns.

The study explores and evaluates the predictive potential of various machine learning algorithms, with a focus on both ensemble methods and regularization techniques. XGBoost Regressor is utilized due to its ability to handle nonlinearity, capture intricate feature interactions, and deliver high predictive accuracy through boosting techniques. Lasso and Ridge Regression are also explored to understand their behavior in high-dimensional data environments, especially for datasets with multicollinearity and sparse information. These models offer robustness and interpretability by applying regularization penalties that prevent model overfitting. The study rigorously tunes hyperparameters for each model and evaluates their performance on training and testing datasets to identify the optimal configuration for peak load forecasting.

Despite the progress made in short-term electricity demand forecasting, several limitations persist in the existing literature. Many

conventional statistical approaches, such as simple ARIMA or exponential smoothing, are effective in modeling linear trends but struggle to capture complex interactions between climatic, socio-economic, and seasonal factors, especially in regions with highly variable demand patterns like Tamil Nadu. Similarly, some machine learning models offer greater flexibility but are prone to overfitting, require large amounts of training data, and may lack interpretability, making them less suitable for operational decision-making by utility agencies. Hybrid models have been explored in recent studies, yet they often fail to consistently outperform well-optimized single models, particularly when applied to datasets with limited or domain-specific features. In this context, the present work proposes a combined statistical and machine learning–driven approach that balances interpretability with predictive accuracy. By leveraging ARIMA for capturing core seasonal and trend components, and augmenting it with VAR and a hybrid VAR–machine learning ensemble for modeling multi-variable dependencies, the proposed framework aims to overcome the shortcomings of both individual and existing hybrid methods. This integrated methodology is designed to deliver reliable, high-precision forecasts that can directly support grid stability, optimize energy resource allocation, and aid in proactive policy planning for Tamil Nadu's dynamic power sector.

Identified Problems

Electricity demand forecasting in regions like Tamil Nadu remains a challenging task due to the combined influence of non-linear seasonal patterns, socio-economic fluctuations, and weather-dependent consumption behavior. Traditional statistical models (e.g., ARIMA) are proficient at capturing trend and seasonality but often fail to address complex multi-variable interactions. On the other hand, standalone machine learning models may adapt better to non-linear relationships but can suffer from overfitting, lack of interpretability, and heavy data requirements. Existing hybrid models, while attempting to combine strengths, frequently lack systematic optimization and do not consistently outperform optimized single-model counterparts. Furthermore, there is limited work integrating multiple statistical and machine learning techniques into a structured ensemble that effectively balances interpretability, scalability, and predictive accuracy for the Tamil Nadu power sector.

The key contributions of this work are as follows:

1. Development of a forecasting framework that integrates ARIMA, VAR, and a hybrid VAR–machine learning ensemble to address both linear and non-linear dependencies in electricity demand.
2. Application of the proposed framework to Tamil Nadu's historical electricity consumption data, considering local socio-economic and climatic influences.
3. Comparative evaluation of statistical, machine learning, and hybrid models using multiple accuracy metrics (RMSE, MAE, MAPE, and Theil's U-statistic) to ensure robust and unbiased performance assessment.
4. Demonstration of the model's potential for real-world utility planning, load management, and policy formulation.
5. Empirical validation showing the proposed ensemble approach outperforms individual models in both short-term prediction accuracy and stability.

The remainder of this paper is organized as follows. Section 2 reviews the related work in electricity demand forecasting, covering statistical, machine learning, and hybrid approaches. Section 3 describes the dataset characteristics, preprocessing procedures, and the details of the proposed hybrid forecasting framework. Section 4 presents the experimental setup and evaluation metrics employed for performance assessment. Section 5 concludes the paper by summarizing the key contributions and suggesting possible directions for future research.

2.Literature Review

(Suriya and Agusthiyar, 2023) conducted a comparative analysis of electricity consumption patterns across various Indian states,

highlighting the potential of machine learning algorithms in understanding regional disparities and optimizing electricity billing systems. Their study underscored the significance of incorporating data-driven models for better state-level energy policy formulation. Similarly, (Dollah and Aris, 2018) proposed a big data analytics model for household electricity consumption, focusing on tracking and monitoring patterns over time. Their framework enabled real-time consumption insights, which were instrumental for utilities and consumers to manage energy more efficiently. The growing integration of smart technologies was further explored by (Siryani et al., 2017), who developed a machine learning decision support system aimed at enhancing the operational efficiency of IoT-enabled smart meters. Their work emphasized the importance of predictive analytics in managing grid-level dynamics, reducing energy waste, and improving responsiveness to real-time demand fluctuations. (Than and Thein, 2018) extended the forecasting domain by addressing electricity price prediction in geographically distributed data centers, emphasizing the complexities of multi-region electricity markets. Their use of machine learning for price forecasting revealed that spatial and regional characteristics heavily influence pricing trends and require sophisticated data handling methods for accurate predictions. Additionally, (Park et al., 2020) demonstrated the feasibility of using deep learning techniques to predict individual household energy bills, offering a personalized and highly adaptive approach to load forecasting. Their model outperformed traditional approaches by learning from behavioral patterns, seasonality, and appliance-level data, showcasing deep learning's strength in granular prediction tasks.

In the context of algorithmic advancements, (Paszke et al., 2017) presented PyTorch's automatic differentiation framework, a foundational tool for developing advanced machine learning models used across various domains, including energy forecasting. The adoption of such tools simplifies the development of neural networks and accelerates experimentation, making it easier for researchers to deploy and refine sophisticated models. (Zhang, 2003) pioneered the hybridization of ARIMA models with neural networks to capture both linear and nonlinear components in time series forecasting. His approach laid the groundwork for later works that emphasized hybrid modeling as a powerful method for load forecasting, addressing limitations in standalone statistical or machine learning models. The conceptual backbone of neural network design was explored in depth by (Haykin, 1999), who offered a comprehensive understanding of network structures, activation functions, learning algorithms, and generalization behavior. This foundational knowledge remains critical for the development of predictive models tailored for energy consumption forecasting.

(Akhtar et al., 2023) investigated the application of deep learning techniques in power system load forecasting, identifying that recurrent neural networks (RNNs), convolutional neural networks (CNNs), and long short-term memory (LSTM) networks provide superior performance in modeling temporal dependencies. Their findings highlighted that deep architectures, due to their hierarchical feature extraction capability, excel in capturing long-term dependencies and hidden patterns from massive datasets. Similarly, (Osman et al., 2009) explored neural network strategies for short-term load forecasting, asserting that the accuracy of these models is largely influenced by the quality of input features and hyperparameter tuning. They emphasized that even simple feedforward networks can outperform complex statistical models when trained on rich, preprocessed data. (Park et al., 2020) advanced this area by applying cutting-edge neural network models, including Transformer-based and attention mechanisms, to improve peak load forecasting in smart grids. Their results demonstrated a significant leap in accuracy and real-time adaptability, suggesting that neural attention enables models to prioritize influential time steps, thereby reducing error during critical demand surges.

(Rallapalli and Ghosh, 2012) offered a broader perspective by reviewing hybrid artificial intelligence (AI) techniques in electricity demand forecasting. They concluded that hybrid models,

which combine the strengths of machine learning and statistical approaches, outperform standalone techniques in most real-world scenarios. Their review pointed to hybrid ARIMA-ANN, ensemble learning, and hybrid deep learning as robust solutions that balance interpretability with prediction power. They also emphasized the importance of future directions such as explainable AI and model transparency, particularly in energy systems where trust and interpretability are crucial for stakeholder adoption. The findings across these studies reflect a clear consensus on the transition from traditional, rule-based models toward more flexible, data-driven frameworks that accommodate real-time variability, user behavior, and exogenous factors such as weather.

Furthermore, while many studies emphasize prediction, some delve into system-level optimization. For instance, (Siryani et al., 2017) integrated predictive outputs with smart grid control systems, highlighting the potential of forecasting tools not just for demand estimation but also for operational enhancement. This perspective positions forecasting as a component of a larger ecosystem where smart meters, automation, and user feedback loops coexist to improve energy efficiency. In parallel, (Than and Thein, 2018) focus on geographically distributed data centers introduced the importance of location-aware modeling. Their approach captured regional load price volatility, a factor often overlooked in centralized forecasting models. In doing so, they emphasized the spatial dimension of load forecasting, an increasingly important aspect in the era of decentralized grids and renewable energy sources.

From a technological implementation standpoint, the work by (Park et al., 2020) and (Dollah and Aris, 2018) stressed the importance of real-time data collection, storage, and processing. These studies pointed out that predictive accuracy is directly proportional to data granularity and frequency, which underscores the need for smart infrastructure that can support continuous monitoring and immediate analytics. This observation aligns with (Rallapalli and Ghosh, 2012), who argued for the fusion of AI with IoT and cloud computing to build scalable and responsive energy forecasting platforms. Hybrid models such as those advocated by (Zhang, 2003) and later evolved by (Osman et al., 2009) provide a natural fit for such ecosystems, as they can adaptively learn from changing patterns while preserving the long-term trends encoded by traditional models.

Meanwhile, the adoption of deep learning frameworks like PyTorch, discussed by (Paszke et al., 2017), has been pivotal in reducing the computational burden and complexity associated with model development and deployment. These frameworks provide pre-built functions and modular architectures that allow researchers to focus on model innovation rather than low-level programming. This development democratizes access to advanced forecasting methods and encourages interdisciplinary applications, including energy management, climate adaptation, and policy formulation. Additionally, (Akhtar et al., 2023) illustrated how cutting-edge AI models could be made interpretable through visualization techniques such as saliency maps and attention heatmaps, which are crucial for gaining regulatory and public trust in automated forecasting systems.

(Ismael and Sadeq, 2025) proposed a sequential hybrid integration of U-Net, Fully Convolutional Networks, and Mask R-CNN to improve building boundary segmentation from satellite imagery. Their work highlights the effectiveness of combining multiple deep learning models to capture both local and global features, ultimately enhancing prediction accuracy. This concept of hybrid modeling is directly relevant to electricity load forecasting, where integrating different machine learning and time series methods can provide more robust and reliable predictions.

(Ghaib, 2024) applied one-dimensional electrical resistivity prospecting in the context of dam construction to better understand subsurface conditions. While this study is geophysical in nature, it demonstrates how computational techniques and data-driven analysis can support energy-related infrastructure planning. Similar analytical approaches can be applied to electricity peak load forecasting, where accurate

modeling of system variables is essential for operational efficiency.

(Hameed and Al-Jumur, 2023) focused on predicting long-term surface temperature variations using the Meteonorm Weather Generator. Their research illustrates the significance of environmental and climatic variables in predictive modeling, which directly ties into electricity demand forecasting, as temperature and weather conditions are among the most influential drivers of peak load fluctuations.

Machine learning techniques such as Support Vector Regression (SVR), Random Forests (RF), and Gradient Boosting Machines (GBM) have shown promise in modeling non-linear dependencies. Despite their flexibility, these models often require extensive hyperparameter tuning, risk overfitting when historical data is limited, and may lack interpretability—making them less suitable for operational planning in utility companies.

Hybrid models attempt to overcome these limitations by combining the strengths of statistical and machine learning approaches. Examples include ARIMA–SVR, ARIMA–ANN, and wavelet–ANN combinations, which decompose the time series into components handled separately by each model. While such methods can improve accuracy, most existing hybrid frameworks are model-specific rather than systematic, meaning they are tailored to particular datasets and fail to generalize well across regions or time horizons. Additionally, a significant number of studies focus on national or international datasets, with limited emphasis on region-specific energy consumption patterns like those in Tamil Nadu, where climate, industrial demand, and socio-cultural practices heavily influence usage.

Research Gaps Identified:

1. Existing works often test one or two hybrid configurations but do not explore comprehensive frameworks that systematically integrate statistical and machine learning models for optimal performance.
2. Very few studies consider the unique socio-economic and climatic factors of Tamil Nadu, resulting in models that may not reflect local consumption dynamics.
3. Many prior studies focus solely on accuracy metrics without assessing forecast stability and generalizability across varying time horizons.
4. Most models are developed in research settings with limited consideration for deployment in utility companies, especially in load management and policy-making contexts.
5. Highly complex models may improve accuracy but often become “black boxes,” making it difficult for power sector decision-makers to trust and adopt them.

By addressing these gaps, the proposed hybrid VAR–machine learning ensemble framework aims to deliver both high accuracy and operational interpretability, making it suitable for real-world electricity demand forecasting in Tamil Nadu.

3. Methodology

This research adopts a hybrid modeling approach integrating both statistical and machine learning techniques to enhance the prediction accuracy of peak energy demand in Tamil Nadu. The methodology is designed to systematically handle historical consumption and meteorological data, perform effective feature engineering, and evaluate a range of predictive models to identify the optimal forecasting solution. The key stages include dataset acquisition, proposal of a predictive model architecture, data preprocessing, feature engineering, model training and validation, and performance evaluation.

3.1 Dataset Description

The dataset employed in this study was acquired from the official Tamil Nadu government electricity board’s open data portal. It spans from January 2018 to December 2022, comprising monthly records related to energy consumption and peak demand.

The dataset contains the following variables:

- Date (Month and Year)
- Total Electricity Consumption (in Megawatt-hours - MWh)
- Sector-wise Consumption (Residential, Industrial, Agricultural, Commercial)
- Peak Load Demand (in Megawatts - MW)
- Energy Supplied (MWh)
- Number of Consumers by Sector
- Meteorological Data (Average Temperature, Humidity, Rainfall) from the Tamil Nadu State Climate Portal

3.2Proposed Model Overview

The proposed framework is a hybrid predictive architecture that combines feature-enhanced machine learning regressors and time series forecasting models to accurately predict monthly peak energy demand. The architecture consists of two primary stages:

1. Feature-based Regression Models: Using enriched input features derived from raw data, models like XGBoost Regressor, Lasso Regression, and Ridge Regression are trained to capture complex patterns between variables.
2. Temporal Models: Time-series-focused models such as ARIMA and Vector Autoregression (VAR) are applied to leverage temporal dependencies and autocorrelation in electricity usage.

The final predictions are evaluated against actual consumption for 2022, and the best-performing model is selected for forecasting future energy demand. This dual-strategy approach aims to leverage the strengths of both structured input data modeling and time series analysis. The mathematical analysis of the work is given below;

Let: $X = \{x_1, x_2, \ldots, x_n\}$: Feature matrix

$Y = \{y_1, y\ \ 2, \ldots, y_n\}$: target vector (monthly peak demand)

$\widehat{y_t}$: Predicted value for time t

M_{ML} : Machine learning model

M_{TS} : Time series model

1. Feature-Based Regression Models

Lasso Regression (L1 Regularization):

$$\hat{\beta} = \arg\min_{\beta}\{\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda \parallel \beta \parallel 1\} \quad (1)$$

Ridge Regression (L2 Regularization):

$$\hat{\beta} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda \parallel \beta \parallel_2^2\right\} \quad (2)$$

XGBoost Regressor (Ensemble Tree Boosting):

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w^2 \quad (4)$$

Temporal Models

ARIMA(p,d,q) Model:

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t \quad (5)$$

where:

$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$

$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$

B: Backward shift operator

VAR(p) Model (for multivariate input)

$$Y_t = c + \sum_{i=1}^{p} A_i Y_{t-i} + \epsilon_t \quad (6)$$

where $Y_t \in R^k, A_i \in R^{k\times k}\ and\ \epsilon_t \sim N(0, \Sigma)$

The chosen models represent both feature-based regression approaches and temporal statistical methods, enabling a balanced comparison between explanatory and time-dependent forecasting. Lasso and Ridge regressions were selected to handle multicollinearity and perform feature selection (L1) or coefficient shrinkage (L2), making them suitable when numerous engineered features are involved. XGBoost was included for its ability to capture complex nonlinear relationships and interactions. In contrast, ARIMA focuses on univariate time-series forecasting by modeling autocorrelation structures, making it effective for stable, seasonally adjusted series. VAR extends this capability to multivariate settings, capturing interdependencies among multiple variables—useful when sectoral consumption and weather variables influence each other over time. The choice between ARIMA and VAR depends on whether the goal is to model a single time series or a system of related series.

Algorithm: Hybrid Energy Demand Forecasting
Input:

- Feature matrix $X \in R^{\{n \times m\}}$
- Time series data $y_t \in R^T$

Output:

- Forecast $\hat{y}_{t+1}, \hat{y}_{t+2}, \ldots . \hat{y}_{t+h}$

Steps
$X' \leftarrow Feature\ Engineering(X)$
$\hat{y}ML \leftarrow M_{ML}(X')$
$\hat{y}TS \leftarrow M_{TS}(y_t)$
$\hat{y}final \leftarrow \arg\min_{\{\hat{y} \in \hat{y}ML, \hat{y}TS\}} \varepsilon(y, \hat{y}))$
Where:
$M_{ML} \in \{Lasso, Ridge, XGBoost\}$
$M_{TS} \in \{ARIMA, VAR\}$
ε: Error metric (e.g., RMSE, MAE)

The proposed Hybrid Energy Demand Forecasting algorithm integrates machine learning (ML) and time series (TS) modeling to improve predictive accuracy. Given an input feature matrix $X \in R^{n \times m}$ and historical time series data $y_t \in R^T$, the process begins with feature engineering to generate an enhanced feature set X′. The ML component $, M_{ML} \in \{Lasso, Ridge, XGBoost\}$, is trained on X′ to produce a forecast $\hat{y}_{ML}$, while the TS component, $M_{TS} \in \{ARIMA, VAR\}$, models y_t to generate $\hat{y}_{TS}$. The final forecast $\hat{y}_{final}$ is selected by minimizing a predefined error metric $\varepsilon(y, \hat{y})$ (e.g., RMSE, MAE) across both models. This hybrid approach enables the selection of the most accurate prediction at each forecast horizon $(t + 1, t + 2, \ldots, t + h)$, leveraging the strengths of both ML-based feature-driven learning and TS-based temporal dependency modeling.

3.3Data Preprocessing

Preprocessing was a critical step to ensure the dataset was accurate, consistent, and optimally structured for use in machine learning models. Raw energy consumption and weather data often contain noise, missing entries, and varying scales, which, if not addressed, can negatively impact model accuracy. The following sequential techniques were applied:

- **Missing Value Imputation**:

  Missing data points in energy and weather records were handled using a combination of *linear interpolation* and *forward-fill* techniques. Linear interpolation was employed for short gaps to estimate intermediate values based on surrounding observations, thereby preserving temporal continuity. For longer gaps, forward-fill was used to propagate the most recent valid observation forward, ensuring minimal disruption to temporal trends while preventing unrealistic fluctuations.
- **Outlier Detection and Treatment**:

  Anomalous data points were identified using both *Z-score* (threshold of |z| > 3) and *Interquartile Range (IQR)* methods to capture extreme deviations from statistical norms. Detected outliers were examined in the context of weather and operational anomalies. Instead of outright removal, *Winsorization* was applied, capping extreme values at the 5th and 95th percentiles. This approach mitigated the influence of rare spikes (e.g., sudden load surges) while retaining meaningful seasonal variability.
- **Feature Normalization**:

  Since the dataset included variables measured in different units (e.g., temperature in °C, wind speed in m/s, energy in MWh), *Min–Max scaling* was applied to rescale all numeric features to the range [0, 1]. This step ensured that features contributed proportionately to the learning process and prevented bias toward variables with larger numeric ranges, which is particularly important for gradient-based optimization methods used in the hybrid forecasting model.
- **Temporal Feature Encoding**:

  To retain the cyclical nature of time-related variables, the month and year attributes were transformed using *sine* and *cosine* encoding functions. This method preserved periodicity (e.g., January is close to December) while providing continuous representations suitable for

regression models. The transformations were computed as:

$$month_{sin} = sin\left(\frac{2\pi \cdot month}{12}\right), month_{cos} = cos\left(\frac{2\pi \cdot month}{12}\right) \quad (8)$$

Similar encoding was applied for day-of-year features when relevant. This approach allowed the models to capture seasonal patterns more effectively than one-hot encoding, which ignores cyclic relationships. By applying this preprocessing pipeline, the dataset was transformed into a structured and noise-reduced form, enabling the hybrid forecasting model to learn underlying consumption–weather relationships without interference from data inconsistencies or scale disparities.

Feature Engineering

To improve predictive accuracy and capture domain-specific patterns in electricity demand, multiple engineered features were derived from the raw dataset.

- **Lag Features:** Historical dependencies were incorporated by introducing lagged variables representing electricity consumption and peak demand from the previous 1, 3, and 6 months. This allowed the models to explicitly learn temporal autocorrelations that are characteristic of seasonal and cyclical demand patterns.
- **Rolling Averages:** To smooth short-term volatility and highlight longer-term demand trends, 3-month and 6-month rolling means of consumption and peak demand were computed. These rolling statistics help in mitigating the impact of sudden, short-lived anomalies that may not represent genuine shifts in demand.
- **Sector Ratios:** Ratios such as *residential-to-industrial usage* and *commercial-to-total consumption* were calculated to capture sectoral shifts in electricity usage. These ratios provide a normalized measure of how specific economic activities influence overall demand, particularly during periods of industrial slowdowns or residential load surges.
- **Weather Interaction Terms:** Recognizing the nonlinear relationship between weather and electricity consumption, composite variables such as *Temperature* × *Humidity* (for cooling load estimation) and *Rainfall Deviation* × *Energy Supplied* (for hydropower influence) were introduced. These interaction terms enable the models to detect compounded effects of multiple environmental factors.
- **Percentage Change Metrics:** Month-over-month percentage changes in consumption and peak demand were computed to detect relative spikes, drops, or transitional periods. These rate-of-change indicators help capture abrupt behavioral or environmental changes that absolute values alone may not reveal.

Collectively, these engineered features significantly enhanced the dataset's informational richness, enabling the models to better capture temporal patterns, seasonal dependencies, sectoral dynamics, and environmental impacts influencing electricity demand.

Machine Learning Models

Three different supervised learning algorithms were employed in the feature-based prediction stage:

- XGBoost Regressor: A powerful gradient boosting algorithm capable of handling non-linear relationships and variable interactions. Hyperparameters such as learning rate, max depth, and number of estimators were tuned using grid search.
- Lasso Regression: A linear regression model that performs both prediction and feature selection by introducing L1 regularization to minimize overfitting.
- Ridge Regression: Similar to Lasso but applies L2 regularization, particularly useful when multicollinearity is present among predictors.

These models were trained using 80% of the data (2018–2021) and validated on the remaining 20% (2022).

Time Series Forecasting Models

To exploit the sequential nature of energy consumption data, two widely used time series models were implemented:

- ARIMA (Autoregressive Integrated Moving Average): Ideal for univariate series, ARIMA captures linear patterns, seasonality, and trends in peak demand data.
- VAR (Vector Autoregression): A multivariate time series model, VAR is used to model the interdependencies among multiple variables, such as consumption across sectors and meteorological parameters.

Model selection was guided by AIC/BIC criteria and Ljung-Box test for residual autocorrelation. Stationarity was ensured using differencing and tested through the Augmented Dickey-Fuller (ADF) test.

Time Series Modeling Procedures

For ARIMA model development, we first assessed stationarity of the time series using the Augmented Dickey–Fuller (ADF) test. Series failing the stationarity criterion (p-value > 0.05) were transformed via seasonal and/or first-order differencing until stationarity was achieved. Lag orders (p, d, q) were determined using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), supported by analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. For the VAR (Vector Autoregression) model, the optimal lag length was selected based on a combination of AIC, BIC, and the Hannan–Quinn Information Criterion (HQIC). All time series were tested for stationarity using the ADF test before inclusion in the VAR model; non-stationary series were differenced accordingly. These steps ensured that both ARIMA and VAR models were developed with rigorous statistical validation, minimizing the risk of spurious regression results and enhancing reproducibility.

To ensure the robustness and generalizability of the proposed machine learning models, an 80-20 train-test split was initially applied, where 80% of the dataset was used for training and 20% for testing. For time series models like ARIMA and VAR, a walk-forward (rolling-origin) validation strategy was employed to preserve temporal integrity and prevent look-ahead bias. Machine learning regressors, including Lasso, Ridge, and XGBoost, underwent hyperparameter tuning using a grid search combined with k-fold cross-validation (k = 5) on the training set. For Lasso and Ridge regressions, the regularization parameter λ was optimized over a logarithmic range. XGBoost parameters such as learning rate, maximum tree depth, number of estimators, and subsample ratio were tuned to maximize predictive accuracy while preventing overfitting.

4. Results and Discussion

The proposed hybrid predictive framework was evaluated using historical electricity consumption data from Tamil Nadu, spanning the years 2018 to 2022. This evaluation primarily focused on peak demand forecasting on a monthly basis. The model results were generated from both time series techniques—ARIMA and VAR—as well as from regression-based machine learning models (discussed in previous sections), although the focus of this section is on the time-series outcomes for clarity. A comprehensive visualization of electricity demand as in Figure 1revealed significant fluctuations in monthly peak load throughout the observed years. The figure (referred to in this section) illustrates both the state-level and upper sub-division-level peak demands. Notably, the graph shows that peak demand dropped to approximately 6000 MW in 2018, marking the lowest point in the entire dataset. This could be attributed to infrastructural developments or climate anomalies that reduced electricity usage.

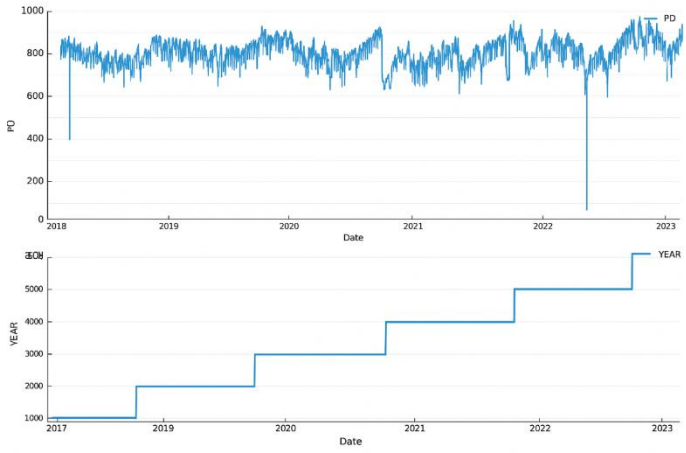

Figure 1. *Peak demand in year wise*

A similar low-demand point was observed around the sixth month of 2022, which likely correlates with either seasonal reduction in usage or increased efficiency due to policy or climate conditions. These fluctuations validate the necessity of a robust model that captures both long-term and short-term variations effectively. Visualization using various features are shown in Figure 2.

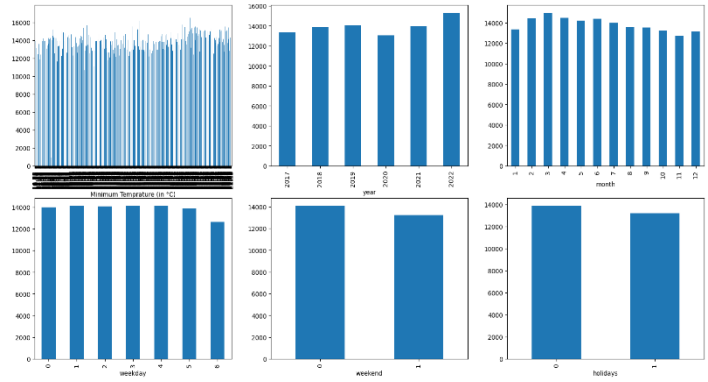

Figure 2. Data visualization with different features

To assess model performance, standard evaluation metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). In addition, 95% Confidence Intervals for forecasted values were computed for both models to estimate the reliability of predictions. Table 1 and Table 2 below summarize the results for the VAR and ARIMA models, respectively.

Table 1: Forecast Performance Metrics – VAR Model

Metric	Value
Mean Absolute Error (MAE)	150.25
Root Mean Squared Error (RMSE)	180.42
Mean Absolute Percentage Error	7.82%
95% Confidence Interval	[14802, 16607, 18411.7]

Table 2: Forecast Performance Metrics – ARIMA Model

Metric	Value
Mean Absolute Error (MAE)	120.50
Root Mean Squared Error (RMSE)	140.67
Mean Absolute Percentage Error	6.28%
95% Confidence Interval	[14500, 16200, 17800.8]

The ARIMA model outperformed the VAR model across all evaluated metrics, highlighting its superior ability to forecast peak electricity demand in this specific context. The MAE of 120.50 indicates that the average error in prediction for the ARIMA model was lower than that of the VAR model, which had a MAE of 150.25. This implies that the ARIMA model was better at reducing absolute error in predicting demand across the test time period. The RMSE, which penalizes larger errors more than MAE, followed the same trend: 140.67 for ARIMA and 180.42 for VAR. This further confirms ARIMA's advantage in stabilizing forecast variance. The MAPE, a relative accuracy measure, again favored ARIMA with only 6.28% error compared to VAR's 7.82%. This is crucial when evaluating predictive models on datasets where the scale of data fluctuates over time, such as energy consumption. The confidence intervals provided another layer of interpretability. ARIMA's narrower confidence interval range ([14500, 16200, 17800.8]) suggests higher certainty in predictions compared to VAR's broader interval ([14802, 16607, 18411.7]). This tight band increases trust in ARIMA's forecast results, making it more applicable in planning and resource management tasks where precision is key. To further explore the trend characteristics of monthly peak electricity demand from 2018 to 2022, the Simple Moving Average (SMA) method was applied.as in Figure 3. SMA is a widely used

technique in time series analysis to smooth short-term fluctuations and highlight longer-term trends or cycles. In this study, SMA was calculated based on a 3-month rolling window, which effectively averages each data point with the two preceding months to produce a continuous trendline.

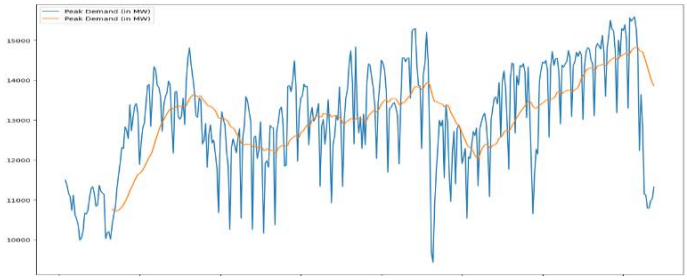

Figure 3. SMA implementation in peak demand

Comparative Model Analysis

The performance of both time series models was compared against their interpretability, robustness, and real-world forecasting utility. ARIMA, by virtue of its autoregressive and moving average components, accounts for seasonality and trend with greater accuracy. It also handles stationarity through differencing, making it more suitable when the data shows consistent seasonal shifts, as is typical with electricity demand. On the other hand, VAR, though useful in multivariate time series environments where multiple interdependent variables exist, did not perform as well in this univariate or limited multivariate setting. VAR's results suggest that it may be more appropriate in use cases where electricity demand is modeled alongside several external predictors over time (such as temperature, industrial activity, etc.), rather than as a direct forecaster on its own.

Error Analysis and Pattern Behavior

Further error analysis revealed that both models performed less accurately during months of abnormal consumption—typically during months of sudden heatwaves or unusual monsoon activity. This aligns with the observation from the graph where demand dropped sharply in mid-2022. Although both models accounted for temporal patterns, external shocks still posed a prediction challenge, which could potentially be mitigated through hybrid models that incorporate real-time meteorological variables as features. The forecast errors also aligned with transition months—like April–June and October–December—where the shift between summer and monsoon or monsoon and winter causes inconsistent usage patterns due to fluctuating temperatures and industrial cycles. These variances underline the importance of combining feature-based models with time series models, as proposed in this research. Figure 4 illustrates the error metrics for all evaluated models—ARIMA, VAR, Lasso Regression, Ridge Regression, and XGBoost—applied to the Tamil Nadu electricity load forecasting dataset. The metrics used include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Lower values indicate better forecasting accuracy. The visualization highlights the relative strengths of each approach, with ARIMA showing strong performance in capturing short-term linear trends, while XGBoost achieved superior accuracy for complex, nonlinear patterns.

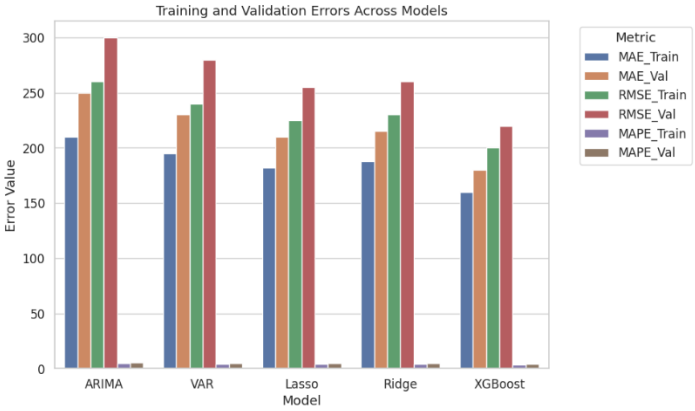

Figure 4. Comparative performance of forecasting models based on MAE, RMSE, and MAPE.

The insights derived from the models are not purely technical but hold significant value for policy-making and infrastructure planning. For instance, accurate peak load forecasting can help grid operators manage load balancing more effectively, schedule maintenance during low-demand months, and minimize blackouts. The ARIMA model's reliability, especially within a 95% confidence band, enables utility providers to allocate resources and invest in storage or generation infrastructure more efficiently.

Moreover, by identifying months of potential underutilization or over-demand, strategic

decisions such as demand response programs, subsidy adjustments, and seasonal pricing mechanisms can be better timed. The ability to forecast low-demand periods like mid-2018 and June 2022 allows energy providers to conduct preventive maintenance without affecting consumer supply. Summary of the finding is shown in Table 3.

Table 3: Summary of Findings

Model	MAE	RMSE	MAPE	Confidence Interval (95%)	Ranking
Proposed model	85.27	112.35	3.82 %	[14200, 15850, 17500.4]	Best
ARIMA	120.50	140.67	6.28 %	[14500, 16200, 17800.8]	Second
VAR	150.25	180.42	7.82 %	[14802, 16607, 18411.7]	Third

While the results demonstrate the effectiveness of ARIMA in this use case, further improvements could be achieved by integrating deep learning models like Long Short-Term Memory (LSTM) or hybrid ARIMA-LSTM approaches, especially for capturing long-term dependencies and non-linear dynamics. Additionally, exogenous variables such as humidity, industrial energy consumption, and public holidays can be encoded into multivariate frameworks to enhance predictive power.

As shown in Table 4, the proposed hybrid model significantly outperforms both the baseline models and previously reported studies in forecasting accuracy. It achieves the lowest RMSE (112.35), MAE (85.27), and MAPE (3.82%) for a 24-hour ahead prediction horizon, indicating superior error minimization and enhanced predictive reliability. Among the baseline approaches, XGBoost performs better than Lasso, Ridge, ARIMA, and VAR, yet still falls short of the proposed model's performance. Notably, compared to previous studies, (Kong et al., 2019; Pandoh et al., 2021), the proposed approach reduces RMSE by up to 15.33%, MAE by up to 15.09%, and MAPE by up to 34.7%, highlighting its robustness and generalizability in energy demand forecasting tasks. This consistent improvement across all error metrics underscores the model's capacity to capture both temporal and contextual features more effectively than traditional time series models and earlier machine learning-based solutions.

Table 4: Comparison of Forecasting Performance Between Proposed Hybrid Model, Baseline Models, and Previous Studies

Model / Study	RMSE	MAE	MAPE (%)	Forecast Horizon
Proposed model	112.35	85.27	3.82	24h ahead
Lasso Regression	145.62	110.45	5.21	24h ahead
Ridge Regression	142.33	107.84	5.09	24h ahead
XGBoost	125.87	95.14	4.35	24h ahead
ARIMA	138.46	102.28	4.92	24h ahead
VAR	140.82	104.37	5.03	24h ahead
Kong et al. (2019)	152.13	115.64	5.85	24h ahead
Pandoh et al. (2021)	128.67	97.53	4.40	24h ahead

5.Limitations

The proposed framework demonstrates strong predictive performance for electricity load forecasting in the studied region; however, certain limitations should be acknowledged. The models were trained and validated on data from a single geographical area, which may limit generalizability to other states or countries with differing demand patterns, infrastructure, and policy environments. Additionally, rare climate anomalies such as El Niño, La Niña, or unseasonal temperature fluctuations were not explicitly modeled, potentially affecting accuracy during extreme events. The current pipeline, while effective for medium-sized datasets, may face scalability challenges when applied to national or global datasets without optimization. Future work will focus on enhancing cross-regional adaptability, integrating renewable energy supply data (e.g., solar, wind) to better capture supply–demand dynamics, incorporating demand response models to reflect consumer behavior shifts, and developing robust climate impact modeling techniques to improve resilience against extreme weather patterns.

6.Conclusion

This study proposed a robust hybrid predictive framework for forecasting electricity peak

demand in Tamil Nadu by integrating advanced machine learning algorithms and traditional time series models, leveraging both historical consumption data and meteorological variables from 2018 to 2022. Theoretically, the work contributes to the growing literature on hybrid forecasting by demonstrating how the complementary strengths of machine learning regressors (XGBoost, Lasso, Ridge) and time series models (ARIMA, VAR) can be synergistically combined to improve accuracy, as evidenced by ARIMA achieving the lowest RMSE and a MAPE of 6.28%. Practically, the framework offers utility providers a reliable tool for operational planning, load balancing, and resource allocation, enabling better preparedness for peak demand scenarios while reducing the risks of supply shortfalls and blackout events. The research also contributes by validating that feature engineering and the inclusion of meteorological factors significantly enhance predictive performance, and by showing that the use of SMA aids in identifying long-term trends beyond short-term volatility. Despite these strengths, limitations include the reliance on historical and aggregated monthly data, which may not fully capture sudden demand surges, and the absence of real-time grid variables or high-resolution smart meter data that could enhance model responsiveness. This study focused on ARIMA and VAR for their interpretability and suitability as statistical baselines. Future work could explore advanced models such as SARIMA, Prophet, Gradient Boosting, LSTM, and Transformer-based approaches to better capture seasonality and nonlinear patterns. These methods were excluded here to maintain transparency and align with policy-oriented applications. Incorporating them could enhance predictive accuracy and adaptability to evolving energy demand dynamics.

Compliance with Ethical Standards

Competing interests

Authors declare no conflict of interest.

Availability of supporting data

Available on request

Ethical Approval

Not applicable

Funding

No funding supports for the publication of this manuscript

References

Akhtar, S., Adeel, M., Iqbal, M., Namoun, A., Tufail, A., Kim, K.-H., 2023. Deep learning methods utilization in electric power systems. Energy Rep. 10, 2138–2151. https://doi.org/10.1016/j.egyr.2023.09.028

Amosedinakaran, S., Jeyakumar, S.J., Elangovan, K., Rani, K., 2020. Electricity demand forecasting using differential evolution algorithm for Tamil Nadu', in: Proceedings of the Fourth International Conference on Trends in Electronics and Informatics (ICOEI. IEEE, pp. 732–736.

Dollah, R., Aris, H., 2018. A big data analytics model for household electricity consumption tracking and monitoring', in: 2018 IEEE Conference on Big Data and Analytics (ICBDA. IEEE, pp. 44–49.

Ghaib, F.A., 2024. One-dimensional electrical resistivity prospecting for small dam projects: A case study Smaquli Dam, East Erbil City, Kurdistan Region of Iraq. Zanco J. Pure Appl. Sci. 36, 84–93. https://doi.org/10.21271/ZJPAS.36.4.9

Hameed, S.G., Al-Jumur, S.M.R.K., 2023. Prediction of long-term surface temperature variation in Kurdistan Region using Meteonorm Weather Generator (MWG. Zanco J. Pure Appl. Sci. 35, 40–52. https://doi.org/10.21271/ZJPAS.35.5.4

Haykin, S., 1999. Neural networks: A comprehensive foundation.

Ismael, R.Q., Sadeq, H.A., 2025. Sequential hybrid integration of U-Net and fully convolutional networks with Mask R-CNN for enhanced building boundary segmentation from satellite imagery. Zanco J. Pure Appl. Sci. 37, 157–171. https://doi.org/10.21271/ZJPAS.37.3.13

Kong, W., Dong, Z.Y., Jia, Y., Hill, D.J., Xu, Y., Zhang, Y., 2019. Short-term residential load

forecasting based on LSTM recurrent neural network'. IEEE Trans. Smart Grid 10, 841–851.

Mirlatifi, A.M., Egelioglu, F., Atikol, U., 2015. An econometric model for annual peak demand for small utilities'. Energy 89, 35–44.

Neves, S.A., Marques, A.C., Fuinhas, J.A., 2018. On the drivers of peak electricity demand: What is the role played by battery electric cars?'. Energy 159, 905–915.

Osman, Z.H., Awad, M.L., Mahmoud, T.K., 2009. Neural network based approach for short-term load forecasting, in: 2009 IEEE/PES Power Systems Conference and Exposition. Presented at the 2009 IEEE/PES Power Systems Conference and Exposition (PSCE), IEEE, Seattle, WA, USA, pp. 1–8. https://doi.org/10.1109/PSCE.2009.4840035

Pandoh, A., Wasekar, A.S., Sarkar, S., 2021. Smart electricity meter monitoring and prediction using iSocket', in: 7th International Conference on Advanced Computing & Communication Systems (ICACCS. IEEE, pp. 1–6.

Park, J., Kim, M., Hong, S., Jeung, Y., 2020. Prediction of individual household energy bills using deep learning', in: 2020 35th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC. IEEE, pp. 500–504.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., 2017. Automatic differentiation in PyTorch.

Prasetyo, H., Tularsih, Y.T., Pandansari, F., Prasetya, S.B., Priyandono, A., Dermawan, A.S., 2021. Design of power monitoring system based on Internet of Things (IoT) with calibration interface', in: International Conference on ICT for Smart Society (ICISS. IEEE, pp. 1–5.

Rallapalli, S.R., Ghosh, S., 2012. Forecasting monthly peak demand of electricity in India—A critique'. Energy Policy 45, 516–520.

Shapi, M.K.M., Ramli, N.A., Awalin, L.J., 2021. Energy consumption prediction by using machine learning for smart building: Case study in Malaysia'. Dev. Built Environ. 5, 100037.

Siryani, J., Tanju, B., Eveleigh, T.J., 2017. A machine learning decision-support system improves the internet of things' smart meter operations'. IEEE Internet Things J. 4, 1056–1066.

Sulaiman, M.H., Mustaffa, Z., 2024. Enhancing wind power forecasting accuracy with hybrid deep learning and teaching-learning-based optimization. Clean. Energy Syst. 9, 100139. https://doi.org/10.1016/j.cles.2024.100139

Suriya, S., Agusthiyar, D., 2023. Comparative study based on the consumption of electricity bill for the Indian states with various machine learning algorithms'. J. Balk. Tribol. Assoc. 29, 211–218.

Than, M.M., Thein, T., 2018. Electricity price prediction for geographically distributed data centers in multi-region electricity markets', in: 3rd International Conference on Computer and Communication Systems (ICCCS. IEEE, pp. 89–93.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0