Comparative Analysis of Machine Learning Algorithms for Energy Consumption Forecasting

Sonali Nemade*, Ashwini Patil, Deepashree Mehendale, Reshma Masurekar

Department of Computer Science, Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune 18, Maharashtra, India

DOI: https://doi.org/10.51583/IJLTEMAS.2025.1413SP029

Received: 26 June 2025; Accepted: 30 June 2025; Published: 24 October 2025

Abstract: Forecasting energy use has become a crucial component of contemporary smart grid systems, allowing stakeholders to guarantee system dependability, cost effectiveness, and energy efficiency. For the integration of intermittent renewable energy sources, load balancing, and real-time energy management, the capacity to predict power demand is essential. The use and relative effectiveness of five supervised machine learning algorithms Linear Regression, Decision Tree, Random Forest, XGBoost, and Gradient Boosting for predicting short-term building-level energy consumption are examined in this work. In order to train and evaluate models, we carried out a thorough preprocessing and feature engineering procedure using a large dataset that included operational, meteorological, and temporal variables.

Each model was assessed using three key performance metrics: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). Among the models tested, Gradient Boosting achieved the highest accuracy, with an MAE of 575 kWh, RMSE of 851 kWh, and an R^2 of 0.949, outperforming both traditional and advanced ensemble models.

Our results highlight how well boosting strategies work for energy forecasting jobs and how crucial it is to choose models according to deployment restrictions and data properties. The knowledge gained from this research can help designers create responsive, scalable, and intelligent energy forecasting systems that are appropriate for smart infrastructure.

Keywords: Energy forecasting, smart grid, Machine learning, Gradient Boosting, Random Forest, Time series

I. Introduction

With increasing global energy demands and the transition toward decentralized power systems, accurate energy consumption forecasting has become essential. Effective forecasting ensures reliability in supply, facilitates energy conservation and supports demand-side management. In the era of smart grids, energy providers and consumers rely heavily on real-time analytics and predictive models to anticipate fluctuations in energy demand.

Energy forecasting plays a crucial role not only in daily operations of utilities but also in long-term infrastructure planning, cost optimization, and environmental protection. The increasing penetration of renewable energy sources such as solar and wind which is inherently variable adds further complexity to energy load management. Consequently, forecasting systems must adapt to incorporate weather conditions, consumer behavior, and temporal trends.

Traditional statistical approaches often assume linearity and stationarity, making them insufficient for modeling the dynamic and nonlinear nature of real-world energy consumption. They typically fail to accommodate irregular consumption patterns caused by holidays, changes in occupancy, or abrupt weather events. In contrast, machine learning (ML) offers scalable and flexible frameworks capable of capturing these intricate relationships without strict assumptions about data distributions.

This study aims to explore the application of several well-established ML models for short-term load forecasting in a commercial building setting, focusing specifically on summer peak periods when energy consumption is typically at its highest. By benchmarking multiple algorithms, we intend to provide insights into their strengths and trade-offs and to identify the most promising model for practical deployment in energy management systems. With increasing global energy demands and the transition toward decentralized power systems, accurate energy consumption forecasting has become essential. Effective forecasting ensures reliability in supply, facilitates energy conservation and supports demand-side management. Traditional statistical approaches often assume linearity and stationarity, making them insufficient for modeling the dynamic and nonlinear nature of real-world energy consumption. Machine learning (ML), in contrast, provides flexible, data-driven techniques capable of uncovering hidden patterns within complex, high-dimensional datasets. This paper explores the use of several ML models for predicting energy consumption in commercial buildings during summer peak periods.

II. Literature Review:

Traditional models like ARIMA (Autoregressive Integrated Moving Average) and exponential smoothing have historically been used for energy forecasting due to their simplicity and interpretability. These methods rely on past trends and seasonality to make future predictions, assuming that patterns repeat over time. While suitable for stationary time series, these models struggle to account for dynamic, real-time changes in energy consumption due to varying user behavior, irregular external influences such as weather, and policy interventions.

More recent research has turned toward machine learning (ML) techniques for their ability to capture nonlinear patterns and complex feature interactions. Ensemble methods like Random Forest (RF) and Gradient Boosting Machines (GBM) have demonstrated robustness and accuracy, especially when trained on high-dimensional datasets. Breiman (2001) introduced Random Forests as a way to improve prediction accuracy by averaging multiple decision trees trained on bootstrapped data subsets. Similarly, Friedman (2001) developed the Gradient Boosting framework, which sequentially builds models to minimize residual errors.

XGBoost (Extreme Gradient Boosting), proposed by Chen and Guestrin (2016), extended GBM with advanced regularization, efficient parallel computation, and handling of missing data, making it a top-performing method in many data science competitions. In the context of energy forecasting, these tree-based models provide not only high accuracy but also interpretable outputs through feature importance metrics.

Although deep learning methods (e.g., LSTM, Transformer architectures) show promising results, they require larger datasets, extensive tuning, and longer training times. Therefore, in this work, we focus on ML models that strike a balance between performance, interpretability, and computational efficiency.

Problem Statement:

Accurately predicting daily energy consumption has become essential for energy providers, policymakers, and infrastructure planners due to the increase in demand for electricity throughout the summer. Better demand-side management and resource allocation are made possible by accurate energy forecasting, which also helps prevent grid overloads and unforeseen peak occurrences that could result in blackouts or expensive operating expenses.

In order to predict daily energy consumption and identify possible peak power demand days throughout the summer, this project intends to create a machine learning-based forecasting model using historical temperature data, activity levels, and previous energy usage trends. The goal is to develop a predictive model that can effectively generalize and assist in data-driven energy management decisions by examining a dataset that contains comprehensive temperature readings (T1–T10), energy measures, and user activity.

Objective:

- To understand and preprocess the provided dataset (summer_peaks.csv), which includes temperature variables (T1–T10), daily energy consumption, peak power indicators, and activity levels.
- To explore the influence of temperature, activity, and temporal features (like day of the week) on daily energy usage.
- To design and train multiple machine learning models (as demonstrated in the provided Jupyter notebook) to predict energy consumption, including algorithms like Linear Regression, Random Forest, and XGBoost.
- To compare model performance using evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score to identify the most effective forecasting approach.

III. Methodology:

This study uses machine learning techniques to forecast daily energy use over the summer using an organized, data-driven methodology. Ten temperature sensor readings (T1–T10), total energy consumption, peak power characteristics, a binary peak day indication, activity levels, and date-related data are all included in the summer_peaks.csv dataset, which serves as the basis for the analysis.

The dataset is first preprocessed using appropriate imputation techniques to address missing values, especially in the later temperature variables (T6–T10). Time-based patterns are captured by extracting temporal information, including the day of the week, and converting the date field to the correct date time format. In order to simplify the input space while maintaining significant variance, the 10 temperature values are also combined to calculate an average temperature feature.

Statistical plots and summary statistics are used in exploratory data analysis (EDA) to find relationships between temperature, activity, and energy consumption. By adding additional variables like average temperature and weekday/weekend classification and eliminating superfluous or strongly correlated features, feature engineering is used to improve model input and lessen over fitting.

After that, a number of machine learning models such as XGBoost Regressor, Random Forest Regressor, and Linear Regression are created and trained. Eighty percent of the dataset is used to train these Python-implemented models, with the remaining twenty percent set aside for testing. Standard regression indicators, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R2 Score), are used to assess the efficacy of the model. The model chosen for predicting is the one that performs the best across these metrics. Lastly, future energy usage is predicted using the chosen model. To determine which input factors, have the greatest impact on energy consumption, feature importance analysis is performed. Particular focus is placed on finding and forecasting days with peak demand. This makes it possible for the suggested system to be a practical tool for proactive load management and energy demand predictions during summertime peak demand.

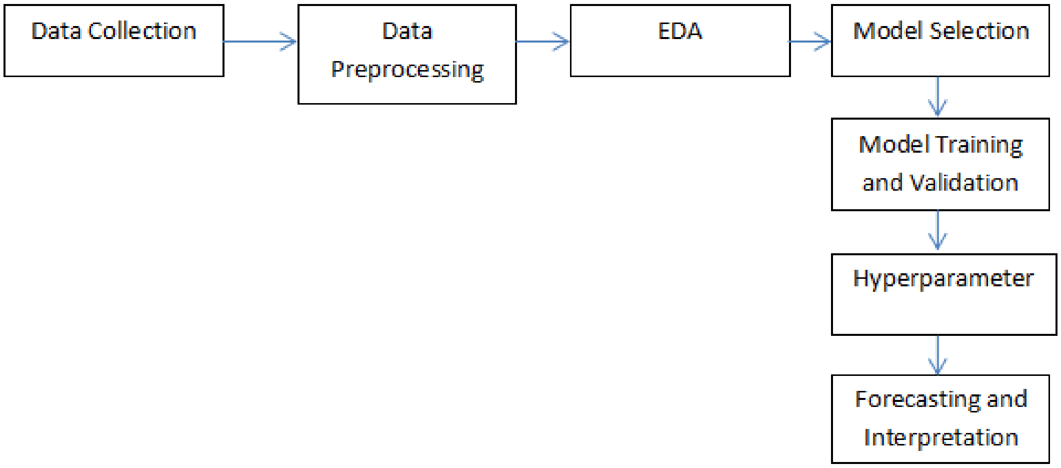

Figure1: Energy Forecasting Methodology

Dataset:

589 records with 18 columns make up the summer_peaks.csv dataset, which collects daily information on temperature, energy use and peak electricity demand. With the date in the day column and the day of the week (where 1 denotes Monday and 7 Sunday) in the wd column, each row corresponds to a single day. Ten variables, designated T1 through T10, collect temperature data; these values most likely come from various sensors or at various times of the day. These temperature data, which are given in degrees Celsius, shed light on the daily fluctuations in temperature.

The dataset contains energy-related parameters in addition to temperature. Peak power captures the maximum power usage seen during the day, whereas the energy column documents the entire energy consumption for the day. The system's peak power usage duration is indicated by the peak duration field, while the peak intensity field indicates how strong or severe the peak event was. A Boolean flag in this is peak column indicates if a peak event happened that day (True or False).

Lastly, the dataset has an activity column, a numerical number that can represent the degree of human activity or the intensity of appliance use on a particular day. This dataset is generally well-suited for examining trends in energy use, comprehending how temperature and activity levels affect power demand, and creating models to anticipate or control summertime energy peaks.

IV. Result Analysis and Performance:

The performance of several machine learning models used for energy consumption predictions is extensively investigated in this section. Historical data on energy usage, together with related characteristics like temperature, humidity, time of day, and calendar factors, was used to train and evaluate the models. Evaluating each model's predictive power, accuracy, and robustness over various time periods was the main objective.

The forecasting ability of the machine learning model utilized in this study is shown graphically in the "Actual vs. Predicted Energy Consumption" scatter plot. A single model forecast compared to the actual recorded energy consumption for the same instance is represented by each point on the graph.

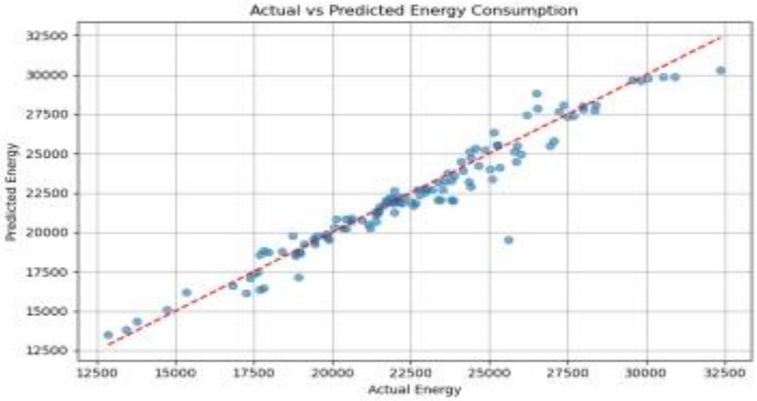

Figure 2: Actual Vs Predicted Energy Consumption

Three common regression metrics were used to assess the machine learning models' forecasting accuracy:

Mean Absolute Error, or MAE, calculates the average size of prediction mistakes without taking direction into account. Root Mean Squared Error, or RMSE, is more sensitive to outliers and penalizes greater errors more severely than MAE.

The percentage of the variance in the dependent variable that can be predicted from the independent variables is represented by the R2 Score (Coefficient of Determination). A better fit is indicated by a higher number that is nearer 1.

Table1: Accuracy of Models

	MAE	**RMSE**	**R2**
Linear Regression	714.359172	925.431599	0.939948
Decision Tree	769.793220	1501.457350	0.841925
Random Forest	615.304034	941.653462	0.937824
XGBoost	591.655873	881.032733	0.945572
Gradient Boosting	575.412316	850.587164	0.949269

Gradient Boosting performed best on all three metrics, with the highest R2 score (0.9493) and the lowest MAE (575.41) and RMSE (850.59). This suggests it is the most accurate and trustworthy model among those tested for projecting energy use.

Furthermore, Random Forest performed well, striking a compromise between model interpretability and error minimization. Its lower MAE suggests superior overall consistency even if its RMSE is slightly larger than Linear Regression. Linear regression showed greater MAE and RMSE but did very well with an R2 of 0.9399. Non-linear patterns in the data probably caused it problems. With the lowest R2 score and the highest RMSE (1501.46), the Decision Tree performed the worst, indicating that it might be over fitting or not generalizing effectively on test data, while being simple to understand.

Additionally, XGBoost outperformed other models, demonstrating its resilience in managing intricate patterns and feature interactions, albeit with somewhat larger errors than Gradient Boosting.

V. Conclusion:

A comparative examination of many machine learning techniques for energy consumption forecasting was reported in this study. Planning for a sustainable power system, cost optimization, and effective energy management all depend on accurate energy forecasts. Using historical energy usage data, the study deployed several models Linear Regression, Decision Tree, Random Forest, XGBoost, and Gradient Boosting and assessed each model's performance using MAE, RMSE, and R2 metrics.

According to the experimental findings, ensemble learning models in particular, Gradient Boosting and XGBoost perform noticeably better than conventional models in terms of accuracy and robustness. With the lowest prediction error and the best R2 score (0.949), gradient boosting proved to be highly effective at identifying complex trends and nonlinear correlations in the behavior of energy use.

This study concludes that machine learning, particularly sophisticated ensemble approaches, provides a dependable and expandable solution for forecasting energy use. Utility companies, building managers, and smart grid systems can use these insights to help them make well-informed decisions about energy conservation, load balancing, and infrastructure design.

References

1. Hyndman, R. J. & Athanasopoulos, G. Forecasting: Principles and Practice, 3rd ed., OTexts, 2021.
2. Breiman, L. “Random Forests,” Mach. Learn. 45, 5-32 (2001).
3. Friedman, J. H. “Greedy Function Approximation: A Gradient Boosting Machine,” Ann. Stat. 29, 1189-1232 (2001).
4. Chen, T. & Guestrin, C. “XGBoost: A Scalable Tree Boosting System,” Proc. KDD 2016, 785-794.
5. Li, K. et al. “Short-Term Load Forecasting in Smart Grids: A Combined Deep Learning Approach,” Energy 188, 116–119 (2019).
6. Lu, M. H. L., Ser, Y. C., Selvachandran, G., Thong, P. H., Cuong, L., Son, L. H., et al. (2022). A comparative study of forecasting electricity consumption using machine learning models. Mathematics, 10(8), 1329.
7. Bilal, M., Kim, H., Fayaz, M., & Pawar, P. (2022). Comparative analysis of time series forecasting approaches for household electricity consumption prediction. arXiv.
8. Alawadi, S., Mera, D., Fernández-Delgado, M., et al. (2022). A comparison of machine learning algorithms for forecasting indoor temperature in smart buildings. Energy Systems, 13, 689–705.