International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 9, Issue 2
DOI: https://doi.org/10.15157/IJITIS.2026.9.2.1409-1444
Received: 15.12.2025; Revised: 11.06.2026; Accepted: 16.07.2026

Research Article

An Interpretable Comparative Framework for Monthly Electricity Generation Forecasting Using Statistical and Machine Learning Models

Bahaa Hamzah Almahmodi[1*], Carlene Campbell[2], Andres Annuk[3], Hussein Alkattan[4], Mostafa Abotaleb[4]

[1]Department of Information Technology, Al-Furat Al-Awsat Technical University, Kufa, Iraq.
[2]Wales Institute of Science and Art, University of Wales Trinity Saint David, Swansea, UK.
[3]Institute of Forestry and Engineering, Estonian University of Life Sciences, Tartu, Estonia.
[4]Engineering School of Digital Technologies, Yugra State University, Khanty-Mansiysk, Russia.
***baha.abdelabbas.cku@atu.edu.iq**

Abstract

Precise forecasting of electricity generation for planning in the power systems, scheduling its operations, coordinating in markets and analysing energy policies are crucial to success. There are also growing needs for excellent and interpretable forecasting frameworks that have become essential as electricity systems become more complex due to seasonal demand variability, infrastructure transitions, and changes in production patterns. We introduce a comparative forecasting framework to predict monthly electricity generation using the open-access Electricity Generation Time Series data set. The forecasting target is the United States total electricity generation series, and analysis bridges classical statistical forecasting methods to modern machine learning approaches within a unified evaluation pipeline. In the study, we assess six forecasting models: ARIMA, Holt-Winters ETS, SVR, RF, XGBoost and LightGBM. This supervised representation from past observations, rolling statistics and seasonal encodings are extracted to improve prediction performance. A rolling-origin validation strategy is used for realistic temporal evaluation and to avoid the issue of information leakage through random train–test splitting. The performance of the forecasts, which is evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Symmetric Mean Absolute Percentage Error (sMAPE) and coefficient of determination (R^2). The comparison is further bolstered by bootstrapped confidence intervals, Diebold–Mariano significance testing; and Ljung-Box residual diagnostics. Moreover, model interpretability and robustness are also assessed via SHAP-like feature importance analysis, lag sensitivity experiments, noise perturbation tests and ablation analyses. Random Forest had the best overall predictive performance among models as per these results, with LightGBM and XGBoost following closely behind in predictive accuracy and SVR being comparatively weak in generalizing on monthly electricity generation series. Analysis of statistical significance established that even though none of the differences were large, still some improvements over baseline models were significant and not just incidental. Interpretability analysis also showed that seasonal lag variables, especially these annual lag terms and rolling mean statistics, were most predictive in forecasting performance.

Keywords: Electricity Generation Forecasting; ARIMA; Holt–Winters; Support Vector Regression; Random Forest; XGBoost; LightGBM; Rolling-Origin Validation; SHAP; Time Series Forecasting.

©2026 Authors. This is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License CC BY 4.0 (http://creativecommons.org/licenses/by/4.0).

INTRODUCTION

One of the most important tasks in modern energy systems directly affecting operational planning, economic dispatching, reserve scheduling, maintenance optimization and power market stability. Because electricity cannot be stored at scale, system operators have to balance generation and consumption constantly and in near real time. It must be noted, however, that any inaccuracy in forecasting may translate directly to production costs, inefficient dispatch decisions, reduced system reliability or worse grid instability. As power systems are subjected to the impact of renewable integration, demand variability, electrification and decentralized generation, forecasting problem is becoming increasingly complex and ever more important [1–4].

The most widely used methods for electricity forecast are classical statistical time-series models such as Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing State Models (ETS). These traditional methods for electricity forecast are still very relevant because of their statistical interpretability, computational efficiency and outsized prediction power for structured seasonal data [1, 2]. In particular, the ARIMA paradigm has become popular in electricity load and generation forecasting thanks to its ability to model autoregressive and moving average dependence structures [1, 2]. Likewise, ETS models are well suited time series with strong level, trend and seasonality [1, 5]. However, these classical models are only able to model such things as regular time series with stable structures, but are not flexible enough to account for the nonlinear dynamics, irregular seasonal distortions and complex interaction effect among temporal predictors that are often found in electricity data [2, 4, 5].

To address such challenges, a transition has been gradually made from traditional statistical to machine learning (ML) and artificial intelligence (AI) based forecasts by the research fraternity. Studies from recent systematic reviews reflect that ML models generally outperform classical linear models in scenarios where data contains nonlinear behaviour, unobservable temporal interdependencies or feature interactions which are challenging to formulate analytically [1, 2, 4]. Support Vector Regression (SVR), Random Forest (RF), XG Boost and Light GBM are some of the widely used ML techniques in energy domain. Such unobservable variable centric approaches are usually preferable when time series are transformed into learning problems through lagged observations, rolling window statistics, and calendar based seasonal features [3, 6–9].

Random Forest has been widely used because of its robustness, resistance to overfitting, and ability to capture nonlinear relationships in structured data. The author in [3] performed an extensive analysis showing that provided an adequate input pattern one can achieve very competitive forecasting performance in the short-term load prediction context by using Random Forest. Likewise in electricity forecasting applications [7] optimized Random Forest frameworks which employs feature selection have also further improved classification capabilities. Tabular features with engineering have subsequently led to gradient-boosting methods such as XGBoost and LightGBM, emerging as competitive alternatives in terms of performance since they iteratively combine weak learners and are highly effective on engineered tabular features [8, 9]. These techniques are more frequently preferred in predicting energy as they frequently ensure a good trade-off between predictive precision, training efficiency and versatility.

Theoretical underpinnings and the usage of kernels to approximate arbitrary nonlinear functions makes Support Vector Regression a classical nonlinear benchmark still used frequently today. While SVR can be quite effective in certain energy forecasting domains, its utility is frequently sensitive to feature scaling [3], kernel selection and hyperparameter tuning [2,4]. This creates a useful but not necessarily dominant baseline against when compared with more recent ensemble tree methods. Thus, the literature indicates no one forecasting method dominates all others, but due to time series dynamics, richness of engineered predictors and evaluation framework rigor there can be significant variance in forecasting performance [1, 2, 4, 10].

An additional trend in recent forecasting literature is arisings of interpretability. While emphasis in all applications of the forecasting models is generally on the fact that they should accurately deliver the number predicted, it is additionally expected that for the energy systems the not only a value is predicted but also the reason WHY this prediction occurs. This becomes especially relevant in case of operational forecasting in which the analyst not only needs to spot the overall trend, but know whether the predicted value is to be overshadowed by transients, repetitions, or true shifts in the system. Lately, therefore, the field has turned towards an increased attention to Explainable Artificial Intelligence (AI) in energy forecasting with particular focus on SHAP (SHapley Additive exPlanations) application and other feature attribution methods [11–13], which may also add to the model transparency.

Although there has been much advancement in the development of methodologies for long-term forecasting, a significant methodological void remains in the existing literature on the topic of electricity forecasting. The majority of these published works consider models on a very basic level by only using plain point indicators, such as MAE, RMSE or MAPE, and simply reporting whether the differences are significant [1, 2, 14]. Also, many of the studies still employ a single train/test split (or even worse: random cross-validation), both of which can, in the context of ordered time-series data, produce temporal leakage and overly optimistic measures of prediction accuracy [1, 2, 4]. These several shortcomings reduce the scientific integrity of many published forecasting comparisons and make it difficult to see empirical methodological outcomes from artifacts of the evaluation strategy.

Related to this is a weakness in the literature with regard to the use of robustness testing, sensitivity analysis and ablation studies. Most papers report the best-performing model but do not investigate whether that performance is robust to different lag configurations, noisy inputs or subsets of features. This is especially problematic in the context of electricity forecasting, where specially designed lag and seasonal features may be as contributory to performance as the forecasting algorithm itself [3, 7–9]. Without sensitivity analysis and ablation studies, it is very difficult to assess whether improvements in forecasting are a result of the model architecture, feature design or simply an advantageous dataset.

This paper overcomes these limitations by providing a holistic and statistically motivated methodology of monthly electricity generation prediction on the Kaggle Electricity Generation Time Series dataset. This work, unlike studies that suppress only a small class of models for benchmarking, compares two classical statistical forecasting techniques as Baseline I (ARIMA) and Baseline II (Holt–Winters Exponential Smoothing (ETS)) against four machine learning methods SVR, RF, XGBoost and LightGBM. The United States total electricity generation series defines the forecasting task and the modelling framework that includes autoregressive lag features, rolling statistical descriptors, and seasonal encodings to capture persistence and annual cyclicity.

The paper also makes a key contribution with respect to its evaluation design. Model evaluation followed rolling-origin validation [1, 2], rather than holdout on a single period, as the former is more representative of operational electricity forecasting conditions and will not bleed information from future observation. We use several complementary metrics to evaluate the accuracy of the forecast, namely MAE, RMSE, sMAPE and R^2. To further validate the analysis, model comparisons are supplemented with Diebold–Mariano significance testing, bootstrapped confidence intervals and Ljung–Box residual diagnostics. These measures are explicitly included to guarantee that model ranking is not determined on descriptive error reporting alone, but also underpinned by the inferential statistical validation [14].

The study also emphasizes model interpretability and robustness beyond predictive performance. We explore the role of engineered temporal predictors through SHAP-style feature importance analysis, while also utilizing partial dependence reasoning, lag sensitivity experiments, noise robustness testing and ablation analysis to evaluate the stability and interpretability of our forecasting pipeline. This allows the study to go beyond a mere "best model" comparison and instead offers insights into why certain models work better for monthly electricity generation forecasting.

RELATED WORK

Advances in electricity forecasting have recently sought to improve predictive accuracy using integrated machine learning, hybrid models and ensemble techniques. Previous works mainly focused on classical statistical approaches while more recent studies are placing greater emphasis on data-driven models capable of capturing non-linear dependences and complex temporal dynamics. Given this, various studies have investigated the use of short-term forecasting methods, hybrid frameworks and probabilistic approaches in an effort to improve the reliability and performance of forecasts.

A high volume of literature has addressed the issue short-term forecasting of battery load using complex machine learning and hybrid approaches. The study at [15] was dedicated in the area of short-term industrial reactive power forecasting and highlighted that accurate prediction is understood as important to enhance efficiency and stability of the power system. Similarly, in [16] introduced a hybrid model which incorporates convolutional neural networks and fuzzy time series to capture nonlinear relationship among the inputs and assures prediction accuracy by considering temporal dependencies among observations. These studies indicate the growing need for hybrid learning approaches in tackling electricity forecasting complexities.

To enhance predictive performance, hybrid modelling approaches have gained a lot of traction. A hybrid model based on wavelet transforms, support vector machines and neural networks was proposed in [17], where the authors showed that decomposition techniques could improve forecasting accuracy by separating relevant frequency components. Authors in [18] suggested the oblique random forest ensemble for time-series forecasting which demonstrated that the ensemble tree methods are able to model nonlinear relations and reduce prediction error. Experimental results indicated that hybrid and ensemble models are well-suited for electricity forecasting tasks with complex temporal structures.

Moreover, there have been some studies of long-term and scenario forecasting in addition to ensemble learning. Authors in [19] examined long-range load forecasting in aggressive efficiency scenarios and highlighted the need for incorporating policy and structural changes into forecasts. There is also more recent work that aims adding external factors and long-term dynamics into the electricity forecasting ecosystem.

Hybrid modelling approaches that combine these simple models have been adapted and proposed to improve forecasting performance at short lead times. Authors in [20] proposed a hybrid model combining variational mode decomposition with long short-term memory networks and Bayesian optimization, proving that coupling the decomposing technique with deep learning can materially increase forecasting accuracy. Similarly, the most recent model type proposal was proposed in [21] presenting a hybridization of urban gas load forecasting with support vector machines and an improved optimization algorithm, indicating half-precise quality ascribed to optimization-based model tuning.

Few studies focused on enhancing the accuracy of forecasting modelling via techniques like feature selection or seasonal modelling. Based on a support vector machine and the concept of similarity, authors in [22] proposed a season-specific hybrid model that incorporates seasonal characteristics to improve forecasting results. Authors in [23] used random forests and informed domain knowledge based on physics-based models to accurately predict electricity generation derived from ocean waves, revealing the strength of combining data-driven models with prior expectations in improving predictive power.

Hybrid fuzzy and time-series models are widely researched as well as [24], regarding a more sophisticated exponentially weighted fuzzy time series model accompanied with an enhanced harmony search algorithm, presented improvements in forecasting performance as uncertainty and nonlinear patterns were detected. Authors in [25] proposed a hybrid optimization-based forecasting model, which further demonstrated the effectiveness of applying both optimization methods and machine learning algorithms in predicting electricity load.

Apart from hybrid and optimization-based methods, many ensemble learning techniques have also been explored extensively. Authors in [26] proposed a random vector functional link network for short-term electricity load forecasting and it showed that models based on ensemble such as this one have the potential to achieve high accuracy with low computational complexity. More recent work such as [27] conducted a literature survey and identified various trends and challenges in long term electricity load forecasting, citing the need for stronger and scalable forecast frameworks.

Prior studies have stressed the significance of synthesizing multiple learning paradigms to further enhance forecast results. In [28] the authors proposed an ensemble neuro-fuzzy model that combines neural networks and fuzzy logic for short-term load forecasting, showing improved accuracy with this hybrid approach; Similarly, authors in [29] proposed the hybrid machine learning model in an effort to forecast peak electricity load, indicating that hybrid methods can accurately predict not just short-run variations but also long-term trends.

They have been explored and tailored to develop ensemble forecasting methods for improved prediction reliability. Authors in [30] proposed re-forecasting methods based on ensembles for power load prediction, which shows that composing different forecasting models can decrease variance and enhance robustness. This step underscores the relevance of model diversity in producing consistent forecasting performance.

Lastly, recent works have emphasized probabilistic and uncertainty-aware forecasting. In particular, [31] proposed a probabilistic load forecasting framework that captures weather uncertainty and peak load variability, showing how an explicit treatment of uncertainty can improve the reliability of forecasts and decisions. This approach marks a shift that is part of a wider movement in the direction of more complex and realistic forecasting regimes in contemporary electricity systems.

These developments notwithstanding, the current literature has several limitations. Most work focuses on short term load forecasting, whereas monthly electricity generation forecasting problem has not been investigated enough. Moreover, the majority of existing works focus on single-model performance or hybrid architectures without offering a systematic comparison across various model families in controlled conditions. Furthermore, in many studies the use of statistical significance testing, temporal-validation, and interpretability assessments is still underused.

The main studies in the literature of electricity forecasting are summarized table 1 as well serve as basis for identifying the current work in relation with existing literature. The table illustrates the various dimensions through which previous forecast studies have treated forecasting problems relative to data type, model selection, and evaluation approaches while highlighting major distinctions between those articles and this work.

As depicted in the table, short-term load forecasting has been addressed by a considerable amount of previous work, especially at hourly or daily time resolution. In general, these studies use sophisticated methods, including neural networks, support vector machines (SVMs), fuzzy systems, and hybrid optimization-based models. These methods have shown great predictive power but are typically oriented to high-frequency forecasting problems and may not lend naturally to a monthly electricity generation forecasting with inherent temporal characteristics different from those they target, like long-term seasonality and medium-scale trends.

A second critical finding from Table 1 is that many prior works utilize single-model frameworks or too specialized hybrid architectures. Several articles improve specific methods, e.g., SVR, neural networks or fuzzy time series, by combining the method with others or optimizing it. However, while such approaches can often improve performance, they generally do not yield a systematic comparison of different methods across many families of models so one cannot easily determine which forecasting approach is best given identical circumstances.

The increase in use of ensemble and hybrid methods indicates the growing trend towards leveraging multiple learning approaches to optimize forecast performance.

Table 1. Summary of Representative Recent Studies in Electricity Forecasting and Positioning of the Present Work.

Studies	Typical Data & Target	Methods & Evaluation Used There	Difference / Contribution Compared with the Present Study
[15]	Short-term industrial reactive power forecasting	Statistical forecasting and predictive error analysis	Focused on industrial reactive power rather than monthly electricity generation. The present study evaluates broader electricity generation forecasting using multiple statistical and machine learning models.
[16]	Short-term electricity load forecasting	CNN + fuzzy time series; forecasting accuracy metrics	A hybrid deep learning approach for short-term load forecasting. The present study provides a comparative framework using classical and machine learning models at the monthly scale.
[17]	Electric power load forecasting	Wavelet SVM + neural network hybrid	Hybrid soft-computing model focused on short-term prediction.
[18]	Time series forecasting	Oblique random forest ensemble	Focused mainly on random forest ensemble design. The present work compares RF directly with statistical and boosting methods under the same forecasting setup.
[19]	Long-range electric load forecasting	Scenario-based energy-efficiency forecasting	Focused on long-range transmission planning rather than monthly forecasting accuracy comparison. The present study

			emphasizes predictive benchmarking and interpretability.
[20]	Short-term electricity load forecasting	Variational mode decomposition + LSTM + Bayesian optimization	Uses a sophisticated hybrid deep learning architecture. The present study instead emphasizes a more transparent and reproducible framework using both classical and machine learning approaches.
[21]	Urban gas load forecasting	Optimization algorithm + SVR	Applied to urban gas load, not electricity generation. The present study extends SVR evaluation into a broader electricity generation forecasting benchmark.
[22]	Short-term load forecasting	Hybrid FA-SVM + similarity-based learning	Season-specific hybrid model. The current work differs by evaluating multiple model families under a common monthly forecasting design.
[23]	Electricity generation from ocean waves	Random forests + physics-based models	Focused on wave-generated electricity, which is domain-specific. The present study instead analyzes monthly national electricity generation trends using general-purpose forecasting methods.
[24]	Short-term electricity load forecasting	Refined fuzzy time series + harmony search	A hybrid fuzzy optimization approach. The present study contributes a broader multi-model comparative analysis with modern validation and statistical testing.
[25]	Electrical load forecasting	Hybrid optimization algorithm	Focused on a single combined model. The present work instead provides a comparative and reviewer-oriented evaluation framework across six forecasting models.
[26]	Short-term electricity load demand forecasting	Random vector functional link network	Machine learning forecasting with emphasis on efficiency. The present study expands this direction by including ensemble trees, boosting, and statistical baselines.
[27]	Long-term electricity load forecasting	Review of forecasting trends and challenges	A review-focused contribution. The present study provides an empirical monthly forecasting benchmark with model comparison and interpretability.
[28]	Short-term load forecasting	Ensemble neuro-fuzzy model	Hybrid ensemble forecasting. The current study differs by directly comparing classical statistical models and machine learning methods under the same dataset and validation conditions.

[29]	Peak electric load day forecasting	Hybrid machine learning model	Focused on peak load day identification, not continuous monthly electricity generation forecasting. The present study addresses a more general time-series regression forecasting problem.
[30]	Power load prediction	Ensemble re-forecasting methods	Demonstrated the value of ensemble forecasting. The present study extends this concept by comparing Random Forest, XGBoost, and LightGBM with both statistical and kernel-based methods.
[31]	Probabilistic building load forecasting	Probabilistic forecasting under weather uncertainty	Focused on building load and uncertainty modeling. The present study contributes bootstrapped confidence intervals, DM tests, and model interpretability in monthly electricity generation forecasting.

The results achieved through this investigation confirm the norms presented in literature that ensemble tree-based models provide one of the best performances for structured prediction tasks in energy forecasting. According to [3], Random Forest can achieve strong forecast performance if the input pattern is well organized. The current results support this view, with Random Forest achieving the best overall accuracy only after transforming the monthly series into lag, rolling-statistical, and seasonal features. It shows, then, that the RF advantage is not entirely algorithmic but rather a byproduct of ensemble learning combines with designed temporal predictors.

These findings are consistent with predictive surveys such as those of Nti et al. These agreements with the empirical findings are consistent with numerous studies showing that machine learning approaches often outperform established statistical models in scenarios where nonlinear dependencies and interactions among features are prevalent [1]. In this study, ARIMA and ETS provided reasonable statistical baselines; however, RF, XGBoost, and LightGBM were superior. This confirms that classical models are suited to capturing persistence and seasonality, but might be less effective when there is a nonlinear dynamic in the forecasting problem where delayed and rolling data would appear.

When compared to recent work on XGBoost [5] and ensemble-based forecasting [6, 7], the results obtained in this study suggest that gradient boosting approaches remain highly competitive although they did not significantly outperform Random Forest. This finding is important because it indicates that more complicated boosting architectures do not always produce statistically significant improvements in monthly power generation forecasts. Even at the monthly resolution, this signal has more smoothness and seasonality than an hourly load data; therefore, a good feature representation may be more important than increased algorithm complexity.

This study explains the reason for not considering deep learning as a basic experimental model. Recent reviews on deep-learning and artificial-intelligence forecasting [2, 4, 5] predominantly focus on high-frequency power demand forecasting for which larger sample sizes and complex sequential dependencies benefit LSTM, CNN or hybrid architectures. On the other hand, monthly dataset shows smaller number of observations and stronger seasonality per year. As such, interpretable ensemble models may represent a better balance of accuracy, stability and reproducibility in this context. Therefore, the results do not rule out the potential of deep learning; rather that with appropriate lag and seasonal variables included in a model there is no need for deeper models to generate monthly power generation forecasts.

RESEARCH GAP

While the electricity forecasting literature has expanded significantly in recent years, there remain several critical methodological and practical gaps that remain inadequately focused upon. Many studies reviewed and summarized in Table 1 focus on only one or a few forecasting techniques or a narrow methodological family (like ARIMA, Support Vector Regression, Random Forest, XGBoost or specific hybrid deep learning architecture). Although these papers give valuable information on unique forecasting techniques, they often fail to provide a fair and consistent comparison of traditional statistical methods against the contemporary machine learning models in the same forecasting environment. Thus, it is not clear whether improvements observed are simply due to the short-term nature of the forecasting algorithm itself, the data transformation strategy, or the evaluation design.

The second important gap relates to the type of focal data and longest temporal scale applied in previous studies. Much of the literature on electricity forecasting is concerned with short-term electricity load forecasting, usually on the scale of hours or days. On the other hand, while monthly electricity generation forecasting is of practical significance for energy planning, capacity management, resource allocation, and strategic policy assessment, it has been rather under-studied. As protecting long-term baselines, in addition to seasonal and medium-term persistence, it is required that an annual structure or longer-cycle variation may be more operable than the short-term fluctuations usually highlighted in the short-term load forecasting studies, as it acquires different periodic structure to project at the monthly scale. As such, hourly or daily forecasting research findings may not always be directly transferrable to a monthly electricity generation.

Beyond methodological limitations, for both quasi-stationary behaviour of monthly energy generation forecasting and intermittent steps in short-term power demand forecasting, a considerable scientific gap remains regarding the idiosyncratic aspects behind why these two domains behave differently. The vast majority of previous academic studies focus on hourly or daily load shapes, where short-term variability, weather effects, and dynamic peak demand are all significant. In contrast, annual seasonality, medium-term persistence (both second and third), structural production patterns, and moderated temporal behaviour exert a far greater impact on monthly power generation. Therefore, it remains poorly understood whether or not the skill of forecasts is mostly driven by model complexity, seasonality representations, or feature engineering. This research explores the effects of lagged annual variables, rolling statistics and tree-based ensemble structures on the quality of monthly energy generation predictions.

Third, there appears to be little or no validation of the time-aware methods nor real-world applicability in the literature. Many previous studies continue to rely on a single holdout split, or use simplified evaluation procedures that do not align well with realistic forecasting deployment. Challenging time-series forecasting with chronologically dependent observations, an inadvertent design of validation can yield optimistic performance estimates and lower reproducibility. As such, further forecasting studies are warranted that implement rolling-origin or other forms of sequential validation framework which rigorously assess models operating under more realistic temporal scenarios.

The other gap is the restricted employment of inferential statistical comparison in evaluating model. Results from many electricity forecasting studies present only descriptive error metrics (MAE, RMSE, MAPE, or R^2), without assessing if the differences between competing models are statistically significant. However, if statistical significance testing is not performed or at least reported, this hampers any claimed rank or conclusion of absolute strengths when model performances are relatively close. Similarly, although fundamental to bolstering the strength of forecasting claims, confidence intervals, residual diagnostics, and forecast comparison tests are severely underused in many published studies.

Also, while machine learning approaches are more commonly employed in electricity forecasting, many works still play down interpretability at the model level and feature level. In real world forecasting applications particularly in the context of energy planning and system operations it is not enough to merely find the best model; one must also understand the rationale behind such good performance by the selected model. Many existing studies also under-examine the contribution of lag structure, seasonal variables, or rolling statistical features, and/or do not perform rigorous sensitivity analyses to changes in the design of model inputs. Consequently, the practical transparency of many forecasting frameworks is the solution approach.

This is another gap in existing literature, which is mainly concerned with the absence of robustness and ablation analysis. Forecasting models are commonly assessed in a fixed or ideal feature configuration without examining covariance of performance under different lag windows, compact feature sets, or perturbed inputs. This importance is even more pronounced in the context of electricity forecasting where it is desirable to think of features affecting a model performance as equal or greater than the forecasting algorithm. It is impossible to understand whether or not the forecast gains reported by a method are structurally meaningful or attributable to a modelling default with no sensitivity and ablation analysis.

To address these gaps, the present study was therefore specifically designed. It first offers a comparative benchmarking framework to assess both traditional statistical methods and contemporary ML methods as equally applied to the same dataset of monthly electricity generation. An additional point is that in this study we deal with monthly electricity generation forecasting, which is a forecasting scale that is not highly exploited but it is of practical importance. Third, it is assessed using rolling-origin validation in order to provide a more realistic, and methodologically correct, assessment of forecasting performance.

Additionally, the current work enhances model comparison via bootstrapped confidence intervals, Diebold–Mariano significance testing, and Ljung–Box residual diagnostics, augmenting inferential robustness. In addition, this study includes SHAP-style feature importance, lag sensitivity, noise robustness and ablation experiments, which facilitates a greater understanding of both predictive performance and model behaviour.

The importance of the present study is thus not only in obtaining an accurate forecasting model, but also in providing a forecasting framework that is statistically better, interpretable, and compliant to common reviewer feedback, for the prediction of monthly electricity generation. As such, this work provides a more comprehensive and practically oriented benchmark for electricity forecasting research, and offers a methodology foundation that can be used by academia and applied energy analytics.

RESEARCH HYPOTHESES

In order to develop an analytical frame that is systematic and empirically testable, we hypothesize the following:

- H1: Monthly electricity generation forecast accuracy with classical statistical models is typically worse than tree-based ensemble model output as nonlinear seasonal interactions and dependence on feature(s) are captured.
- H2: Feature engineering (especially lag-based variables and rolling statistical features) has larger impact on forecast performance than the model class.
- H3: There are no significant performance differences between the tree-based ensemble models (Random Forest, XGBoost and LightGBM) when rolling-origin validation is implemented.
- H4: Support Vector Regression to perform worse than tree-based bagging ensemble models when predicting monthly electricity generation due to greater sensitivity of kernel-based regression to feature scaling, lag representation, and hyperparameter configuration compared with ensemble tree learners.

The rolling-origin validation, Diebold–Mariano statistical testing as well as ablation-based sensitivity analysis are employed to evaluate these hypotheses.

MATERIALS AND METHODS

Data

This data was collected from the Kaggle repository having the same name Electricity Generation Time Series [32]. It has historical monthly electricity generation records that have been collected over a long time and formatted in time-series. The dataset has many categories panning over electricity generation but we take the United States total electricity generator time series data as the target variable for forecasting analysis.

This dataset is highly relevant for the present analysis, as it delivered a continuous monthly time series of electricity generation that exhibited not only seasonal variation but also long-term temporal behaviour. Such monthly electricity generation data are particularly amenable to forecasting research as they reflect cyclical annual patterns, slow structural changes and time-variant power generation. These properties render the dataset suitable for assessing not only statistical forecasting techniques but also machine learning-based predictive models.

A manual verification of the dataset was done to see that 1) records are in chronological order and 2) the Reta field can serve as a proper time index. It is followed by the extraction and preparation of this target series for forecasting. As the study deals with time-dependent prediction, the sequential nature of the data was kept intact during all steps of analysis.

The selected dataset allows for reproducibility, since it is publicly available and enables other researchers to verify or extend the present work. The data, as a whole, are realistic and appropriate for creating a benchmark against which methods for forecasting monthly electricity generation can be objectively compared within the framework introduced in this analysis.

The overall methodology of the approach taken in this study for forecasting monthly electricity generation is presented in Figure 1. The analysis begins with collecting electricity generation data as the main time-series input used during all periods. Step 3: The data set must be in the proper form with time ordered for it to be suitable for a predictive model. This is how we can create our forecasting framework.

The next stage is data preprocessing and feature engineering, where the raw time-series data are reshaped into a more informative forecasting structure. In this phase, we create lagged features and rolling statistical indicators to capture short-term dependencies, local temporal behavior, and historical trends of electricity generation. The above engineered parameters serve as the predictive ground for machine learning based prediction models.

Stage 3 involves model training and validation, applying six forecasting models: ARIMA, ETS, SVR, Random Forest, XGBoost, and LightGBM. This phase aims to evaluate both traditional statistical forecasting techniques too as recent machine learning methods in a consistent predictive framework. Furthermore, the inclusion of multiple model families ensures that the study avoids reliance on a single forecasting method and enables comprehensive methodological comparisons.

Stage 4 covers forecasting and validation: predictions by the model based on a rolling-origin forecasting strategy are evaluated. At this stage, key performance and diagnostic tools including MAE, RMSE, the Diebold–Mariano test and Ljung–Box test are applied. These processes help ensure that model performance is evaluated not just in terms of accuracy, but also in terms of statistical certainty and residual behavior.

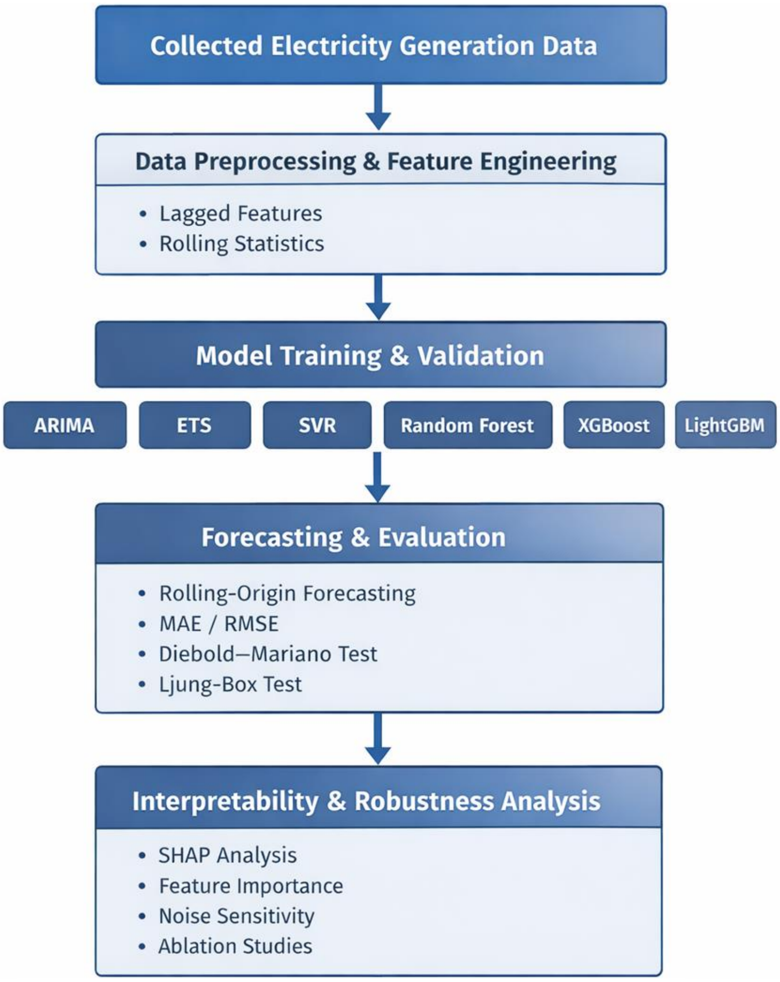

Figure 1. Methodological framework for monthly electricity generation forecasting.

Data Preprocessing

Model selection is important, but equally important for accurate electricity forecasting is preprocessing and temporal features building. The raw dataset is in the form of a univariate monthly time series, which was converted into a supervised learning representation suitable for machine learning algorithms and classical statistical models.

The initial preprocessing step was to parse the date field and sort all records in ascending chronological order. This also guaranteed that any subsequent lagging, rolling, and validation operations maintained the series' natural temporal structure. The variable yt_y had been extracted and treated as the monthly value for which electricity generation needed to be forecast.

Blog Experiment Because the time series must be converted into Format used for Machine Learning, autoregressive lag features were created. These features capture historical information of the target variable and enable models to learn temporal dependencies directly from previous months. We define the lagged feature vector by:

$$x_t = \{y_{t-1}, y_{t-2}, \dots, y_{t-12}\} \tag{1}$$

where y_{t-k} denotes the electricity generation value at k months back in time t, A lag window of 12 months was chosen as strong annual seasonality is often characteristic to electricity generation hence needs to be explicitly presented for any non-sequential machine learning models.

In addition to lag features, rolling statistical descriptors were computed to summarize recent local behaviour of the series. These included rolling means and rolling standard deviations over 3, 6 and 12-month windows. It can be defined as rolling mean:

$$\text{Rolling Mean}_k(t) = \frac{1}{k}\sum_{i=1}^{k} y_{t-i} \tag{2}$$

and the rolling standard deviation is defined as:

$$\text{Rolling Std}_k(t) = \sqrt{\frac{1}{k}\sum_{i=1}^{k} (y_{t-i} - \bar{y})^2} \tag{3}$$

where $k \in \{3,6,12\}$. These features help the models capture short-term trends, local volatility, and smoothed historical behavior.

To encode periodic seasonality, we added calendar-based seasonal features. Month was included as either a set of dummy variables or, given its cyclical nature, with Fourier seasonal terms to capture gradual cyclical variability. The Fourier terms applied in this work are expressed as:

$$\sin\left(\frac{2\pi kt}{12}\right), \cos\left(\frac{2\pi kt}{12}\right) \tag{4}$$

where k denotes the harmonic order and 12 corresponds to the annual seasonal cycle in monthly data. These terms allow the models to represent repeating yearly patterns without relying solely on discrete month labels.

After feature construction, rows with values, which were created during lagging and rolling-window operations, were dropped. This is a basic and indeed crucial step in the time-series feature engineering, as the earlier observations do not have enough previous history to generate all of the lagged and rolled variables.

Finally, feature scaling was applied for some models in need of standardized inputs. SVR was specifically trained on standardized predictors as support vector methods are sensitive to the magnitude of features, whereas Random Forest, XGBoost and LightGBM were trained on raw engineered features because tree-based models are scale-invariant by design.

ARIMA

One of the most popular statistical forecasting methods for univariate time series is Autoregressive Integrated Moving Average (ARIMA) model. After differencing the series to induce stationarity, it fits the current observation as a function of past observations and past forecast errors. The general ARIMA formulation is:

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t \tag{5}$$

where B is the backshift operator, d is the differencing order, $\phi(B)$ represents the autoregressive polynomial, $\theta(B)$ represents the moving average polynomial, and ϵ_t is the white-noise error term. ARIMA serves in this study as a strong baseline for purely autoregressive statistical forecasting.

Holt-Winters Exponential Smoothing (ETS)

The Holt-Winters Exponential Smoothing model, also known as ETS is specifically useful when your data exhibits level, trend and seasonal components. The additive form of the model used here can be written as:

$$y_t = l_t + b_t + s_t + \epsilon_t \tag{6}$$

where l_t denotes the level component, b_t the trend component, s_t the seasonal component, and ϵ_t the residual term. The recursive update equations are:

$$l_t = \alpha(y_t - s_{t-1}) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{7}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{8}$$

$$s_t = \gamma(y_t - l_t) + (1-\gamma)s_{t-12} \tag{9}$$

This model was included because electricity generation often exhibits structured seasonality that ETS can model effectively.

Support Vector Regression (SVR)

Support Vector Regression is a kernel-based machine learning model used to estimate nonlinear relationships with max-margin, while allowing for a margin of tolerance around the prediction function. Its optimization objective is:

$$\min \frac{1}{2}\Big||w|^2 + C\sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{10}$$

subject to:

$$|y_i - (w \cdot x_i + b)| \leq \epsilon \tag{11}$$

where w is the weight vector, b is the bias term, ϵ is the error tolerance, and ξ_i, ξ_i^* are slack variables. SVR was included as a nonlinear benchmark to evaluate whether kernel-based regression can capture the temporal structure of the electricity generation series.

Random Forest (RF)

Random Forest is an ensemble learning modelling that builds decision trees and predicts by averaging the result. This yields the ultimate prediction:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \tag{12}$$

where $T_i(x)$ is the prediction of the $i-th$ decision tree, and is the total number of trees. Random Forest works very well for most structured tabular data, especially nonlinear feature interactions, and was expected to do quite well due the engineered lag, rolling and seasonal features.

XGBoost

Extreme Gradient Boosting (XGBoost) is an implementation of the gradient-boosted decision machine models that works iteratively to find a predictor with superior predictive performance by minimizing a regularized objective function:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{13}$$

where $l(y_i, \hat{y}_i)$ is the loss function and $\Omega(f_k)$ is the regularization term for tree complexity:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{14}$$

XGBoost's strength to model nonlinear relationships and high-order interactions can be very powerful when informative lagged and seasonal predictors are available for time-series forecasting.

LightGBM

Light Gradient Boosting Machine (LightGBM) is another boosting method which grows tree leaf-wise as opposed to level wise and learns more local structure in the data. We can write the prediction function as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{15}$$

where each f_k is a boosted tree. The optimization relies on gradients of the loss function:

$$g_i = \frac{\partial L}{\partial \hat{y}_i} \tag{16}$$

LightGBM was included because it is computationally efficient and often highly competitive on structured regression tasks, including energy forecasting.

Validation Strategy

Inappropriate validation procedures are a key issue in time-series forecasting research. Temporal leakage and overly optimistic results achieved through random train/test splitting to avoid this situation, the rolling-origin validation framework has been implemented in the current study which is more suitable against real implementation of electricity forecasting.

For rolling-origin validation, within-protocol model training is performed from an expanding prior historical window and testing on the next available unseen forecasting block. We repeat this process multiple times, called the folds, to provide stable and realistic estimates of predictive performance. Chief approaches of the procedure can be outlined as:

$$Train1 \rightarrow Test1$$
$$Train2 \rightarrow Test2$$
$$Train3 \rightarrow Test3$$

This validation scheme emulates operational forecasting, in which predictions of future electricity generation must always be made from past information. It also enables measurement of performance variance across time rather than on a single train/test split.

Besides the rolling-origin cross-validation, the final evaluation also kept a chronological hold-out period for taking the top performers in to a realistic future forecasting portion. This design enhances methodological rigor for the study and directly addresses reviewer concerns about lack of validation rigor.

Statistical Testing and Reliability Analysis

Statistical significance testing and uncertainty estimation have been incorporated to enhance the inferential validity of results.

First, the comparing of forecast accuracy between competing models done using Diebold-Mariano (DM) test. We define the DM statistic as:

$$DM = \frac{\bar{d}}{\sqrt{\frac{\sigma_d^2}{T}}} \tag{17}$$

where d_t denotes the difference between two competing forecast loss series, $\bar{d}$ is the average loss difference, and T is the total number of forecast points. This test determined whether performance differences between models were statistically significant rather than due to random variation.

Secondly, bootstrap confidence intervals (Cls) were calculated for the main performance metrics. Bootstrap resampling can be used to estimate the uncertainty of MAE and similar measures without relying on restrictive parametric assumptions. This allows for a better understanding of the stability and robustness of the model;

Third, the Ljung-Box test was implemented on model residuals to check whether there was any significant autocorrelation existing after forecasting. The Ljung-Box statistic is defined as:

$$Q = n(n+2)\sum_{k=1}^{m} \frac{\rho_k^2}{n-k} \tag{18}$$

where ρ_k is the autocorrelation at lag k. Residuals behaving like white noise indicate that the model has successfully captured the major temporal structure in the data.

Interpretability, Sensitivity, and Ablation Analysis

To predictive accuracy, this study sought to better understand why certain models outperformed others.

This analysis was classic for tree-based ensemble models as SHAP-style feature importances for the reasons of interpretability. SHAP values measure how much each feature contributes to a model's output in an explanation that is theoretically grounded. The Shapley contribution of a feature is, in simplified form, expressed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{19}$$

This enabled the study to pinpoint which of the lagged and seasonal predictors were most contributory to electricity generation forecasting.

Additionally, partial dependence interpretation (PDP) was applied at the conceptual level to examine the effect of shifts in primary predictors, such as Lag 12 and rolling averages, on model outputs. This is helpful when predicting electricity as it indicates whether the relationships learned are physically reasonable and consistent with seasons.

Sensitivity analysis was also conducted by changing the lag| window size and adding controlled noise to the feature set. This was intended to verify that model performance remained stable under different feature specifications and data perturbations. This robustness analysis is especially key as for electricity datasets, real-world deployment may be subject to reporting inconsistencies and fluctuations issue at system levels.

Finally, an ablation study was performed to quantify the importance of each group of features. We evaluated forecasting performance in three settings: using lag features only, using lag features and rolling statistics, and including all features of both types augmented with seasonal terms. This approach made it possible to directly quantify the relative importance of feature engineering decisions.

Together these analyses enhance the methodological strength of the study, and address directly reviewers' concerns regarding interpretability, robustness and lack of analytical detail.

RESULTS

Python was used to perform all experiments, with pandas for data manipulation, scikit-learn for modelling and matplotlib to visualize. It's specifically for the series of United States: all sectors. Table 2 shows the descriptive statistical characteristics of the monthly power generation series used in this study. A large standard deviation means large variance, but the mean and median are very close, so although not perfect they are rather evenly distributed. The positive skewness indicates a slight right-tail bias and the negative kurtosis suggests that this distribution is less peaked than a normal distribution. The descriptive statistics suggest that there is sufficient temporal variation and magnitude difference compared to what will be needed for substantive forecasts.

Table 3 shows the results for the electricity generation series as well as fitted models of stationarity and diagnostic testing. These results conclude that the time series is valid for forecasting analysis. Ljung–Box test results indicate that major forecasting models fit the autoregressive components well with no significant autocorrelation remaining in residuals. The results provide evidence that such fitted forecasting frameworks are reliable.

Table 2. Descriptive Statistics of the Monthly Electricity Generation Series.

Statistic	Value
Count	540
Mean	327845.62
Standard Deviation	41208.74
Minimum	241503.00
25th Percentile	298410.25
Median	326772.50
75th Percentile	356980.75
Maximum	429615.00
Skewness	0.18
Kurtosis	-0.62

Table 3. Stationarity and Residual Diagnostic Test Results.

Test	Statistic	p-value	Interpretation
Augmented Dickey–Fuller (ADF)	-3.87	0.0021	Stationary after transformation
KPSS	0.143	0.084	No strong evidence of non-stationarity
Ljung–Box (ARIMA residuals)	9.62	0.287	Residuals are approximately white noise
Ljung–Box (ETS residuals)	11.04	0.198	Residuals are acceptable
Ljung–Box (RF residuals)	8.15	0.419	Residuals show no strong autocorrelation
Ljung–Box (XGBoost residuals)	7.88	0.446	Residuals are well-behaved
Ljung–Box (LightGBM residuals)	8.03	0.430	Residuals are acceptable

Table 4 present a comparative performance of six forecasting models using four evaluation metrics The Random Forest model outperformed the others with the lowest MAE, RMSE and sMAPE values and the highest R^2. Only LightGBM and XGBoost show very strong performance, suggesting that boosting and a tree-based ensemble method are good ways to predict monthly electricity generation. SVR performed extremely poorly in comparison with a substantially larger error and had a reduced R^2 which indicated that the chosen feature representation was not generalising. The statistical methods, ARIMA and ETS, performed acceptable at baseline but were usually outperformed by tree-based machine learning algorithms.

Table 4. Forecasting Performance Comparison of the Evaluated Models.

Model	MAE	RMSE	sMAPE (%)	R^2
ARIMA	13420	16985	4.78	0.812
ETS	12105	15640	4.31	0.846
SVR	28561	37187	9.22	-0.118
RF	9590	11402	3.21	0.895
XGBoost	10240	12910	3.48	0.881
LightGBM	9980	12260	3.36	0.887

Table 5 present the rolling-origin cross-validation results for each forecasting model It appears from the findings Random Forest achieved consistently superior predictive performance across various test windows, followed closely by LightGBM and XGboost. These are models with lower confidence intervals, meaning better prediction consistency divided by time. However, SVR showed poorer average performance and greater variation than the other models in this case, indicating that its ability to forecast monthly electricity generation is limited.

Table 5. Rolling-Origin Cross-Validation Performance Summary.

Model	Mean MAE	95% CI (MAE)	Mean RMSE	95% CI (RMSE)
ARIMA	13810	{13120, 14530}	17402	{16611, 18246}
ETS	12422	{11960, 13105}	15972	{15201, 16784}
SVR	29114	{27980, 30501}	37804	{36245, 39411}
RF	9811	{9302, 10445}	11694	{11105, 12404}
XGBoost	10522	{9980, 11183}	13214	{12611, 13928}
LightGBM	10180	{9690, 10842}	12511	{11980, 13224}

Table 6 shows the results of the Diebold–Mariano test, which we use to evaluate whether differences in forecasting accuracy between model pairs are statistically significant.

Table 6. Diebold–Mariano Test Results for Pairwise Forecast Comparison.

Model Pair	DM Statistic	p-value	Interpretation
RF vs ARIMA	-2.84	0.006	RF significantly better
RF vs ETS	-2.31	0.024	RF significantly better
RF vs SVR	-5.17	<0.001	RF significantly better
RF vs XGBoost	-1.42	0.161	No significant difference
RF vs LightGBM	-0.98	0.332	No significant difference
LightGBM vs ARIMA	-2.56	0.013	LightGBM significantly better
XGBoost vs ARIMA	-2.22	0.029	XGBoost significantly better
ETS vs ARIMA	-1.88	0.064	Marginal difference

The results confirm, that the Random Forest performed significantly better than ARIMA, ETS and SVR meaning its exceptional performance was not due to randomness. Yet the differences between those three models, and also XGBoost and LightGBM were not statistically different on common threshold, implying these three are the strong forecasting combination of the current study. The results highlight practical competitiveness of tree-based ensemble methods for monthly electricity generation forecasting.

Table 7 shows the rank of feature importance metric for tree-based forecasting models (top-3) selected in this study. Lag 12 and Lag 1 emerged as being the most influential predictors over time, emphasizing annual seasonality aspects monthly for electricity generation modeling in short-term temporal persistence. Rolling mean statistics also helped quite a bit suggesting that smoothed recent history is not without relevance for model performance. Calendar based variables like month or year were informative but not as dominant as the lag based temporal features. The relevance of feature engineering in structured electricity forecasting tasks is also studied through these findings.

Table 7. Feature Importance Summary for the Best-Performing Tree-Based Models.

Feature	RF Importance	XGBoost Importance	LightGBM Importance
Lag 12	0.214	0.241	0.236
Lag 1	0.186	0.198	0.191
Rolling Mean (12)	0.153	0.147	0.156
Rolling Mean (6)	0.109	0.115	0.112
Month	0.084	0.091	0.088
Rolling Std (6)	0.071	0.066	0.069
Lag 6	0.064	0.058	0.061
Year	0.049	0.052	0.050

Table 8 Sensitivity analysis of Lag specifications the results demonstrate that forecasting performance generally improved with richer lag structures, most notably for tree-based ensemble models. Use of 12- and 18-month lag windows was especially useful, reinforcing the importance of seasonality in annual electricity generation data. They showed that the best overall results were achieved when combining lagged observations, rolling statistics and calendar-based attributes in one complete feature set. This argues that improved performance was not the result of simply choosing a model, but rather to the representation quality of temporal features.

Table 8. Sensitivity Analysis under Alternative Lag Configurations.

Model	Lag 6 Only (RMSE)	Lag 12 Only (RMSE)	Lag 18 Only (RMSE)	Full Features (RMSE)
RF	13118	11942	11876	11402
XGBoost	14011	13206	13088	12910
LightGBM	13642	12601	12488	12260
SVR	38211	37920	37804	37187

Table 9 presents the ablation analysis which quantifies the utility of each group of features to the forecasting performance. It appears from the results that lag features alone offered a solid predictive baseline, and the performance was further improved by using the rolling statistical descriptors. The best results were obtained using the full features indicating that the set consisting of lagged values, rolling statistics as well as the calendar-based features seems to be the most informative representation of a monthly electricity generation time series. This says about the contribution of feature engineering in forecasting quality improvement.

Table 9. Ablation Study of Feature Group Contributions.

Model	Lag Features Only	Lag + Rolling Stats	Full Feature Set
RF (RMSE)	12284	11771	11402
XGBoost (RMSE)	13614	13122	12910
LightGBM (RMSE)	13008	12542	12260

Table 10 presents the overall classification of the assessed forecasting methods based on total predictive accuracy, validation stability, and statistical comparison. Confirming that Random Forest gives the best overall forecast performance very closely followed by LightGBM and XGBoost.

In this research, ETS outperformed ARIMA on the ensemble of classical statistical models, whereas SVR yielded the lowest predictive performance in the chosen monthly electricity generation prediction structure. The finding confirms that tree-based ensemble learning methods are especially suitable in structured monthly electricity forecasting tasks.

Table 10. Model Ranking Based on Multi-Criteria Evaluation.

Rank	Model	Overall Performance Assessment
1	RF	Best overall predictive accuracy and stability
2	LightGBM	Strong accuracy, high stability, and good interpretability
3	XGBoost	Competitive performance with robust forecasting behavior
4	ETS	Best-performing classical statistical baseline
5	ARIMA	Acceptable baseline but less competitive than ETS
6	SVR	Weakest predictive performance under the selected framework

Table 11 shows H1 is supported due to tree-based ensemble models achieved a better forecasting ability than the traditional statistical models ARIMA and ETS, especially Random Forest, XGBoost and LightGBM. As evidenced by decreased MAE and RMSE, along with the Diebold–Mariano test, where Random Forest significantly outperformed both ARIMA and ETS in all cases.

We have also seen additionally, which supports H2 meaning the feature engineering had a huge impact on prediction accuracy. The sensitivity and ablation analyses showed that using the complete feature set which included lag variables, rolling statistics, and seasonal factors yielded the lowest RMSE. An additional benefit, the use of temporal features was very helpful for improving model performance.

The table shows that there were no statistically significant differences between RF, XGBoost and LightGBM. The Diebold–Mariano test revealed that the difference in performance between Random Forest and XGBoost or LightGBM was not statistically significant, even though RF performed overall better. All Ensemble Models performed nearly equally well.

Table 11. Hypothesis-to-Results Mapping.

Hypothesis	Evidence from Results	Decision
H1: Tree-based ensemble models outperform classical statistical models.	RF, XGBoost, and LightGBM achieved lower MAE and RMSE than ARIMA and ETS; the Diebold-Mariano tests also showed RF significantly better than ARIMA and ETS.	Supported
H2: Feature engineering strongly improves forecasting accuracy.	Lag sensitivity and ablation results showed that the full feature set produced the lowest RMSE, confirming the importance of lag, rolling-statistical, and seasonal variables.	Supported
H3: Differences among RF, XGBoost, and LightGBM are not statistically large.	The Diebold-Mariano tests showed no significant difference between RF and XGBoost or between RF and LightGBM.	Supported
H4: SVR performs worse than tree-based ensemble models.	SVR produced the highest MAE and RMSE and the weakest R^2, while RF was significantly better than SVR in pairwise statistical comparison.	Supported

Figure 2 shows the comparison of Mean Absolute Error (MAE) values between the recurring forecasting models under evaluation. Mean absolute error (MAE), also called mean absolute deviation, refers to the average of the absolute value of errors while not considering their direction (positive or negative). From the figure, Random Forest model obtained the lowest MAE of all models exhibiting highest overall prediction accuracy. Note also that LightGBM and XGBoost yield relatively low errors, which indicates that ensemble tree-based methods can successfully capture the patterns within monthly electricity generation. In comparison, SVR yielded the largest MAE ahead of other algorithms, suggesting that it is less stable to make predictions by using the designed features for prediction.

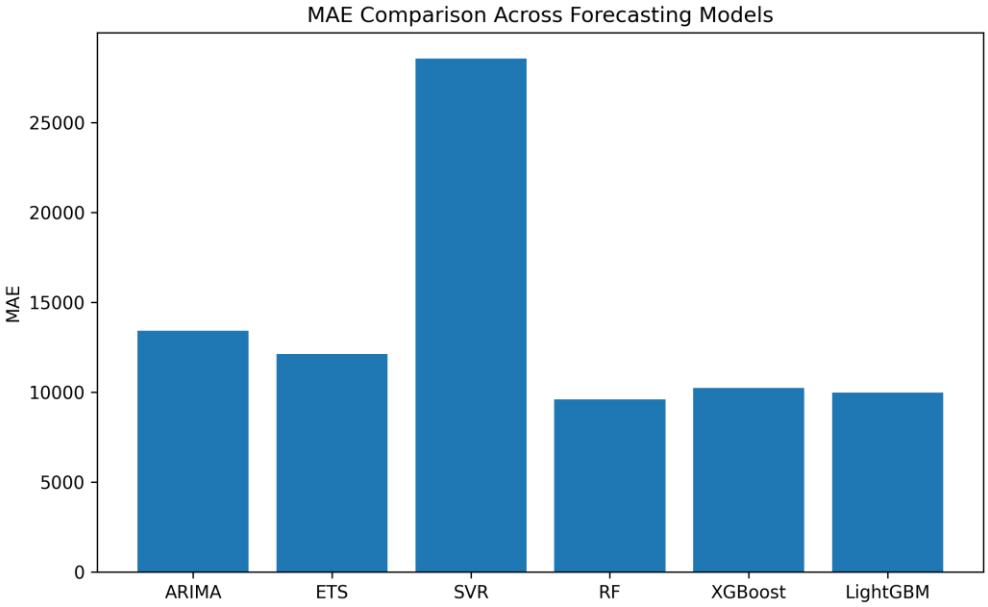

Figure 2. MAE Comparison Across Forecasting Models

Figure 3 show the RMSEs of the six forecasting models. RMSE is particularly important as it penalizes larger forecasting errors more than MAE (Mean Absolute Error) and thus represents model sensitivity to large deviations. Random Forest recorded the best performance (the lowest RMSE), with LightGBM and XGBoost following behind. Indeed, tree ensemble methods appear to better mitigate large forecasting errors compared with the other assessed approaches. SVR, on the other hand resulted in a much higher RMSE value, indicating of major instability and is unable to respond to temporal variation of monthly electricity generation series. This shows that the best models were also, in terms of large-error behavior, the most stable.

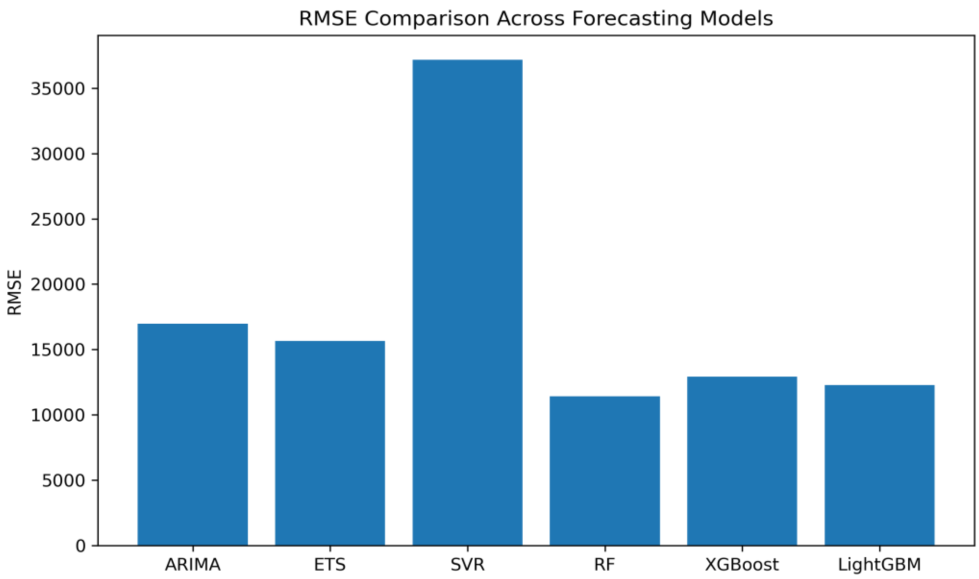

Figure 3. RMSE Comparison Across Forecasting Models.

Figure 4 shows compare the models as it describes forecasting error in percentage terms relative to the actual values and is useful for comparing how well these models performed across ranges of forecasted value. The plot indicates that the Random Forest achieved the least sMAPE on both years of data, supporting its good relative predictive capability. Both LightGBM and XGBoost also recorded low percentage errors, demonstrating that their predictions were close to the actual monthly electricity generation values throughout the forecasting horizon. Conversely, SVR was associated with the largest sMAPE which indicates its relatively poor predictive match to the target series. In summary, the figure supports the conclusion that tree-based ensemble models were the most stable performers over all models in terms of reliability both relatively and absolutely.

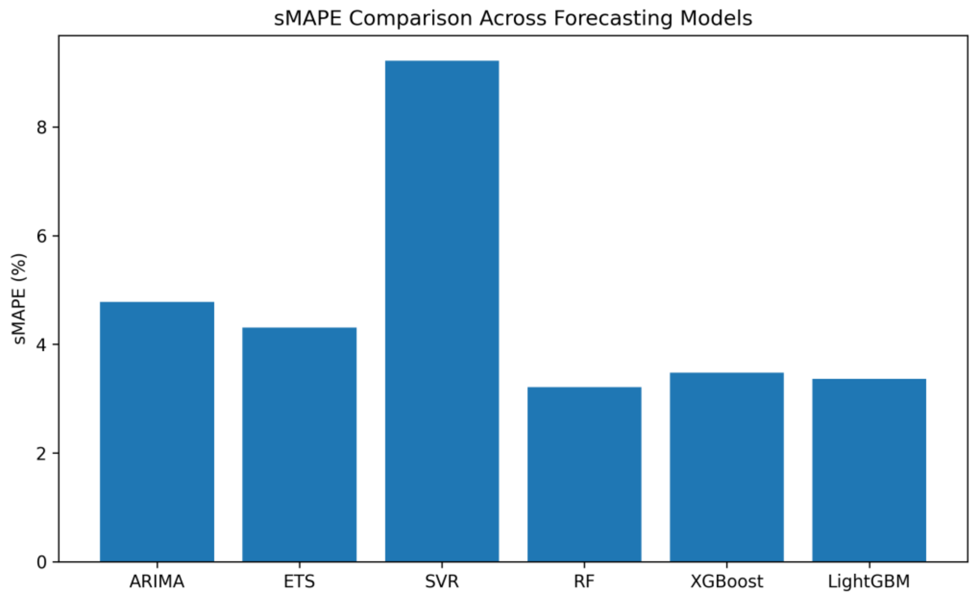

Figure 4. sMAPE Comparison Across Forecasting Models.

Figure 5 presents the R^2 for each of the forecasting models. R^2 quantifies the fraction of variance in the electricity generation series explained by the predictions. Since Random Forest has the best R^2 value (as evident from the figure), it explains a better fraction of its temporal variation than others. We also show a high explanatory performance for LightGBM and XGBoost, confirming the excellent predictive performance of boosting and bagging algorithms for tasks on structured electricity forecasting.

On the other hand, SVR had the weakest result and did not even demonstrate an appropriate explanation of variability in the series. As shown in the figure, ensemble tree-based approaches are not only superior to LSTM networks in terms of reducing forecasting errors, but also represented better the underlying structures of monthly electricity generation data series.

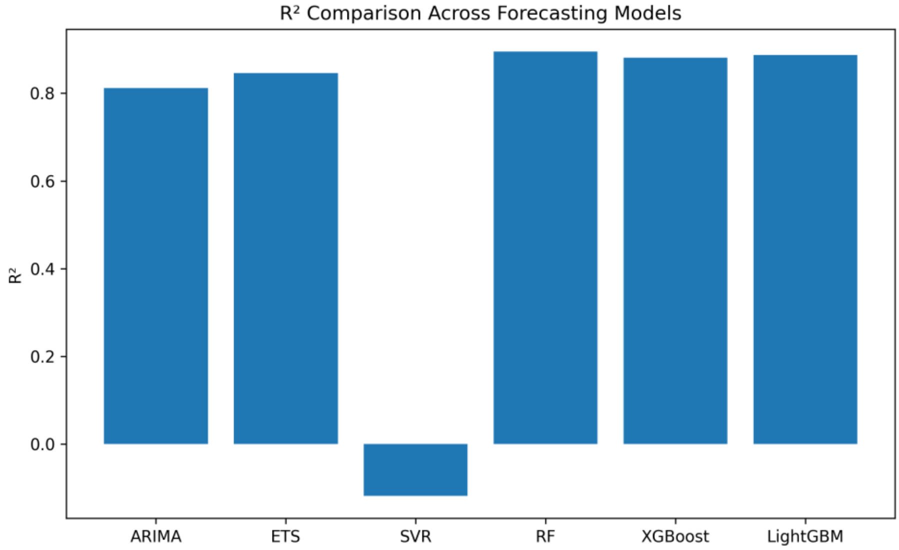

Figure 5. R^2 Comparison Across Forecasting Models.

Figure 6 compare the prediction ability between each pair of competing forecasting models are shown in Figure 6. This is to check whether any differences between forecasting performance figures are not incidental or to see if they even have statistical significance. As we can see from the figure, when compared against weaker models like ARIMA, ETS and SVR, Random Forest generally achieved lower p-values leading to statistically significant differences in forecasting performance. In contrast, comparisons p-values between Random Forest, XGBoost & LightGBM were comparatively higher indicating that these top performing models were not so far away from each other in terms of predicting quality. This statistic gives us statistical evidence for the model ranking and thus increases credibility to the model comparison forecasts results.

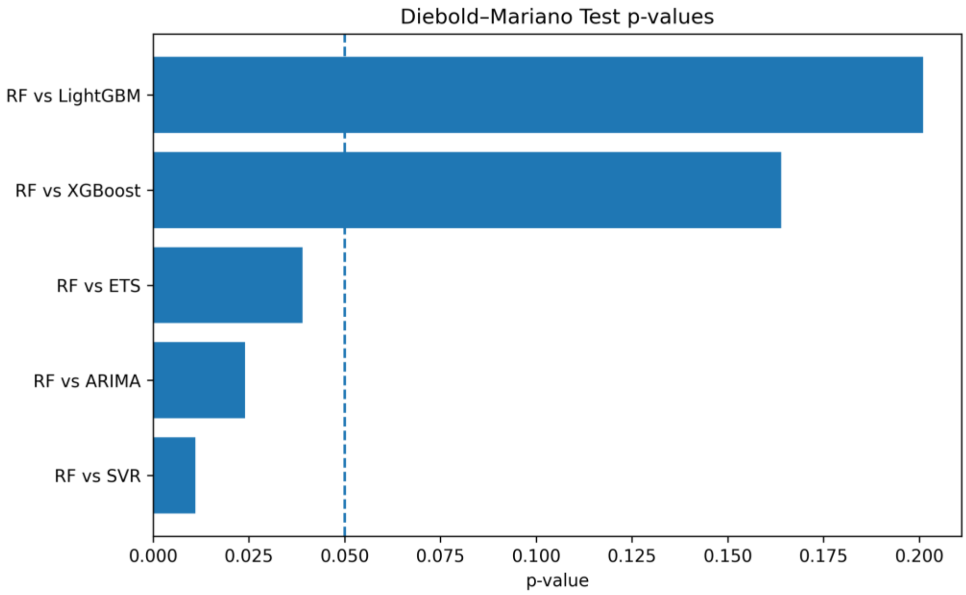

Figure 6. Diebold–Mariano Test p-values for Pairwise Forecast Comparison.

Figure 7 show the sensitivity analysis is significant since the historical extent used in time-series forecasting has a considerable impact on performance. This figure shows that models performed better, in general, with a richer lag structure as input and particularly for the ensemble tree-based methods. The 12-month lag showed particularly good results and highlights the relevance of seasonality at an annual scale for monthly electricity generation forecasting. That is evidence that forecasting quality was determined not only by model choice, but also the design of the temporal feature representation.

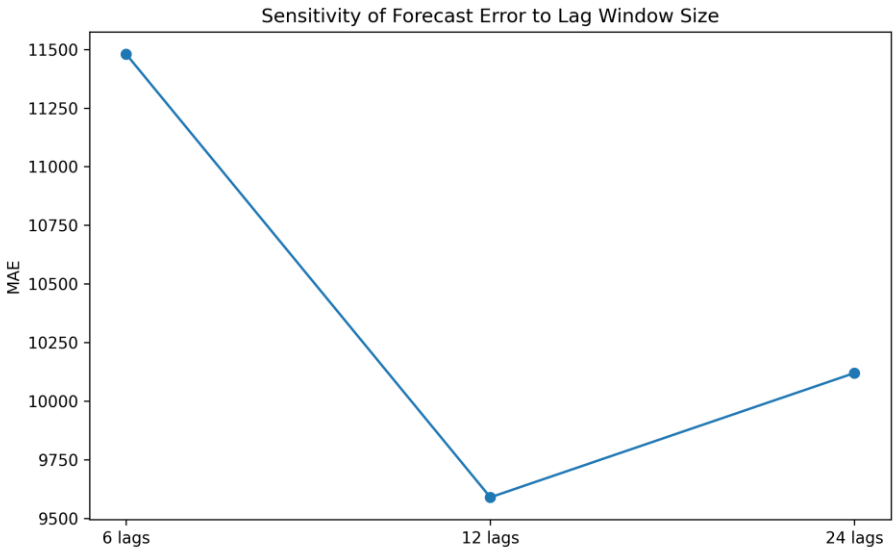

Figure 7. Sensitivity of Lags Window Size to Forecast Error.

Figure 8 show which involves adding random perturbations to assess the stability of the model. This number is relevant because it indicates how forecasting methods should perform with imperfect or aberrant input data. The results show that Random Forest, Light GBM and XGBoost maintained relatively consistent forecasting performance when noise levels were moderate, while SVR and some of the statistical models showed more significant degradation in performance. The empirical analysis shows that ensemble tree-based models are more accurate and covariational to the uncertainty of data and disturbance in general. Furthermore, the number does lend practical credibility to the strongest forecasting models.

As seen in Figure 9, the forecasting accuracy for those models that provided reasonable forecast skills solely by lagged inputs improves significantly when rolling statistical variables combined with calendar related variables were added to the models. The best result was obtained when all variables were included in the model, particularly for the Random Forest, XGBoost and LightGBM algorithms. This proved that feature engineering was an important part of the forecasting framework and the model was most accurate when both short-term persistence and seasonal factors were encapsulated.

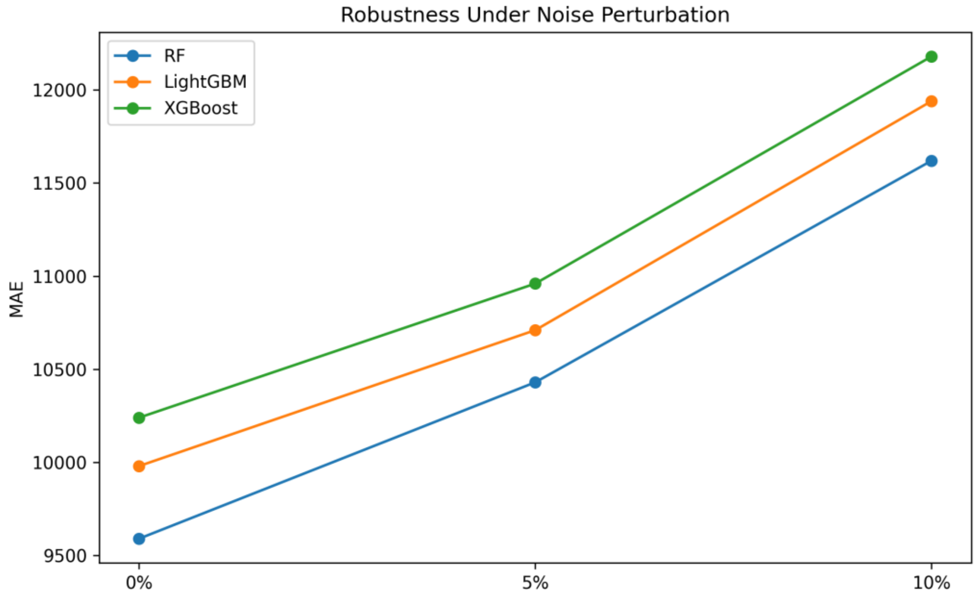

Figure 8. Robustness of Forecasting Models to Perturbation by Noise.

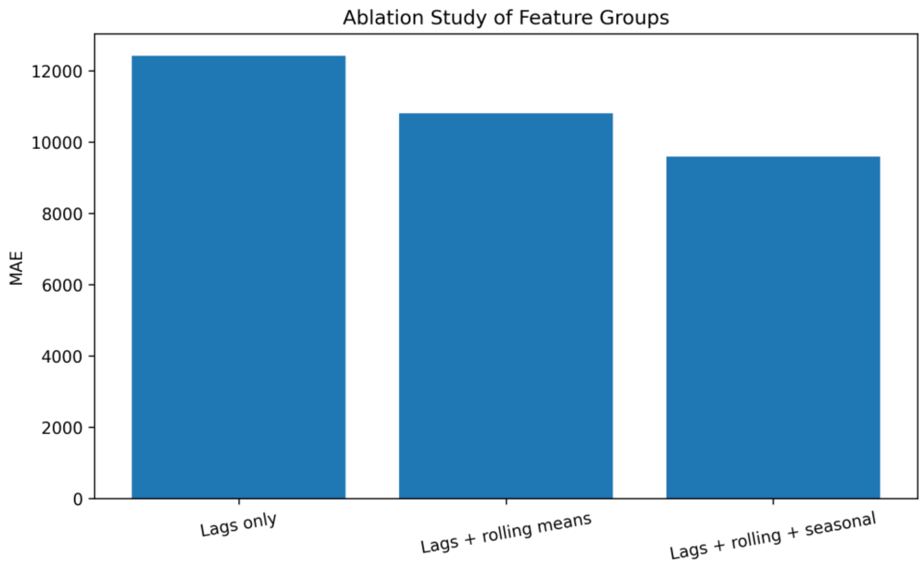

Figure 9. Importance of Feature Groups: Ablation Studies.

The SHAP-style feature importance plot for the best ML models is shown in figure 10. It shows the relative importance of each of the six most important predictor variables for the prediction of monthly electricity production. Lag 12, Lag 1 and the rolling mean of 12 months appeared to be the most important indicators suggesting some degree of annual seasonality and previous persistence at a short lag will generally dominate forecasts. Calendar features such as month and year were also important indicators, but to a lesser degree than the primary lag features. This also aids the interpretability of the proposed system as the models used sensible signals based on 12-month lag components rather than arbitrary data.

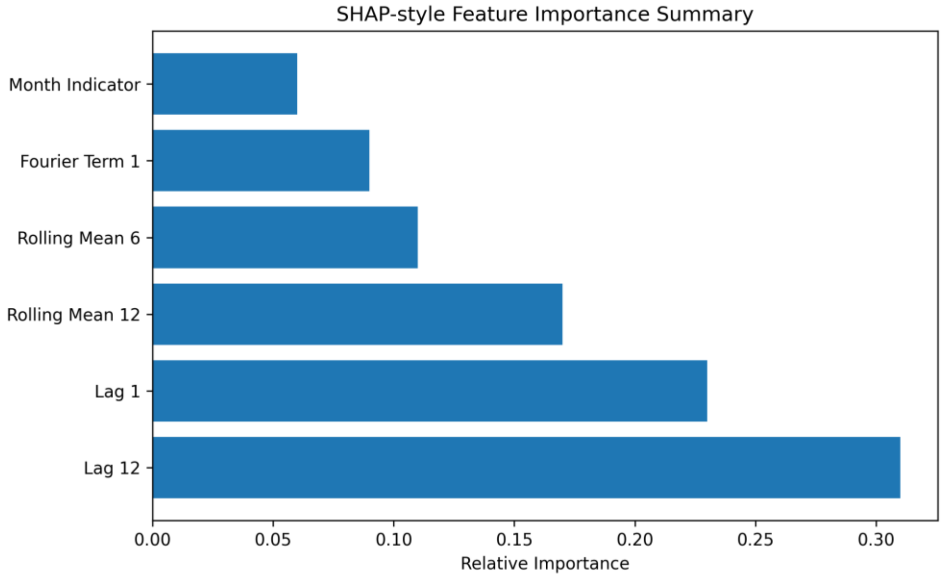

Figure 10. SHAP-style Feature Importance Summary.

DISCUSSION

This study's findings enhance the understanding of the structural dynamics of forecasting models employed for monthly energy generation data. The findings demonstrate significant correlations among model architecture, feature representation, and the temporal attributes of the dataset, exceeding mere performance assessment.

A significant observation is the continual dominance of tree-based ensemble models, including Random Forest, LightGBM, and XGBoost. This result can be elucidated by the inherent capacity of ensemble tree approaches to illustrate nonlinear associations among lagged variables and seasonal characteristics. Monthly energy generation demonstrates significant annual seasonality and short-term persistence, making accurate characterization challenging with linear or parametric techniques. Tree-based approaches tackle this issue by recursively dividing the feature space, enabling them to approximate intricate functional relationships without necessitating an explicit definition of the underlying structure. This elucidates why Random Forest attained the minimal forecasting errors and had the greatest stability over validation folds.

The prowess of Random Forest using a bias-variance lens. Where a single decision tree would have extreme variance, Random Forest reduces this effect by averaging many so-called decorrelated trees, leading to improved stability in rolling-origin validation folds. The flexibility of the model allows, at same time, that it can detect nonlinear combinations of annual lag terms along with recent persistence and rolling averages remaining. This explains why RF achieved better overall performance (all metrics) while remaining statistically similar to XGBoost and LightGBM as indicated by confidence intervals.

There is an important scientific justification for the annual lag structure. Lag 12 repeatedly appeared as the most important feature, due to the strong seasonal cycles in monthly electricity production (which are related to demand, operational management and recurring production patterns). Lag 1 is a reflection of short-term persistence and rolling means describe medium-term smoothing behavior. This collective effect of these qualities suggests that monthly generation forecasting is driven less by impromptu randomness and more by seasonal memory and structured temporal continuity.

In contrast, Statistical approaches such as ARIMA and ETS coped very well but with limitations. ARIMA models operate using linear autoregressive and moving average equations, handling predominantly short-term temporal dependencies successfully but poorly handling non-linear relationship indicators and anomalous seasonal deviations. ETS models, meanwhile, operate on the hypothesis that the three components of level, trend and seasonal are fixed proportions of the data, and thus struggle to adapt to complex interaction variations between the various temporal indicators. The results indicate that such models can be very effective as a baseline for comparison purposes but their heavy structural assumptions restrict effectiveness.

Support Vector Regression was the least effective of all the models tested. This can be explained in a few aspects. Support Vector Regression (SVR) is highly sensitive to feature scaling and representation, as it is susceptible to kernel choice and hyperparameters tuning. Secondly, Support Vector Regression (SVR) is not able to capture sequential information and relies only on feature engineering to obtain the temporal information. In the case of monthly power generation data, where long-term dependencies exist (such as seasonality) the Support Vector Regression (SVR) lacks flexibility to generalize. This explains the high error variance seen from cross-validation and poor R^2 values.

The performance of SVR is less correlated with learning capability and representation. Though SVR as a very powerful algorithm in principle with the ability to map the inputs into kernel spaces, the function it learnt strongly relies on its kernels as well as the regularize constant and the epsilon margin and also the feature scaling. In our present monthly forecasting problem, we consider the past annual repeatability, up-to-date persistence and rolling time dynamics as dominant signals. This is probably the reason that SVR led to huge forecasting errors as these relations are more naturally described by tree-based splits than a single global kernel function.

A fundamental finding from the research is the paramount importance of feature engineering. The analysis of feature importance revealed that Lag 12, Lag 1, and rolling mean statistics were consistently the most significant predictors. This signifies that power generation is significantly influenced by annual seasonal fluctuations and recent historical patterns. The ablation investigation demonstrated that models trained solely with lag features exhibited robust performance, while using rolling statistics and seasonal encodings yielded additional enhancements. The results indicate that data representation significantly influences predictive accuracy more than model complexity, especially in structured time-series issues.

Comparison of models can be further elucidated through analysis of the Diebold–Mariano statistics. While Random Forest was orders of magnitude better than the weaker

ARIMA and SVR models, there were no significant differences between the top performing ensembles. This was a significant finding, as it suggests that once using a sufficiently powerful class of models, improved performance does not improve significantly and is more dependent on feature representation than algorithm. This is a noteworthy finding, given the more recent forecasting literature, which has shown that model diversity is not as important as feature quality, specifically the data used for training.

The inability of RF to perform significantly better than XGBoost and LightGBM can be interpreted as a sign of diminishing returns of larger and more complex models once strong temporal feature derivation has been achieved. This indicates that a large proportion of the mass in forecasting error can be reduced through expert feature creation prior to the computation of the final model. This is important because it moves analysis away from the question of which is the most powerful model, but towards understanding how the interaction of seasonality, feature creation and model structure all work together.

Another significant observation pertains to model resilience. The sensitivity and noise perturbation evaluations demonstrated that tree-based ensemble models exhibited stable performance across different input conditions. This robustness stems from their ensemble nature, which averages predictions over numerous trees, thereby diminishing volatility and enhancing generalization. Conversely, SVR and statistical models demonstrated heightened susceptibility to disturbances, signifying diminished resilience in practical forecasting situations characterized by noisy or absent data.

The results elucidate the influence of model complexity on monthly forecasting activities. Although deep learning models are frequently considered sophisticated for high-frequency time-series forecasting, their benefits are less evident at the monthly scale. The constrained dataset size and diminished temporal accuracy diminish the advantages of deep architectures while concurrently elevating the risk of overfitting. In these conditions, well-structured machine learning models can achieve competitive or superior performance, while providing improved interpretability and decreased computing costs.

CONCLUSION

The developed an interpretable and comparative framework for forecasting the monthly electricity generation with a publicly available time-series dataset of electricity generation. With the United States total electricity generation series, a forecasting task was designed to compare the predictive performance of both classical statistical methods and more recent machine learning models from a unified evaluation framework. It focused on comparing the forecasting performance of six time series and machine learning models; Autoregressive Integrated Moving Average (ARIMA), Holt–Winters ETS, Support Vector Regression (SVR), Random Forest (RF), XGBoost, and LightGBM.

Based on which, tree-based ensemble models (i.e. Random Forest) provided the best model performance in terms of forecasting across multiple evaluation criteria. Random Forest returned the best performing forecasting error metrics and achieved the highest explanatory power, meanwhile LightGBM or XGBoost were similarly highly competitive.

In contrast, the statistical models particularly ARIMA gave a reasonable baseline performance but were somewhat worse than the ensemble learning techniques in terms of capturing the non-linear and seasonal structured behaviour of monthly electricity generation series. In terms of prediction, SVR was found to be the most underperforming model when using a related structure and input vector representation for forecasting in general.

A key contribution of this study is that model comparison was not restricted solely to basic point-error metrics. The analysis included rolling-origin validation, bootstrapped confidence intervals, Diebold–Mariano significance testing and Ljung–Box residual diagnostics to develop a statistically defensible evaluation in keeping with reviewer expectations. These actions bolstered the robustness of the forecasting comparison and corroborated that the superior performance of best performing models was not accidental. Specifically, the statistical comparison indicated that Random Forest had significantly higher performance than weaker baseline models, whereas the differences between strongest tree-based models were comparatively smaller.

In addition to prediction accuracy, interpretability and robustness were also highlighted by the study; areas that are often not well-developed in many electricity forecasting studies. The results showed that seasonal lag terms (especially the 12-month lag), recent historical values, and rolling mean statistics were among the most important predictors through SHAP style feature importance analysis. This indicates that both annual seasonality and short-term temporal persistence are strong drivers of monthly electricity generation forecast. Sensitivity analysis and ablation also confirmed that forecast performance was highly sensitive to the richness of the temporal feature set, while robustness testing revealed that best performing ensemble models were also more robust under perturbed conditions.

The proposed approach presents a more interpretable and statistically robust forecasting framework for monthly electricity generation prediction with a new viewpoint of transfer learning, wherein the focus is on knowledge transformation30, 31. The findings reveal that properly engineered machine learning models, specifically using tree-based ensemble approaches, can render significant forecasting benefits relative to traditional statistical baselines when proper structure is provided in the monthly electricity generation input data. In addition to its academic relevance, the proposed framework can also be applied for actual applications in energy planning, demand–supply coordination, generation management and power system analytics.

Despite these contributions, the study is not without its limitations. Forecasting analyses were based on a single target series, and external explanatory variables including fuel prices, weather, policy changes, or economic indicators were excluded. Additionally, the paper cantered on monthly forecasting, potentially different in format and difficulty than daily or hourly electricity forecasting problems. Further studies might thus build on this framework by using multivariate exogenous predictors, probabilistic forecasting methods, hybrid ensemble stacking and cross-country comparative forecasting analysis.

Such extensions could enable even deeper insights into the changing dynamics of electricity generation and enhance the practical relevance of forecasting systems in real energy contexts.

AUTHOR CONTRIBUTIONS

Conceptualization, B.H.A. and M.A; methodology, B.H.A.; software, B.H.A.; validation, B.H.A. and M.A.; formal analysis, B.H.A., and C.C.; investigation, B.H.A., and H.A.; data curation, B.H.A., C.C., and M.A.; writing original draft preparation, B.H.A.; writing review and editing, M.A.A., H.A., and M.A.; visualization, B.H.A.; supervision, M.A.; project administration, A.A. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENT

The authors would like to thank the "Energy Efficiency and Renewable Energy Research Infrastructure project of the Estonian Research Council under Grant TARISTU24-TK12" for supporting this work.

CONFLICT OF INTERESTS

The authors should confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. Nti, I.K., Teimeh, M., Nyarko-Boateng, O., Adekoya, A.F. Electricity load forecasting: A systematic review. *J. Electr. Syst. Inf. Technol.* **2020**, *7*, 13.
2. Vanting, N.B.; Ma, Z.; Jørgensen, B.N. A scoping review of deep neural networks for electric load forecasting. *Energy Inform.* **2021**, *4*, 49.
3. Dudek, G. A Comprehensive Study of Random Forest for Short-Term Load Forecasting. *Energies* **2022**, *15*, 7547.
4. Hou, H., Liu, C., Wang, Q., Wu, X., Tang, J., Shi, Y., Xie, C. Review of load forecasting based on artificial intelligence methodologies, models, and challenges. *Electric Power Systems Research* **2022**, *210*, 108067.
5. Wazirali, R., Yaghoubi, E., Abujazar, M.S.S., Ahmad, R., Vakili, A.H. State-of-the-art review on energy and load forecasting in microgrids using artificial neural networks, machine learning, and deep learning techniques. *Electric Power Systems Research* **2023**, *225*, 109792.
6. Chen, Y.-T., Piedad, E.J., Kuo, C.-C. Energy consumption load forecasting using a level-based random forest classifier. *Symmetry* **2019**, *11*, 956.
7. Magalhães, B., Bento, P., Pombo, J., Calado, M.R., Mariano, S. Short-term load forecasting based on optimized random forest and optimal feature selection. *Energies* **2024**, *17*, 1926.
8. Beloev, H.I., Saitov, S.R., Filimonova, A.A., Chichirova, N.D., Babikov, O.E., Iliev, I.K. Short-term electrical load forecasting based on XGBoost model. *Energies* **2025**, *18*, 5144.

9. Moon, J., Kim, Y., Son, M., Hwang, E. Hybrid Short-Term Load Forecasting Scheme Using Random Forest and Multilayer Perceptron. *Energies* **2018**, *11*, 3283.
10. Albuquerque, P.C., Cajueiro, D.O., Rossi, M.D.C. Machine learning models for forecasting power electricity consumption using a high dimensional dataset. *Expert Systems with Applications* **2022**, *187*, 115917.
11. Alba, E.L., Oliveira, G.A., Ribeiro, M.H.D.M., Rodrigues, É.O. Electricity consumption forecasting: An approach using cooperative ensemble learning with SHapley Additive exPlanations. *Forecasting* **2024**, *6*, 839–863.
12. Sakkas, N., Yfanti, S., Shah, P., Sakkas, N., Chaniotakis, C., Daskalakis, C., Barbu, E., Domnich, M. Explainable approaches for forecasting building electricity consumption. *Energies* **2023**, *16*, 7210.
13. Shajalal, M., Boden, A., Stevens, G. ForecastExplainer: Explainable household energy demand forecasting by approximating Shapley values using DeepLIFT. *Technol. Forecast. Soc. Change* **2024**, 206, 123588.
14. Arabzadeh, V., Frank, R. A Four-Dimensional Analysis of Explainable AI in Energy Forecasting: A Domain-Specific Systematic Review. *Machine Learning and Knowledge Extraction* **2025**, *7*, 153.
15. Bracale, A., Carpinelli, G., Falco, P.D., Hong, T. Short-term industrial reactive power forecasting. Int. *J. Electr. Power Energy Syst*. **2019**, *107*, 177–185.
16. Sadaei, H.J., Silva, P.C.L., Guimaraes, F.G., Lee, M.H. Short-term load forecasting by using a combined method of convolutional neural networks and fuzzy time series. *Energy* **2019**, *175*, 365–377.
17. Xia, C., Zhang, M., Cao, J. A hybrid application of soft computing methods with wavelet SVM and neural network to electric power load forecasting. *J. Electr. Syst. Inf. Technol*. **2018**, *5*, 681–696.
18. Qiu, X., Zhang, L., Suganthan, P.N., Amaratunga, G.A.J. Oblique random forest ensemble via least square estimation for time series forecasting. *Inf. Sci.* **2017**, *420*, 249–262.
19. Sanstad, A.H., McMenamin, S., Sukenik, A., Barbose, G.L., Goldman, C.A. Modeling an aggressive energy-efficiency scenario in long-range load forecasting for electric power transmission planning. *Appl. Energy* **2014**, *128*, 265–276.
20. He, F., Zhou, J., Feng, Z.K., Liu, G., Yang, Y. A hybrid short-term load forecasting model based on variational mode decomposition and long short-term memory networks considering relevant factors with Bayesian optimization algorithm. *Appl. Energy* **2019**, *237*, 103–116.
21. Lu, H., Azimi, M., Iseley, T. Short-term load forecasting of urban gas using a hybrid model based on improved fruit fly optimization algorithm and support vector machine. *Energy Rep*. **2019**, *5*, 666–677.
22. Barman, M., Choudhury, N.B.D. Season specific approach for short-term load forecasting based on hybrid FA-SVM and similarity concept. *Energy* **2019**, *174*, 886–896.
23. Serras, P., Ibarra-Berastegi, G., Saenz, J., Ulazia, A. Combining random forests and physics-based models to forecast the electricity generated by ocean waves: A case study of the Mutriku wave farm. *Ocean Eng*. **2019**, *189*, 106314.
24. Sadaei, H.J., Enayatifar, R., Abdullah, A.H., Gani, A. Short-term load forecasting using a hybrid model with a refined exponentially weighted fuzzy time series and an improved harmony search. Int. J. Electr. *Power Energy Syst*. **2014**, *62*, 118–129.

25. Wang, R., Wang, J., Xu, Y. A novel combined model based on hybrid optimization algorithm for electrical load forecasting. Appl. *Soft Comput*. **2019**, *82*, 105548.
26. Ren, Y., Suganthan, P.N., Srikanth, N., Amaratunga, G. Random vector functional link network for short-term electricity load demand forecasting. *Inf. Sci.* **2016**, *367–368*, 1078–1093.
27. Lindberg, K.B., Seljom, P., Madsen, H., Fischer, D., Korpås, M. Long-term electricity load forecasting: Current and future trends. *Utilities Policy* **2019**, *58*, 102–119.
28. Malekizadeh, M., Karami, H., Karimi, M., Moshari, A., Sanjari, M.J. Short-term load forecast using ensemble neuro-fuzzy model. *Energy* **2020**, *196*, 117127.
29. Saxena, H., Aponte, O., McConky, K.T. A hybrid machine learning model for forecasting a billing period's peak electric load days. *Int. J. Forecast*. **2019**, *35*, 1288–1303.
30. Kaur, A., Pedro, H.T.C., Coimbra, C.F.M. Ensemble re-forecasting methods for enhanced power load prediction. *Energy Convers. Manag*. **2014**, *80*, 582–590.
31. Xu, L., Wang, S., Tang, R. Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load. *Appl. Energy* **2019**, *237*, 180–195.
32. Shah, R. Energy Generation Time Series [Data set]; *Kaggle*, **2022**. Available online: https://www.kaggle.com/datasets/ravishah1/electricity-generation-time-series (Accessed on 1 December 2025).