Article

Not peer-reviewed version

Combining Statistical and Machine Learning Methodologies in Energy Consumption Forecasting for Electric Vehicles

Vasileios Pitsiavas , Georgios Spanos * , Sofia Polymeni , Antonios Lalas , Konstantinos Votis , Dimitrios Tzovaras

Posted Date: 6 March 2025

doi: 10.20944/preprints202503.0161.v1

Keywords: Machine Learning; Statistical Analysis; Energy Consumption Forecasting; XGBoost; Random Forest; Regression

 doi:10.20944/preprints202503.0161.v1

Article

Combining Statistical and Machine Learning Methodologies in Energy Consumption Forecasting for Electric Vehicles

Vasileios Pitsiavas, Georgios Spanos *, Sofia Polymeni, Antonios Lalas, Konstantinos Votis and Dimitrios Tzovaras

Information Technologies Institute, Centre for Research and Technology—Hellas, Georgios Spanos
* Correspondence: gspanos@iti.gr

Abstract: Achieving the Sustainable Development Goals (SDG) requires a transition from conventional fossil-fuel-powered vehicles to alternative energy sources, such as electricity. However, accurately forecasting energy consumption remains a critical challenge in the widespread adoption of Electric Vehicles (EVs), as it directly impacts operational efficiency, route planning, and charging strategies. To address this, a novel approach is proposed, combining advanced machine learning models—such as XGBoost, Random Forest, and regression-based techniques—with innovative dataset manipulation using statistical methods. The methodology integrates feature engineering to incorporate vehicle-specific metrics, including driving patterns and environmental conditions, ensuring models dynamically adapt to real-world scenarios. The proposed framework demonstrates high accuracy and robustness in predicting energy consumption, providing valuable insights for sustainable transportation and efficient energy management toward SDG achievement.

Keywords: Machine Learning; Statistical Analysis; Energy Consumption Forecasting; XGBoost; Random Forest; Regression

1. Introduction

The global shift from traditional fossil fuels to renewable energy sources marks a pivotal step in order to achieve Sustainable Development Goals (SDG). In particular, toward sustainable transportation, the key component is the transition to the electric vehicles (EVs) as they offer a cleaner alternative to internal combustion engine vehicles [1]. However, despite their growing adoption, several challenges prevent their widespread use, such as driving range, charging time, battery cost, and battery weight [2].

Energy consumption plays a crucial role in the operation of electric vehicles, as it affects their efficiency, the planning of their routes, and the scheduling of refueling [3]. Energy consumption in EVs is highly dynamic and its influenced by factors such as driving patterns, traffic conditions, weather, and vehicle-specific metrics. Hence, one of the most prominent challenges pertains to the inaccuracy in the prediction of energy consumption under varied driving conditions [4]. Accurate energy consumption forecasts can help drivers better manage their vehicle's performance, contributing to more efficient driving and energy usage [5] and in general, to optimize of the overall use of electric vehicles. Hence, poor predictions of energy consumption result in inefficient charging infrastructure, higher range anxiety for the driver, and inefficient energy management strategies.

Several machine learning models have been employed to improve prediction accuracy, however, existing approaches often fail to capture the temporal dependencies and variability inherent in real-world scenarios [6]. For instance, many models rely on static datasets that do not account for the fluctuating energy demands caused by stop-and-go traffic, elevation changes, or extreme weather

 doi:10.20944/preprints202503.0161.v1

conditions. This limitation not only reduces the reliability of predictions but also undermines user confidence in EVs, slowing their adoption.

Energy consumption forecasting, particularly for long-term predictions, is essential for effective infrastructure planning [7]. Accurate forecasting is crucial for optimizing the placement of charging stations and managing energy distribution efficiently. As the EV market continues to expand, existing infrastructure will face increasing strain, necessitating the development of reliable energy consumption models [8]. Addressing these challenges is fundamental to ensuring the long-term viability of electric vehicles and facilitating a seamless transition toward sustainable transportation.

In this article, a novel framework for EV energy consumption forecasting is proposed leveraging advanced machine learning models and innovative data preprocessing techniques containing statistical analysis. More specifically, the suggested approach using a publicly available dataset addresses the energy consumption forecasting problem by i)incorporating feature engineering to capture vehicle-specific metrics and environmental conditions, ensuring adaptability to diverse real-world scenarios, ii)leveraging statistical analysis of the extracted features, and iii)employing machine learning algorithms such as XGBoost, Random Forest, and regression-based models. The purpose of the current research is to deliver high-accuracy predictions that support sustainable transportation and efficient energy management.

The remainder of this paper is structured as follows. Section 2 provides a review of related work, highlighting current advancements and limitations in EV energy consumption forecasting. Section 3 details the methodology, including dataset preprocessing, feature engineering, and model implementation. Section 4 presents the experimental results and discusses their implications. Finally, Section 5 concludes the paper and outlines potential future directions.

2. Related Work

In recent years, the rise of EVs has driven extensive research into optimizing their energy consumption. This has led to the development of machine learning models for predicting energy usage, along with applications designed to help users manage their vehicles more efficiently. In the following paragraphs representative research studies focusing on energy consumption forecasting are presented.

Huang et al. [9] introduced a novel approach to understanding the factors influencing EV energy consumption. The study addresses a key limitation of previous research studies, which primarily examined statistical correlations without exploring causal relationships. By applying double/debiased ML combined with bootstrap-of-little-bags inference, the authors inferred how various factors affect energy consumption.

Petkevicius et al. [10] tackled the growing challenge of EV energy consumption prediction, particularly in long-distance route planning. The authors proposed a two-tier architecture for forecasting both energy consumption and travel time. The first tier consists of a routing and travel-time prediction subsystem that generates a suggested route and predicts speed variations. The second tier predicts energy consumption based on speed profiles, weather conditions, and road slope.

Hua et al. [11] proposed a method for accurately estimating EV energy consumption, emphasizing its importance for optimizing charging station deployment, promoting eco-friendly driving practices, and extending EV range. Their approach comprises three key components: segmentation-assisted trajectory granularity, knowledge transfer from internal combustion engine/hybrid electric vehicles to EVs, and time-series estimation using a bidirectional recurrent neural network.

Athanasakis et al. [12] introduced a novel methodology for forecasting the State of Charge (SoC) in EVs. Their approach integrates lightweight regression-based models with feature engineering to enhance SoC prediction accuracy. To validate the effectiveness of their methodology, experiments were conducted on both a private dataset of automated EVs and a public EV dataset. The results demonstrated the efficiency of their lightweight approach and its superiority over more computationally intensive methods, such as deep learning (DL) models and gradient boosting algorithms.

 doi:10.20944/preprints202503.0161.v1

Chen et al. [13] addressed the challenges of energy consumption estimation in EVs, particularly in scenarios where certain features are unavailable. To account for uncertainties introduced by transient factors, the authors implemented an XGBoost prediction model, which outperformed other approaches in point-value energy consumption estimations.

Tang et al. [14] presented a predictive energy management approach that optimizes energy consumption for plug-in hybrid electric vehicles using travel route data. Their methodology incorporates battery temperature as an optimization factor in the energy cost function and employs an extreme learning machine as a short-term speed predictor. By training the speed predictor on historical real-world speed data and integrating travel route information, the study achieved higher forecasting accuracy than traditional driving cycles.

Polymeni et al. [15] developed a machine learning-based approach for energy consumption forecasting in fuel cell electric vehicles. Their methodology incorporated techniques from three distinct machine learning families—statistical-based models, ensemble methods, and deep learning. Using an artificial dataset of hydrogen vehicles generated through the Future Automotive Systems Technology Simulator [16] and considering real-world booking records from an on-demand public transportation service in the Geneva Canton region, they achieved high forecasting accuracy, with errors remaining below 10%.

This work distinguishes itself from related studies that focus solely on forecasting energy consumption or State of Charge (SoC) by introducing a novel methodology that leverages machine learning models—including Random Forest, XGBoost, and regression-based techniques—applied across multiple time intervals (1-minute, 5-minute, and 10-minute) for energy consumption prediction. By capturing temporal dependencies at different time resolutions, the proposed approach enhances predictive model robustness. Additionally, the methodology employs diverse strategies for selecting input variables, incorporating historical energy consumption, external vehicle-related features, or a combination of both, representing a novel contribution to the field. Unlike previous studies, this approach systematically evaluates the impact of exogenous factors such as elevation, vehicle speed, acceleration, and air conditioning power on energy consumption, ensuring a more comprehensive analysis of EV energy dynamics. Furthermore, the statistical analysis of these contributing factors, as presented in this study, has not been previously explored in related research.

3. Dataset Description

Zhang et al. [17] introduced the extended Vehicle Energy Dataset (eVED), an enhancement of the well-established Vehicle Energy Dataset [18], designed to provide more accurate vehicle trip data for large-scale energy consumption analysis. The eVED dataset ensures data sufficiency for machine learning applications by incorporating precise GPS coordinates and enriched features. It includes a comprehensive set of variables that capture both internal vehicle dynamics and external environmental factors, essential for understanding vehicle behavior and traffic patterns. The dataset consists of multiple electric vehicles with diverse operational characteristics, and for this study, three EVs (vehicle 10, vehicle 455, and vehicle 541) were utilized. Structurally, eVED is formatted as a time series, representing vehicle routes and their corresponding data over time.

Compared to other datasets commonly used in energy consumption research, such as simulated datasets [15] or aggregated trip-level records [19,20], eVED offers a higher temporal resolution by capturing real-world driving conditions at minute-level intervals. This fine granularity enables detailed temporal analysis and improves machine learning models' ability to capture short-term fluctuations in energy usage. Key features influencing energy consumption predictions include vehicle speed, air-conditioning power, heater power, and smoothed elevation data. These variables provide insights into both vehicle performance and the external factors affecting energy consumption. As a result, the eVED dataset serves as a valuable and diverse resource for robust energy consumption modeling across various driving conditions and vehicle types.

 doi:10.20944/preprints202503.0161.v1

3.1. Preprocessing and Data Preparation

To effectively utilize the dataset and ensure its compatibility with the selected predictive algorithms, several key preprocessing steps were applied, constituting one of the main novelties of this study. As previously mentioned, the dataset includes three EVs: vehicle_455, vehicle_10, and vehicle_541. Initially, a separation based on vehicle ID was performed to distinguish between the three electric vehicles. Additionally, the original timestamps, recorded in milliseconds, were aggregated into one-minute, five-minute, and ten-minute intervals, generating three distinct datasets for each vehicle to facilitate the calculation of energy consumption and acceleration.

Following the initial data filtering, a statistical analysis of fundamental metrics was conducted. This analysis, inspired by relevant literature [21,22], aimed to better capture variable distributions. By computing and recording various statistical measures—including mean, median, minimum, maximum, and standard deviation for each feature—the dataset's structure and key characteristics were more thoroughly examined.

Standardizing the raw timestamps into consistent minute-based intervals ensured uniform time series data across all vehicles. Within these intervals, feature-specific metrics were recalculated, providing a more detailed and accurate temporal representation of the dataset. The described preprocessing approach is essential for maintaining consistency across the time series while also improving the accuracy and reliability of the energy consumption predictions. Table 1 shows the initial variables of the dataset and Table 2 depicts the dataset after the preprocessing procedure.

Table 1. eVED Dataset Raw Data Variables.

Variable	Type	Description
DayNum	Numeric	Numerical representation of the day
VehId	Numeric	Unique vehicle identifier
Trip	Numeric	Identifier for each trip
Timestamp (ms)	Numeric	Time recorded in milliseconds
Latitude, Longitude	Numeric	Geographical coordinates
Vehicle Speed (km/h)	Numeric	Speed of the vehicle
OAT (°C)	Numeric	Outside ambient temperature
AC Power (W), Heater Power (W)	Numeric	Power usage of AC and heater
HV Battery Current (A)	Numeric	High-voltage battery current
HV Battery SOC (%)	Numeric	State of charge of the battery
HV Battery Voltage (V)	Numeric	Voltage of the battery
Elevation Raw/Smoothed (m)	Numeric	Raw and processed elevation values
Gradient	Numeric	Road slope gradient
Energy Consumption (kWh)	Numeric	Energy used per trip
Matched Latitude Longitude	Numeric	Matched GPS coordinates
Match Type	Categorical	Type of GPS match
Speed Limit (km/h)	Numeric	Road speed limit
Speed Limit (Direction) (km/h)	Numeric	Speed limit considering direction

 doi:10.20944/preprints202503.0161.v1

Table 2. eVED Dataset Processed Data Variables.

Variable	Type	Description
1min Interval	Numeric	Time interval in minutes
Energy Consumption (kWh)	Numeric	Total energy consumed in the interval
Acceleration (Mean, Max, Min, Median, Std)	Numeric	Statistical measures of vehicle acceleration
Speed (Mean, Max, Min, Median, Std)	Numeric	Statistical measures of vehicle speed
OAT (Mean, Max, Min, Median, Std)	Numeric	Outside ambient temperature statistics
AC Power (Mean, Max, Min, Median, Std)	Numeric	Air conditioning power usage statistics
Heater Power (Mean, Max, Min, Median, Std)	Numeric	Heater power usage statistics
Elevation (Mean, Max, Min, Median, Std)	Numeric	Statistical measures of elevation

4. Methodology

In this paper, an energy forecasting framework for EVs is presented, aiming to provide accurate predictions and insights about energy usage. Considering their particular high performance to different prediction problems [5,15,23–26], Random Forest [27], XGBoost [28] and regression-based algorithms [29] were selected in the suggested methodology. The selection of ML algorithms from different families is in accordance with the literature [15] for generalization purposes of the proposed methodology.

4.1. Machine Learning Models

As mentioned above, XGBoost, a Regression-based model, and a Random Forest model were implemented. Wang et al. [30] highlighted the Random Forest algorithm as a promising tool for energy prediction, demonstrating stability, strong fitting capabilities, and the feasibility of homogeneous ensemble learning in energy forecasting. The Random Forest model, an ensemble learning method, was selected for its ability to reduce overfitting while maintaining accuracy through multiple decision trees. It was trained using preprocessed features such as vehicle speed, air-conditioning power, heater power, acceleration, and smoothed elevation, aggregated into one-minute, five-minute, and ten-minute intervals. This model effectively captured nonlinear relationships between these features and energy consumption, ensuring reliable predictions even with a high-dimensional feature space.

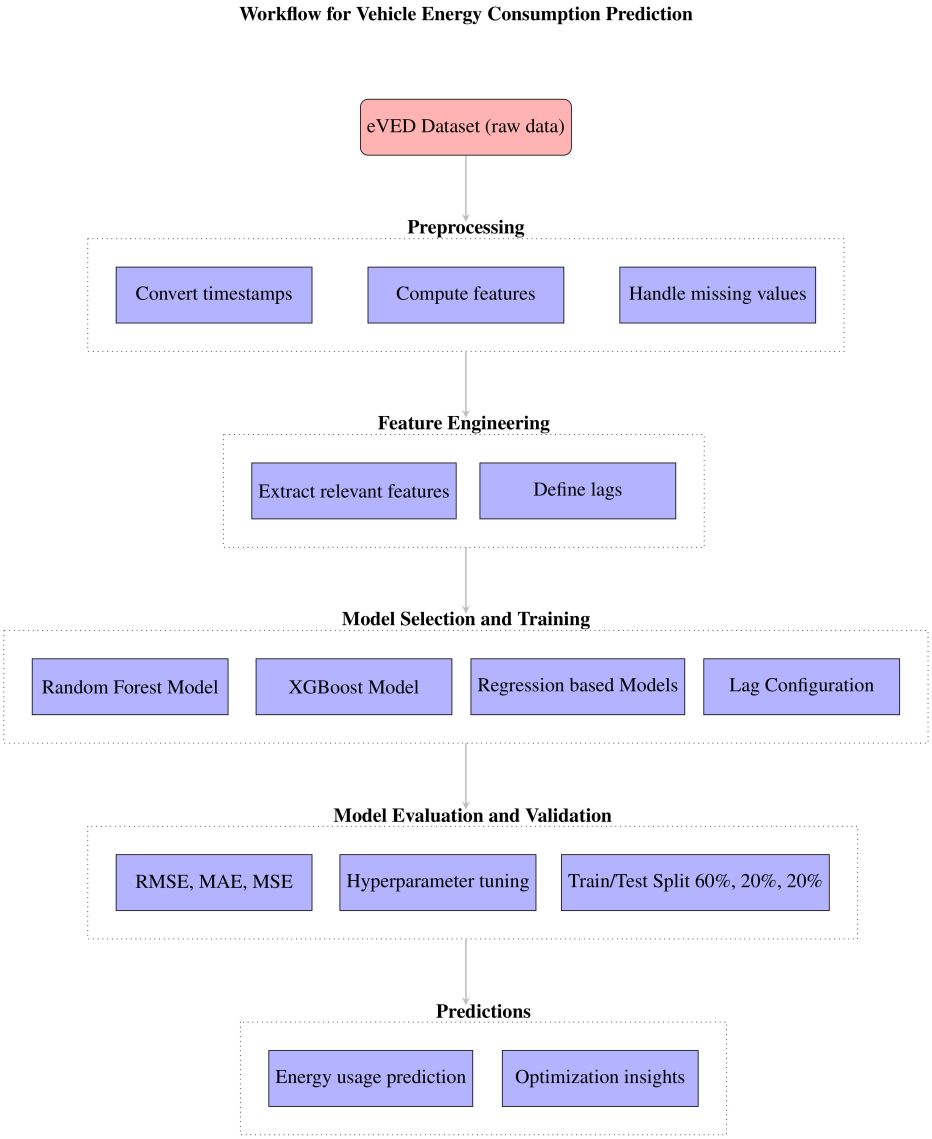
XGBoost, a gradient boosting algorithm, was also employed in this study due to its efficiency in handling large datasets and its iterative learning process, which enhances predictive performance. XGBoost includes built-in regularization to prevent overfitting, making it particularly suitable for datasets with diverse driving patterns. The model was trained using the same feature set and time intervals and underwent hyperparameter tuning, optimizing parameters such as learning rate, maximum tree depth, and the number of boosting rounds. Ma et al. [31] demonstrated the superiority of XGBoost for wind power forecasting, showcasing its effectiveness in handling complex time-series data. Additionally, XGBoost efficiently optimizes hyperparameters using techniques like the Tree-Structured Parzen Estimator (TPE) [32], ensuring robust and precise results. With these advantages, XGBoost is well-suited for forecasting the energy consumption of electric vehicles in real-world applications.

A third model used in this study is Dynamic Linear Regression, a well-established technique known for its simplicity and interpretability. This regression model, widely employed in predictive analysis for time-series data [4], integrates lagged features of both the target variable (energy consumption) and external factors such as acceleration, vehicle speed, and air-conditioning usage. By incorporating historical patterns through a lagged structure, the model captures temporal dependen-

 doi:10.20944/preprints202503.0161.v1

cies and the influence of exogenous factors on energy usage. The model's simplicity allows for easy hyperparameter tuning and residual analysis, enabling the detection of temporal patterns or anomalies. As a result, Dynamic Linear Regression serves as a useful baseline for identifying and quantifying the dynamic factors affecting energy consumption in electric vehicles.

Additionally, an AutoRegressive Integrated Moving Average (ARIMA) model was deployed. ARIMA has proven effective in handling complex datasets by capturing temporal dependencies within time-series data. As discussed by Moslemi et al. [33], ARIMA primarily focuses on autoregressive and moving average components but can also handle non-stationary data through differencing, making it a suitable choice for energy consumption forecasting.



4.2. Different Input Space

As mentioned in the Introduction, three different approaches were followed regarding the input space. Specifically, to capture the temporal dependencies in energy consumption, three distinct feature sets were implemented:

1. **Energy Consumption Lags Only**: This configuration relies solely on past values of energy consumption to predict future consumption, identifying auto-correlations within the dataset [34].

 doi:10.20944/preprints202503.0161.v1

2. **Feature Lags Only**: his setup includes only lagged values of relevant features (e.g., vehicle speed, air-conditioning, and environmental factors), excluding past energy consumption values. This approach evaluates the influence of external factors on energy consumption over time [35].
3. **Combined Lags**: This configuration integrates both energy consumption lags and feature lags, providing a comprehensive analysis of the relationships between past consumption patterns and vehicle dynamics.

These configurations were designed to assess the predictive strength of different feature structures and identify the most effective approach for modeling temporal energy consumption patterns. For each model, three distinct versions were implemented to examine the relationship between the target variable (energy consumption at one-, five-, and ten-minute intervals) and additional features, as well as their impact on prediction accuracy.

For the XGBoost model, three distinct versions were developed: i) Learnable Edge Collaborative Filtering (LECF) [36] that incorporates lags from both the target variable (energy consumption) and external vehicle features, ii) a version utilizing only historical values of energy consumption, and iii) a version considering only historical values of external features. A similar approach with XGBoost was followed for the Random Forest algorithm to implement these three different configurations. For the regression-based models, a Dynamic Regression model was implemented to i) capture historical values from both energy consumption and external features (representing the LECF version) ii) capture only exogenous features (analogous to the LF version). Finally, an ARIMA model was deployed to focus exclusively on historical energy consumption values, without incorporating exogenous variables.

4.3. Configurations

To optimize model performance using the eVED dataset, a grid search was employed to identify the most suitable hyperparameters for each model. Grid search [37] is an exhaustive search technique that evaluates multiple combinations of hyperparameters, ensuring the selection of an optimal configuration to enhance predictive accuracy.

For the Random Forest model, 300 decision trees were selected to enhance robustness and reduce variance. The maximum tree depth was set to 20, balancing complexity and generalization, allowing the model to learn patterns without overfitting to the training data. Additionally, the minimum number of samples required to split an internal node was set to 2, ensuring trees grow to their full potential. To guarantee reproducibility, a fixed random seed was used.

For the XGBoost model, various hyperparameter combinations were tested, with the following configuration proving most effective. The maximum tree depth was set to 3 to prevent overfitting while capturing essential patterns. A learning rate of 0.2 was chosen to balance learning speed and error minimization without overfitting. The number of boosting rounds was set to 300, providing sufficient iterations while incorporating early stopping to prevent overfitting. The objective function was set to minimizing squared error, a common loss function for regression problems. A fixed random seed was used to ensure reproducibility.

For dynamic regression models, an extensive grid search was conducted to determine the optimal lag selection for both the target variable (energy consumption) and exogenous variables such as vehicle speed, acceleration, elevation, and air conditioning. The most effective configuration included up to four lagged values, allowing the model to capture temporal dependencies and the dynamic effects of external factors over time. This approach enabled the model to learn both short-term variations and long-term trends in energy consumption.

A similar hyperparameter tuning approach was applied to the ARIMA model to determine the most suitable configuration for the dataset. After an extensive grid search, an (3,0,1) order was found to be the most effective across different vehicles and datasets. The autoregressive (AR) component of 3 uses three lagged observations of energy consumption to predict future values, effectively capturing dependencies over longer intervals. The differencing order of 0 indicates no differencing was required

 doi:10.20944/preprints202503.0161.v1

to achieve stationarity, avoiding unnecessary complexity while maintaining predictive accuracy. The moving average (MA) component of 1 uses a single lagged forecast error to correct future predictions, smoothing short-term fluctuations in energy consumption. Additional hyperparameter tuning was performed to optimize the differencing order and enhance the model's fit to the dataset's unique temporal patterns. By focusing on autoregressive components and smoothing forecast errors, the ARIMA model effectively captures energy consumption variations while leveraging historical data. This makes it particularly useful for analyzing temporal patterns, even when external exogenous variables are not considered.

Beyond hyperparameter tuning, lag selection was optimized for each algorithm through an exhaustive search to determine the most suitable lag number for each sub-dataset. This process provided a comprehensive understanding of feature distributions and variability, ensuring each model was tailored to the specific characteristics of the dataset. Finally, the dataset was spitted into training, validation and testing set following a distribution of $60:20:20$ split aligned with the literature [24,25]. All models were trained with different number of lags in order to identify the most effective lag number for every model. Lag observation differs from dataset to dataset as three different sub-datasets were utilized for every vehicle (one-minute, five-minute, and ten-minute intervals). The overall methodology is depicted in Figure 1.

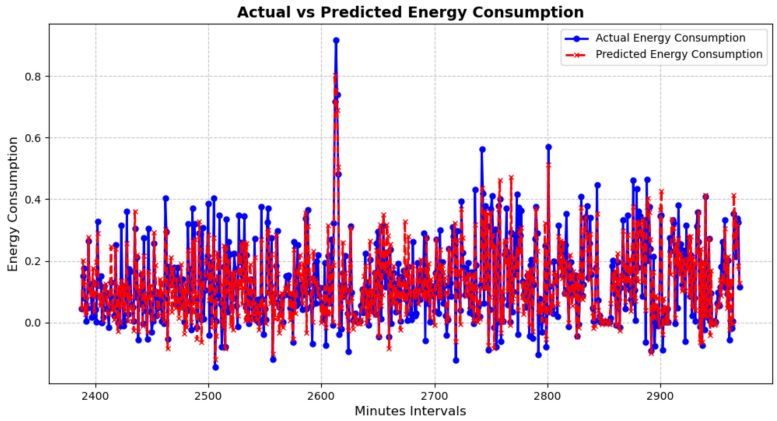


Figure 1. XGBoost Model LECF, Vehicle 455, 1-minute intervals.

4.4. Evaluation Metrics

In order to evaluate the model performance and monitor how they behave in different conditions and different vehicles, the following well-established evaluation metrics were used: Mean Absolute Error (MAE), Median Absolute Error (MdAE), and RMSE, along with their normalized versions. The selection of the aforementioned metrics is aligned with the respective literature [15,38,39]. The following formulas show how these metrics are calculated:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{1}$$

where e_i is the error between the actual and predicted energy consumption values, whereas n is the number of values.

$$MdAE = median(e_i) \tag{2}$$

expressing the median error, thus more robust and resilient against outliers compared to MAE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2} \tag{3}$$

 doi:10.20944/preprints202503.0161.v1

the most stricter metric, as penalizes the big errors more due to the square power

$$nMAE = \frac{MAE}{max(y) - min(y)} \tag{4}$$

where y is the vector of the actual energy consumption values

$$nMdAE = \frac{MdAE}{max(y) - min(y)} \tag{5}$$

$$nRMSE = \frac{RMSE}{max(y) - min(y)} \tag{6}$$

5. Experimental Results

In this paper, we presented an Automated Refueling Service aimed at optimizing energy consumption management for Electric Vehicles (EVs) through predictive analytics and real-time monitoring. By leveraging machine learning algorithms and real-time data collected from vehicle diagnostics, our system successfully provides accurate energy consumption forecasts and refueling notifications, contributing to more efficient fuel usage and enhanced driver convenience.

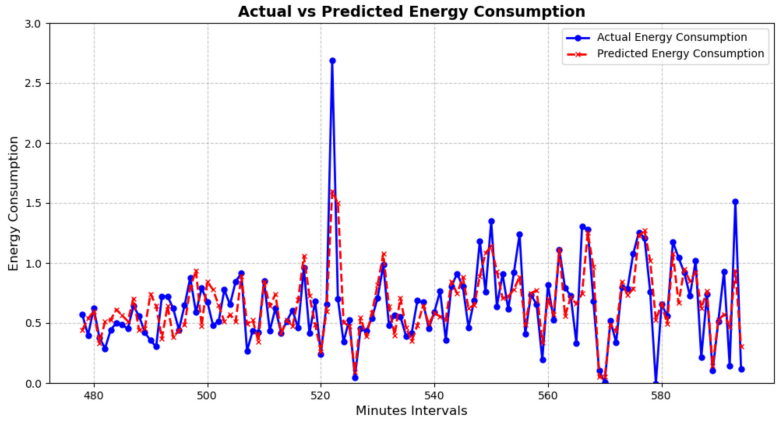


Figure 2. XGBoost Model LECF, Vehicle 455, 5-minute intervals.

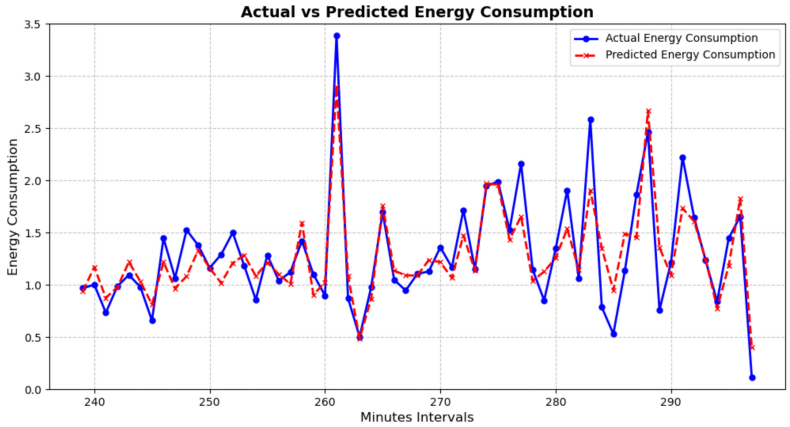


Figure 3. XGBoost Model LECF, Vehicle 455, 10-minute intervals.

This section aims to provide extensive results of the experiments conducted using XGBoost model, Random Forest (RF) and Dynamic Regression (DR) as well as ARIMA, utilizing every electric vehicle

 doi:10.20944/preprints202503.0161.v1

from eVED dataset. As described above, three different sub-datasets for every vehicle were created (1-minute, 5-minute, and 10-minute).

Table 3. Model Metrics for Each Vehicle at 1-minute, 5-minute, and 10-minute Intervals.

1-minute Intervals

Model	Vehicle 455					Vehicle 10					Vehicle 541				
	MAE	RMSE	MDAE	nMAE	nRMSE	nMDAE	MAE	RMSE	MDAE	nMAE	nRMSE	nMDAE	MAE	RMSE	MDAE	nMAE	nRMSE	nMDAE
XGBoost LECF	0.04152	**0.05786**	0.03044	0.02817	0.03925	0.02065	0.05220	0.07156	0.03411	0.05469	0.07499	0.03574	0.12641	0.13683	0.11047	0.12912	0.13977	0.11284
XGBoost LEC	0.11507	0.14201	0.10952	0.08429	0.10403	0.08023	0.10364	**0.13502**	0.08598	0.11233	0.14635	0.09319	0.13739	0.18289	0.16045	0.14034	0.18682	0.16389
XGBoost LF	0.04526	**0.06080**	0.03404	0.03315	0.04454	0.02494	0.05711	0.07676	0.04182	0.06191	0.08320	0.04533	0.15867	0.19568	0.10816	0.16208	0.19988	0.11049
RF LECF	0.04902	**0.06909**	0.03449	0.03569	0.05031	0.02511	0.05853	0.07951	0.04358	0.06134	0.08332	0.04566	0.12842	0.14573	0.10596	0.13117	0.14886	0.10823
RF LEC	0.05222	**0.07489**	0.03411	0.03802	0.05453	0.02484	0.06271	0.08597	0.04301	0.06572	0.09008	0.04507	0.09538	0.10608	0.10296	0.09742	0.10836	0.10517
RF LF	0.04903	**0.06902**	0.03451	0.03570	0.05026	0.02513	0.05822	0.07919	0.04464	0.06101	0.08298	0.04678	0.12667	0.14672	0.10579	0.12939	0.14987	0.10806
DR LEFC	0.05798	**0.07540**	0.04469	0.04222	0.05490	0.03254	0.06053	0.08225	0.04671	0.06343	0.08602	0.04915	0.13086	0.17769	0.12376	0.13367	0.18150	0.12642
DR LF	0.05870	**0.07574**	0.04754	0.04274	0.05515	0.03462	0.06117	0.08196	0.04718	0.06410	0.08588	0.04944	0.13086	0.17769	0.12376	0.13367	0.18151	0.12642
ARIMA	0.11360	0.13865	0.10988	0.08322	0.10157	0.08049	0.10770	0.14059	0.08901	0.11768	0.15361	0.09726	0.12946	0.15594	0.11722	0.13649	0.16442	0.12359

5-minute Intervals

Model	Vehicle 455					Vehicle 10					
	MAE	RMSE	MDAE	nMAE	nRMSE	nMDAE	MAE	RMSE	MDAE	nMAE	nRMSE	nMDAE
XGBoost LECF	0.16136	**0.23006**	0.11206	0.05966	0.08507	0.04144	0.18298	0.23141	0.15553	0.08897	0.11252	0.07563
XGBoost LEC	0.29470	**0.42124**	0.21269	0.10897	0.15577	0.07865	0.38189	0.47821	0.34496	0.18570	0.23253	0.16774
XGBoost LF	0.15091	**0.21649**	0.10313	0.05580	0.08005	0.03813	0.18313	0.23135	0.14905	0.08905	0.11250	0.07247
RF LECF	0.15562	**0.22589**	0.12717	0.05118	0.07506	0.04175	0.18412	0.23863	0.16020	0.08875	0.11503	0.07722
RF LEC	0.15588	**0.22861**	0.12717	0.05118	0.07506	0.04175	0.18716	0.24248	0.12935	0.09022	0.11688	0.06235
RF LF	0.15603	**0.22721**	0.10634	0.05123	0.07460	0.03491	0.18207	0.23699	0.14648	0.08777	0.11424	0.07061
DR LEFC	0.18911	0.24729	0.14322	0.06209	0.08119	0.04702	0.18482	**0.21523**	0.15754	0.08909	0.10375	0.07594
DR LF	0.18911	0.24729	0.14322	0.06209	0.08119	0.04702	0.18482	**0.21523**	0.15754	0.08909	0.10375	0.07594
ARIMA	0.27009	0.36427	0.23336	0.10093	0.13612	0.08720	0.31910	0.39825	0.28685	0.15516	0.19365	0.13948

10-minute Intervals

Model	Vehicle 455					Vehicle 10					
	MAE	RMSE	MDAE	nMAE	nRMSE	nMDAE	MAE	RMSE	MDAE	nMAE	nRMSE	nMDAE
XGBoost LECF	0.18887	**0.25073**	0.12800	0.03725	0.04945	0.02524	0.27514	0.31889	0.27567	0.08223	0.09530	0.08238
XGBoost LEC	0.57992	0.72800	0.49361	0.11438	0.14359	0.09736	0.52889	**0.63396**	0.50195	0.15806	0.18946	0.15001
XGBoost LF	0.19562	**0.26456**	0.13190	0.03858	0.05218	0.02601	0.27167	0.34555	0.21579	0.08119	0.10327	0.06449
RF LECF	0.19891	**0.26294**	0.14714	0.03414	0.04514	0.02526	0.30063	0.37340	0.20712	0.08984	0.11159	0.06190
RF LEC	0.18024	**0.23738**	0.12965	0.03094	0.04075	0.02225	0.29087	0.35814	0.21356	0.08693	0.10703	0.06382
RF LF	0.20089	**0.26366**	0.15228	0.03448	0.04526	0.02614	0.30308	0.37439	0.22552	0.09057	0.11189	0.06740
DR LEFC	0.24259	**0.34969**	0.19803	0.04164	0.06003	0.03399	0.28963	0.35169	0.25501	0.08655	0.10510	0.07621
DR LF	0.24699	0.35302	0.18761	0.04240	0.06060	0.03220	0.29148	**0.35145**	0.26734	0.08711	0.10503	0.07989
ARIMA	0.46426	0.57490	0.40522	0.09157	0.11339	0.07992	0.49861	0.62677	0.42740	0.14901	0.18732	0.12773

At the 1-minute interval, XGBoost consistently proved to be the best-performing algorithm across vehicles and configurations. For example, in the LECF configuration, XGBoost achieved a MAE of 0.04152, an RMSE of 0.05786, and an nRMSE of 0.03925 for Vehicle 455, demonstrating its capacity to accurately capture short-term variations in energy consumption. This trend was consistent across other

 doi:10.20944/preprints202503.0161.v1

vehicles, with Vehicle 10 achieving an RMSE of 0.07156 in the same configuration, significantly better than other models. Random Forest, while robust, exhibited slightly higher errors. For instance, Vehicle 455 under the Random Forest LECF (lags energy consumption and features) configuration recorded an RMSE of 0.06909, which was higher than XGBoost's performance. Similarly, for Vehicle 10, Random Forest under the LECF configuration achieved an RMSE of 0.07951, reflecting its slightly weaker performance in capturing complex patterns at this short interval. Dynamic Regression (DR) models performed comparably well, particularly for Vehicle 455, where an RMSE of 0.07540 was recorded in the LECF configuration, and for Vehicle 10, an RMSE of 0.06053 was observed. However, DR models generally fell behind XGBoost in accuracy, as seen with Vehicle 541, where DR models recorded higher errors compared to XGBoost. ARIMA, on the other hand, showed mixed results. While it achieved reasonable accuracy for Vehicle 455 with an RMSE of 0.13865, its performance deteriorated for Vehicle 10, where an RMSE of 0.14059 was observed. This inconsistency highlighted ARIMA's sensitivity to data variability.

At the 5-minute interval, data aggregation resulted in increased errors across all models, but XGBoost maintained its dominance. For instance, in the LF configuration, XGBoost recorded an RMSE of 0.21649 for Vehicle 455, the lowest among all models for this interval. For Vehicle 10, XGBoost LECF achieved an RMSE of 0.23141, showcasing its robustness even at this aggregated level. Random Forest continued to deliver respectable results, though it trailed behind XGBoost. For Vehicle 455, Random Forest LF recorded an RMSE of 0.22721, while for Vehicle 10, its performance was slightly less accurate, with an RMSE of 0.23699. Dynamic Regression models remained competitive but slightly inconsistent. For Vehicle 10, DR models recorded an RMSE of 0.21523 under both LF and LECF configurations, which were comparable to XGBoost but higher than Random Forest in some cases. However, for Vehicle 541, DR models produced higher errors, revealing limitations in handling data variability. ARIMA demonstrated greater variability in its performance, particularly as the interval length increased. While it achieved competitive results for Vehicle 455, with an nRMSE of 0.13612, its accuracy declined for other vehicles at longer intervals. For instance, Vehicle 10 under ARIMA recorded a significantly higher RMSE of 0.19365, highlighting the model's increased errors and its inherent difficulty in handling aggregated data and extended prediction intervals.

At the 10-minute interval, the loss of granularity further increased the gap between the models. XGBoost continued to dominate, particularly in the LECF configuration, where it achieved an RMSE of 0.25073 for Vehicle 455. This result demonstrated its ability to maintain accuracy even in longer-term predictions. For Vehicle 10, XGBoost recorded an RMSE of 0.31889, highlighting its superior adaptability. Random Forest provided slightly less accurate predictions. For Vehicle 455, Random Forest with the LECF configuration recorded an RMSE of 0.26294, while for Vehicle 10, the RMSE reached 0.37340, indicating a slight decrease in performance at this interval compared to XGBoost. Dynamic Regression models exhibited moderate accuracy at this interval. For Vehicle 455, DR under the LECF configuration recorded an RMSE of 0.34969, while for Vehicle 10, it achieved 0.35169, demonstrating a stable but less competitive performance compared to XGBoost. ARIMA struggled significantly at the 10-minute interval. Its errors for Vehicle 10, for example, escalated to an RMSE of 0.62677, reflecting its difficulty in adapting to longer intervals with aggregated data. Even for Vehicle 455, where ARIMA typically performed better, its RMSE reached 0.57490, substantially higher than XGBoost.

Across all intervals, **XGBoost** consistently **outperformed** other models, making it the most reliable choice for energy consumption predictions. For shorter intervals, particularly the 1-minute interval, XGBoost in the LECF configuration demonstrated remarkable accuracy, leveraging both feature and energy consumption lags. For longer intervals, the LF configuration of XGBoost provided a good balance of simplicity and performance. Random Forest proved to be a dependable alternative, offering robust predictions but consistently trailing XGBoost. Its performance was particularly competitive at the 5-minute interval but weaker at the 10-minute interval. Dynamic Regression (DR) models

 doi:10.20944/preprints202503.0161.v1

offered strong competition, especially at shorter intervals, but lacked the consistency of XGBoost across different vehicles and intervals. ARIMA, while effective in specific cases, exhibited significant variability and struggled with aggregated data at longer intervals. Its performance was often surpassed by both XGBoost and Random Forest. These results highlight the robustness and adaptability of XGBoost, making it the most suitable model for energy consumption prediction across varying time intervals and vehicle data.

6. Conclusions

This research contributes to progress toward SDGs, particularly in sustainable transportation, by offering a robust and adaptive service for optimizing energy consumption in electric vehicles, thereby supporting decarbonization goals and the broader adoption of eco-friendly mobility solutions.

This research study proposed a novel methodology-framework for predicting energy consumption in electric vehicles, based on an advanced combination of machine learning models and state-of-the-art statistical data analysis techniques. The presented approach enabled the identification of temporal dependencies and facilitated more robust predictions by structuring the eVED dataset into multiresolution time intervals and applying lagged features.

Experimental results demonstrated that XGBoost, Random Forest, and regression-based models, with different configurations in lag structure, achieved high prediction accuracy. Additionally, the 1-minute intervals provided the most detailed insights, while 5-minute intervals offered a practical balance between accuracy and computational efficiency. The findings indicate that integrating time-interval-based energy consumption modeling with machine learning enhances prediction accuracy across various driving scenarios and vehicle characteristics. This improvement supports better energy management strategies, contributing to eco-driving, refueling optimization, and range estimation.

Future work will integrate real-time data streams to enhance the applicability of the proposed methodology in dynamic environments. In particular, ML operations will be incorporated into the best-performing model, XGBoost, which utilizes both historical energy consumption and external features transforminf the methodology into a service. This service will be deployed in a real-life environment within the context of the SINNOGENES project.

Funding: This work was funded by the European Union's Horizon Europe project SINNOGENES (Storage innovations for green energy systems), under Grant Agreement No. 101096992.

References

1. Jose, P. S., A. Natarajan, S. Karthikeyan, and T. Bogaraj. Environmental and social impact of electric vehicles. In *Advanced Technologies in Electric Vehicles*. Elsevier, 2024, pp. 107–125.
2. Sanguesa, J. A., V. Torres-Sanz, P. Garrido, F. J. Martinez, and J. M. Marquez-Barja. A review on electric vehicles: Technologies and challenges. *Smart Cities*, Vol. 4, No. 1, 2021, pp. 372–404.
3. Chen, T., B. Zhang, H. Pourbabak, A. Kavousi-Fard, and W. Su. Optimal routing and charging of an electric vehicle fleet for high-efficiency dynamic transit systems. *IEEE Transactions on Smart Grid*, Vol. 9, No. 4, 2016, pp. 3563–3572.
4. Alanazi, F. Electric vehicles: benefits, challenges, and potential solutions for widespread adaptation. *Applied Sciences*, Vol. 13, No. 10, 2023, p. 6016.
5. Akshay, K., G. H. Grace, K. Gunasekaran, and R. Samikannu. Power consumption prediction for electric vehicle charging stations and forecasting income. *Scientific Reports*, Vol. 14, No. 1, 2024, p. 6497.
6. Longo, M., D. Zaninelli, F. Viola, P. Romano, and R. Miceli. How is the spread of the Electric Vehicles? In *2015 IEEE 1st International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI)*. IEEE, 2015, pp. 439–445.
7. Klyuev, R. V., I. D. Morgoev, A. D. Morgoeva, O. A. Gavrina, N. V. Martyushev, E. A. Efremenkov, and Q. Mengxu. Methods of forecasting electric energy consumption: A literature review. *Energies*, Vol. 15, No. 23, 2022, p. 8919.

 doi:10.20944/preprints202503.0161.v1

8. Chudy-Laskowska, K. and T. Pisula. An Analysis of the Use of Energy from Conventional Fossil Fuels and Green Renewable Energy in the Context of the European Union's Planned Energy Transformation. *Energies*, Vol. 15, No. 19, 2022, p. 7369.
9. Huang, H., B. Li, Y. Wang, Z. Zhang, and H. He. Analysis of factors influencing energy consumption of electric vehicles: Statistical, predictive, and causal perspectives. *Applied Energy*, Vol. 375, 2024, p. 124110.
10. Petkevicius, L., S. Saltenis, A. Civilis, and K. Torp. Probabilistic deep learning for electric-vehicle energy-use prediction. In *Proceedings of the 17th International Symposium on Spatial and Temporal Databases*. 2021, pp. 85–95.
11. Hua, Y., M. Sevegnani, D. Yi, A. Birnie, and S. McAslan. Fine-grained RNN with transfer learning for energy consumption estimation on EVs. *IEEE Transactions on Industrial Informatics*, Vol. 18, No. 11, 2022, pp. 8182–8190.
12. Athanasakis, E., G. Spanos, A. Papadopoulos, A. Lalas, K. Votis, and D. Tzovaras. a Comprehensive Leakage-Free Forecasting Pipeline for Segmented Time Series: Application to Cross-Trip State-of-Charge Prediction in Automated Electric Vehicles. *IEEE Transactions on Intelligent Vehicles*.
13. Chen, X., Z. Lei, and S. V. Ukkusuri. Prediction of road-level energy consumption of battery electric vehicles. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 2550–2555.
14. Tang, X., T. Jia, X. Hu, Y. Huang, Z. Deng, and H. Pu. Naturalistic data-driven predictive energy management for plug-in hybrid electric vehicles. *IEEE Transactions on Transportation Electrification*, Vol. 7, No. 2, 2020, pp. 497–508.
15. Polymeni, S., V. Pitsiavas, G. Spanos, Q. Matthewson, A. Lalas, K. Votis, and D. Tzovaras. Toward sustainable mobility: AI-enabled automated refueling for Fuel Cell Electric Vehicles. *Energies*, Vol. 17, No. 17, 2024, p. 4324.
16. Brooker, A., J. Gonder, L. Wang, E. Wood, S. Lopp, and L. Ramroth. *FASTSim: A model to estimate vehicle efficiency, cost and performance*. Tech. rep., SAE technical paper, 2015.
17. Zhang, S., D. Fatih, F. Abdulqadir, T. Schwarz, and X. Ma. Extended vehicle energy dataset (eVED): an enhanced large-scale dataset for deep learning on vehicle trip energy consumption. *arXiv preprint arXiv:2203.08630*.
18. Oh, G., D. J. Leblanc, and H. Peng. Vehicle energy dataset (VED), a large-scale dataset for vehicle energy consumption research. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 4, 2020, pp. 3302–3312.
19. Mediouni, H., A. Ezzouhri, Z. Charouh, K. El Harouri, S. El Hani, and M. Ghogho. Energy consumption prediction and analysis for electric vehicles: A hybrid approach. *energies*, Vol. 15, No. 17, 2022, p. 6490.
20. Yang, X., C. Zhuge, C. Shao, Y. Huang, J. H. C. G. Tang, M. Sun, P. Wang, and S. Wang. Characterizing mobility patterns of private electric vehicle users with trajectory data. *Applied Energy*, Vol. 321, 2022, p. 119417.
21. Spanos, G., K. M. Giannoutakis, K. Votis, and D. Tzovaras. Combining statistical and machine learning techniques in IoT anomaly detection for smart homes. In *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. IEEE, 2019, pp. 1–6.
22. Spanos, G., K. M. Giannoutakis, K. Votis, B. Viaño, J. Augusto-Gonzalez, G. Aivatoglou, and D. Tzovaras. A lightweight cyber-security defense framework for smart homes. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 2020, pp. 1–7.
23. Meire, M., M. Ballings, and D. Van den Poel. The added value of auxiliary data in sentiment analysis of Facebook posts. *Decision Support Systems*, Vol. 89, 2016, pp. 98–112.
24. Aivatoglou, G., M. Anastasiadis, G. Spanos, A. Voulgaridis, K. Votis, and D. Tzovaras. A tree-based machine learning methodology to automatically classify software vulnerabilities. In *2021 ieee international conference on cyber security and resilience (csr)*. IEEE, 2021, pp. 312–317.
25. Aivatoglou, G., M. Anastasiadis, G. Spanos, A. Voulgaridis, K. Votis, D. Tzovaras, and L. Angelis. A RAkEL-based methodology to estimate software vulnerability characteristics & score-an application to EU project ECHO. *Multimedia Tools and Applications*, Vol. 81, No. 7, 2022, pp. 9459–9479.
26. Rathore, H., H. K. Meena, and P. Jain. Prediction of ev energy consumption using random forest and xgboost. In *2023 International Conference on Power Electronics and Energy (ICPEE)*. IEEE, 2023, pp. 1–6.
27. Breiman, L. Random forests. *Machine learning*, Vol. 45, 2001, pp. 5–32.

 doi:10.20944/preprints202503.0161.v1

28. Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.
29. Pankratz, A. *Forecasting with dynamic regression models*. John Wiley & Sons, 2012.
30. Wang, Z., Y. Wang, R. Zeng, R. S. Srinivasan, and S. Ahrentzen. Random Forest based hourly building energy prediction. *Energy and Buildings*, Vol. 171, 2018, pp. 11–25.
31. Ma, Z., H. Chang, Z. Sun, F. Liu, W. Li, D. Zhao, and C. Chen. Very short-term renewable energy power prediction using XGBoost optimized by TPE algorithm. In *2020 4th International Conference on HVDC (HVDC)*. IEEE, 2020, pp. 1236–1241.
32. Watanabe, S. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*.
33. Moslemi, Z., L. Clark, S. Kernal, S. Rehome, S. Sprengel, A. Tamizifar, S. Tuli, V. Chokshi, M. Nomeli, E. Liang, et al. Comprehensive Forecasting of California's Energy Consumption: A Multi-Source and Sectoral Analysis Using ARIMA and ARIMAX Models. *arXiv preprint arXiv:2402.04432*.
34. Sun, Q., J. Liu, X. Rong, M. Zhang, X. Song, Z. Bie, and Z. Ni. Charging load forecasting of electric vehicle charging station based on support vector regression. In *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*. IEEE, 2016, pp. 1777–1781.
35. Saly, G., F. Szauter, and S. Kocsis Szürke. Comprehensive Analysis of the Factors Affecting the Energy Efficiency of Electric Vehicles and Methods to Reduce Consumption: A Review. *Engineering Proceedings*, Vol. 79, No. 1, 2024, p. 79.
36. Xiao, S., Y. Shao, Y. Li, H. Yin, Y. Shen, and B. Cui. LECF: recommendation via learnable edge collaborative filtering. *Science China Information Sciences*, Vol. 65, No. 1, 2022, p. 112101.
37. Liashchynskyi, P. and P. Liashchynskyi. Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv preprint arXiv:1912.06059*.
38. Hyndman, R. J. and A. B. Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, Vol. 22, No. 4, 2006, pp. 679–688.
39. Polymeni, S., G. Spanos, D. Tsiktsiris, E. Athanasakis, K. Votis, D. Tzovaras, and G. Kormentzas. everWeather: A Low-Cost and Self-Powered AIoT Weather Forecasting Station for Remote Areas. In *Environmental Informatics*. Springer, 2023, pp. 141–158.