



PREDICTING DAILY ENERGY CONSUMPTION IN EV CHARGING STATIONS USING TREE-BASED MODELS

Merve **ERTARĞIN*** 

¹University of Munzur, Department of Electrical and Electronics Engineering, Tunceli, Türkiye

*Corresponding author's e-mail: merveboydak@munzur.edu.tr

Abstract: Electric vehicles (EVs) have emerged as a cornerstone in the transition toward sustainable transportation, offering a cleaner alternative to conventional fossil fuel-based mobility. Their integration into modern transportation systems not only supports global decarbonization efforts but also requires robust energy management solutions to address the growing demands on charging infrastructure. Ensuring efficient energy consumption forecasting is vital for optimizing EV charging station operations and maintaining grid stability. This study aims to predict the daily energy consumption of an EV charging station using a public dataset. Key features such as charging duration, vehicle count, mean and median connection time were utilized for the prediction. Both the original dataset and its z-score-cleaned version were employed to evaluate the models' performance under different data conditions. Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) models were implemented, with hyperparameter optimization conducted via GridSearchCV. The results revealed that all four models achieved high accuracy in predicting daily energy consumption across both datasets, with R^2 values exceeding 0.97. Specifically, for the z-score-cleaned dataset, LightGBM demonstrated the best performance with the lowest RMSE (37.35 kWh) and MAE (28.93 kWh), as well as the highest R^2 value (0.9727) followed by XGBoost, RF and DT.

Keywords: Electric vehicles, energy consumption forecasting, LightGBM, XGBoost, random forest.

Received: April 12, 2025

Accepted: July 29, 2025

1. Introduction

The transition toward EVs is a key component of the global shift to sustainable and green energy solutions [1]. As governments and industries push for reduced carbon emissions and increased reliance on renewable energy, the adoption of EVs has been accelerating [2]. While the transition from fossil fuels to renewable energy in vehicular technologies is still in its infancy, it is driven by the shared objective of fostering green energy and achieving sustainable mobility [3]. However, it also presents challenges, particularly in managing the demand for charging infrastructure and optimizing energy consumption patterns.

The increasing demand for EV charging stations has highlighted the need for efficient energy distribution and infrastructure planning [4]. Accurate forecasting of energy consumption at charging stations is critical to preventing resource shortages, minimizing energy waste, and ensuring a seamless

user experience. By understanding charging behaviors and anticipating future demand, stakeholders can implement strategies to improve grid stability and optimize resource allocation.

Optimization and prediction models have become essential tools in addressing these challenges. While optimization models focus on the effective allocation of resources and scheduling of charging sessions, prediction models aim to forecast charging demand based on historical data. Together, these approaches play a pivotal role in maintaining a balance between energy supply and demand, supporting grid reliability and enhancing the scalability of charging networks. Advanced data-driven techniques, such as machine learning and statistical modeling, have emerged as effective methods for solving these complex problems.

Recent advancements in predictive and optimization methods have provided promising solutions for these challenges. For instance, advanced Artificial Neural Network (ANN) models, including Feed-Forward Neural Network (FFNN) and General Regression Neural Networks (AGRNN, IGRNN), are utilized to predict hourly photovoltaic (PV) energy generation and solar irradiance for off-grid PV-powered charging stations [5]. By integrating these predictions with Energy Storage Systems (ESS), the study reports significant improvements in load balancing and a reduction in peak energy demand by up to 20 kW, ensuring efficient and reliable operations of solar-powered charging networks. Similarly in [6], the importance of multi-feature data fusion is highlighted to enhance energy demand forecasting accuracy at EV charging stations. The study uses Long Short-Term Memory (LSTM) models, integrating meteorological data, such as temperature and wind speed, with real-world energy demand observations. Such an innovative approach achieves a 3.29% prediction error, underscoring the effectiveness of deep learning models in capturing complex temporal patterns in dynamic datasets. Furthermore in [7], a Hybrid Particle Swarm Optimization (HPSO) framework, combined with ETOPSIS, is proposed to address the spatial and temporal dimensions of charging demand by optimizing the siting and sizing of EV charging stations. By incorporating trip chain analysis and EV users' state-of-charge (SOC) data, the study develops a multi-objective optimization approach that balances grid load variance, electricity costs, and user convenience. The proposed framework demonstrates its effectiveness in a simulated 33-node distribution and traffic network, providing valuable insights for future infrastructure planning. While these methods address specific predictive and optimization needs, reinforcement learning has emerged as a versatile approach for dynamic decision-making challenges.

Reinforcement learning (RL) has become a transformative approach, particularly suited for dynamic and adaptive optimization tasks within EV infrastructure management. For instance in [8], a Soft Actor-Critic (SAC) algorithm is utilized to optimize EV-PV coordination by incorporating short-term predictions of PV energy output and EV arrivals, enhancing profitability and grid stability. Similarly, [9] leverages SAC to optimize the placement of EV charging stations, integrating traffic density and public transport data to strategically position stations in high-demand areas. In [10], an RL framework based on a sequence-to-sequence model tackles the Charging Station Deployment Problem (CSDP), balancing transportation and power grid constraints to minimize costs while ensuring network reliability. Additionally, [11] introduces the Hyperopia SARSA (HSA) algorithm for real-time pricing and scheduling, achieving significant profitability improvements and mitigating peak grid loads. Collectively, these studies demonstrate RL's versatility in addressing both static planning problems and dynamic operational challenges, paving the way for smarter, more efficient, and sustainable EV infrastructure systems.

In contrast to dynamic optimization methods like RL, tree-based models focus on addressing predictive challenges within EV infrastructure planning. Their scalability, interpretability, and ability to handle complex datasets make them indispensable tools in this domain. For instance, [12] highlights the effectiveness of LightGBM in ultra-short-term energy demand forecasting, where Bayesian optimization and advanced feature extraction techniques are utilized to enhance prediction accuracy and robustness.

Similarly, [13] employs Random Forest (RF) to forecast charging demand based on traffic and weather data, integrating these predictions into a multi-objective optimization framework using Particle Swarm Optimization (PSO) to optimize the placement and sizing of charging stations. In addition, [14] applies an enhanced RF algorithm to predict daily and real-time EV charging loads, achieving a mean absolute percentage error (MAPE) of less than 11%, thereby validating its capability in spatial and temporal load forecasting. Moreover, [15] presents an advanced implementation of the XGBoost regression tree algorithm for EV energy load forecasting, tested on three real-world datasets. This study demonstrates XGBoost's superiority over eight state-of-the-art algorithms, including ANN, CNN, and LSTM, in terms of RMSE and MAE, showcasing its prediction accuracy and generalization capabilities across diverse charging environments.

Tree-based models have gained prominence in predictive modeling for their ability to balance accuracy and interpretability, making them highly suitable for energy demand forecasting in EV charging infrastructure. This study explores the use of a public dataset to predict daily energy consumption, focusing on features such as charging duration, vehicle count, mean and median connection time, which are critical indicators of energy usage. To evaluate the models' robustness and their response to outliers, the predictions were conducted using both the original dataset and a z-score-cleaned version. DT, RF, XGBoost and LightGBM were selected as the predictive models, with their hyperparameters optimized via GridSearchCV.

2. Materials and Method

2.1. Dataset

The effective management of EV charging infrastructure relies on accurate energy consumption forecasts, which require robust models and comprehensive datasets. The quality and structure of input data play a critical role in shaping model performance. In this study, a detailed dataset of daily electric vehicle charging activities [16] was employed, which was initially derived from 185 days of real-world observations at a public charging station located on the Caltech campus [17]. This dataset was extended to represent 29,600 days using Conditional Tabular Generative Adversarial Networks (CTGAN) [18] and Kernel Density Estimation (KDE). The generative approach ensured the preservation of key correlations, such as the relationship between charging duration and energy consumption, while KDE was utilized to replicate the distributional characteristics of EV connection times, reflecting realistic charging behaviors and patterns.

The dataset [19] comprises several key features that provide granular insights into EV charging behavior. For instance, the 'connectionTime' attribute captures the specific times EVs are connected to charging stations, formatted within a 24-hour scale. Additionally, 'chargingDuration' and 'kWhDelivered' offer critical data on the duration of charging sessions (in hours) and the corresponding energy consumption (in kWh). To facilitate chronological grouping, the 'dayIndicator' column aggregates all records sharing the same date, allowing for effective temporal analysis.

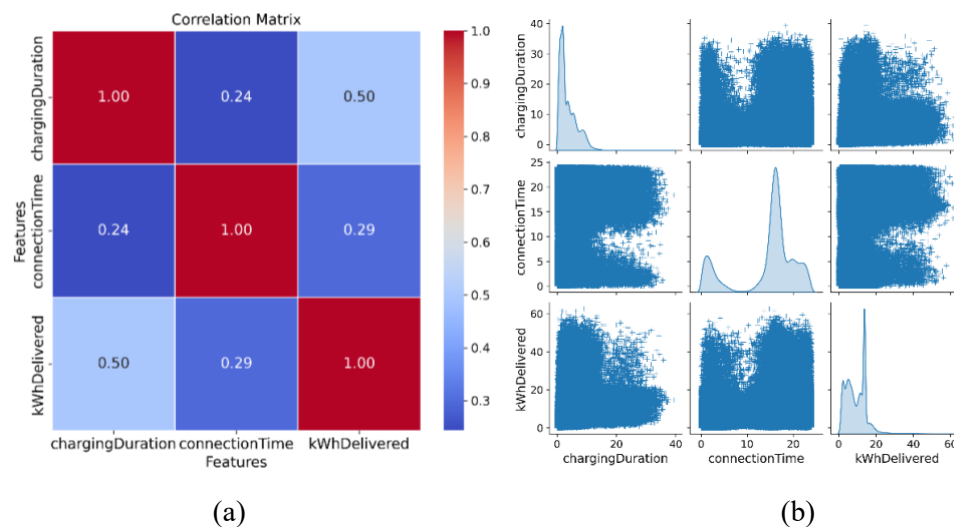


Figure 1. (a) Correlation matrix (b) Pair plot of the features.

In Figure 1 (a), the correlation matrix illustrates the relationships between features in the dataset. It highlights the degree of correlation between variables such as charging duration, connection time, and energy delivered. Notably, the matrix shows a stronger correlation between energy delivered and charging duration than between connection time and energy delivered, emphasizing the impact of active charging time on energy consumption compared to mere connection time.

In Figure 1 (b), a pair plot of the features "chargingDuration," "connectionTime," and "kWhDelivered" is presented. This visualization allows us to observe the distributions and relationships between each pair of variables. Each feature's histogram on the diagonal shows its distribution. "chargingDuration" and "kWhDelivered" are highly skewed towards lower values, indicating that most charging events have short durations and deliver lower amounts of energy. "connectionTime" shows distinct peaks, suggesting specific typical connection durations. The off-diagonal plots display the relationships between pairs of features. "chargingDuration" and "kWhDelivered" show a positive relationship, aligning with the correlation matrix findings that longer charging durations generally deliver more energy. "connectionTime" appears more dispersed with both "chargingDuration" and "kWhDelivered," suggesting a weaker linear relationship, as also indicated by the correlation matrix. This figure complements the correlation matrix by providing a more detailed view of the feature distributions and how they relate to one another visually.

When examining the basic statistical properties of the synthetic and Caltech datasets in terms of connection time, the mean values for both datasets appear to be quite similar. The average connection time in the synthetic dataset is 14.486 hours, while in the Caltech dataset it is calculated as 13.435 hours. For the charging duration variable, the mean values are 3.800 hours (synthetic) and 3.280 hours (Caltech), respectively. Similarly, for the energy consumption variable, the average values are 9.437 kWh in the synthetic dataset and 8.998 kWh in the Caltech dataset. The closeness of the mean values across all three variables indicates that the synthetic data successfully reflects the overall trends of the real data, and that the fundamental statistical structures were preserved during the data generation process.

However, in some cases, the standard deviation values suggest a more balanced or narrower distribution in the synthetic data. For instance, the standard deviation for connection time is 6.427 in the synthetic dataset and 7.006 in the Caltech dataset. For charging duration, these values are 3.387 (synthetic) and 3.060 (Caltech), respectively. For energy consumption, the standard deviations are 5.700 (synthetic) and 6.638 (Caltech). These observations suggest that the synthetic data tends to generate

values closer to the mean and represents extreme values in a more controlled manner. Overall, the findings indicate that the synthetic data is statistically consistent with the real data.

This study applied the z-score method to remove outliers, not as a primary objective, but to evaluate the impact of outlier removal on model performance. This process was intended to compare how the models perform with and without outliers, rather than to improve the dataset itself. The z-score method was applied specifically to the 'ChargingDuration' column to identify and exclude extreme values. Figure 2(a) illustrates the original dataset, while Figure 2(b) displays the cleaned dataset. In both figures, Charging Duration is represented on the x-axis and Energy Delivered on the y-axis. The total number of data points in the dataset decreased from 1965,239 in the original dataset to 1943,825 after the cleaning process. This indicates that approximately 21,414 data points were eliminated as a result of removing outliers.

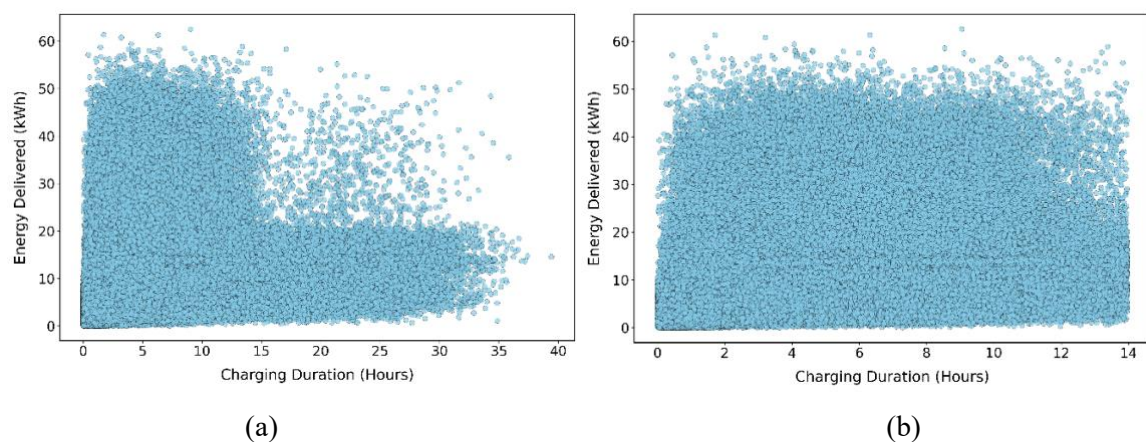


Figure 2. Charging duration vs. energy delivered for the (a) original dataset (b) cleaned dataset.

Table 1. Daily summaries of electric vehicle charging activities.

Day Indicator	Vehicle Count	Charging Duration	Mean connection Time	Median connection Time	kWh Delivered
1	43	156.648593	15.093014	15.924727	414.863606
2	59	216.930493	15.847774	16.025315	557.503552
3	54	228.013671	15.543766	16.390652	634.367046
4	55	252.782241	15.438984	16.243412	589.584292
5	47	179.476879	13.955142	15.885304	517.117013
6	38	186.018397	15.960160	16.796690	423.034731
7	60	233.466146	15.351743	16.233350	621.262425
8	70	327.139617	15.088815	15.992321	695.648204
9	20	84.293136	15.529071	16.355810	214.051579
10	20	75.959999	14.050489	16.309324	146.638061

The dataset was aggregated on a daily basis to generate comprehensive summaries of electric vehicle charging activities. All EVs with the same day indicator value were grouped together, representing the charging activities for that specific day. The total number of vehicles charged each day was determined by counting the entries with the same day indicator value, and this count was recorded as the “vehicle_count” variable. The daily total charging duration was calculated as the sum of the “chargingDuration” values, providing a measure of the total time spent charging vehicles on a given day. Similarly, the total daily energy consumption was calculated by summing the “kWhDelivered” values, representing the total energy delivered to all vehicles for each day. Furthermore, the daily average connection time was computed as the mean of the “connectionTime_decimal” values, while the median connection time was calculated to provide a robust measure of central tendency, capturing typical charging patterns. Table 1 presents the results for the first 10 days of the aggregated daily cleaned data.

2.2. Method

Figure 3 summarizes the proposed method. The dataset was split into training and testing subsets, allocating 80% for training and 20% for testing. Decision Tree, Random Forest [20], XGBoost [21], and LightGBM [22] algorithms were used to predict daily energy consumption. Figure 4 illustrates the working mechanisms of these models. All algorithms utilize decision trees but differ in how they construct and optimize these trees. DT represents the simplest form of a tree-based model, where a single tree is built by recursively splitting the data based on feature values to minimize prediction error. While DTs are easy to interpret and computationally efficient, they tend to overfit the training data and often lack the generalization capability required for complex tasks. RF employs a bagging approach, combining multiple DTs trained on bootstrapped subsets of the data, which helps reduce overfitting and improve model stability. In contrast, XGBoost and LightGBM use gradient boosting [23]. In the gradient boosting method, trees are trained sequentially, with each new tree correcting the errors of the previous ones. This sequential training makes gradient boosting algorithms slower compared to bagging methods, such as RF, where trees are trained independently, allowing for parallel processing and faster computation. However, LightGBM and XGBoost, which specifically implement the gradient boosting methodology, are highly effective for complex datasets, capturing nonlinear relationships, and producing more precise models. RF, on the other hand, excels in simpler and balanced datasets due to its bagging approach, which aggregates predictions from independently trained trees.

In this study, the GridSearchCV method, which combines hyperparameter search and cross-validation, was employed to optimize the model's performance. During this process, all possible combinations from a predefined hyperparameter grid were systematically tested. To tune the hyperparameters, a 5-fold cross-validation was conducted exclusively on the training set, ensuring that the test set remained unseen throughout the optimization process. This approach effectively prevented data leakage and allowed the test set to serve as an independent benchmark for assessing the model's generalization capabilities. Furthermore, the large size of the dataset mitigated potential concerns about information loss caused by splitting the data into training and test sets. With 80% of the data allocated to the training set, there was ample data to perform effective hyperparameter tuning and achieve robust model performance. As a result, the typical trade-off of not utilizing the entire dataset was negligible in this case, as the training data provided a comprehensive representation of the overall data distribution. Table 2 provides hyperparameters for LightGBM, XGBoost, RF and DT. The hyperparameter were selected through a combination of manual exploration and systematic refinement. While these ranges were informed by preliminary testing to capture strong performance, computational efficiency was also a key consideration. Moreover, the hyperparameter ranges were not kept identical across models, as each algorithm has different learning dynamics and requirements. For example, Random Forest

generally benefits from a higher number of estimators to stabilize its performance, especially in large datasets, due to its bagging-based nature. In contrast, boosting algorithms such as LightGBM and XGBoost typically achieve convergence with fewer estimators, owing to their sequential learning structure. Therefore, each range was tailored to reflect the operational characteristics of the respective algorithm while ensuring a fair and efficient comparison.

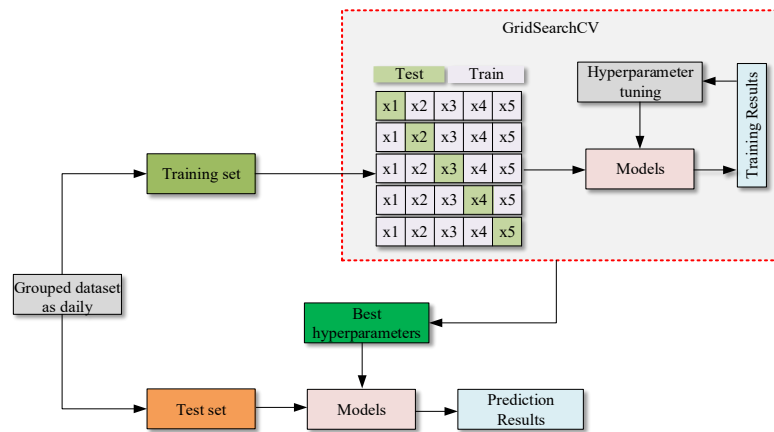


Figure 3. Proposed method.

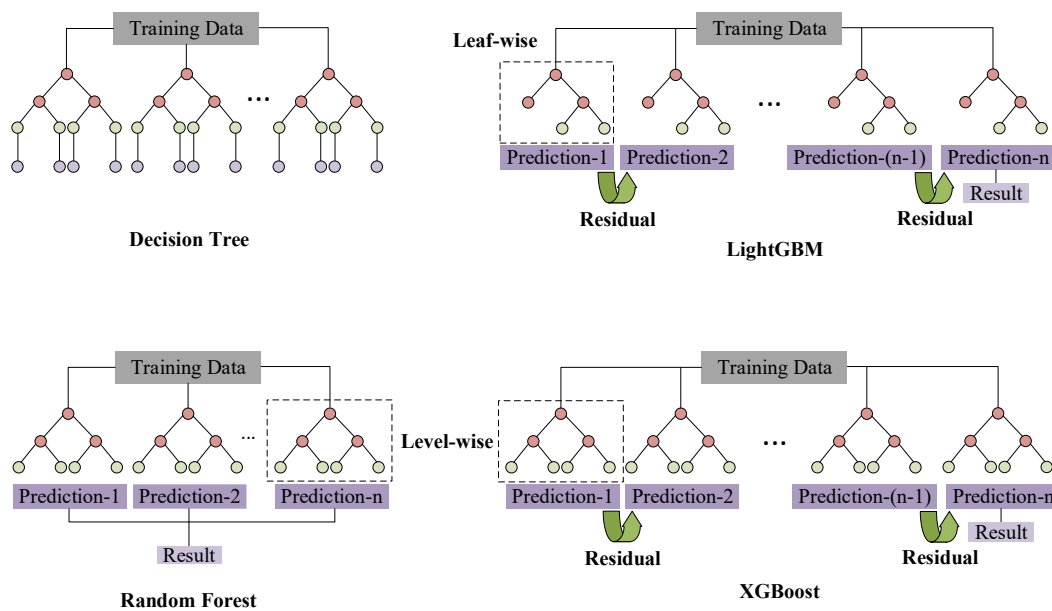


Figure 4. Working mechanisms of Decision Tree, Random Forest, XGBoost and LightGBM algorithms.

Table 2. Model hyperparameters for LightGBM, XGBoost, RF and DT.

Hyperparameter	LightGBM	XGBoost	RF	DT
num_leaves	[20, 50, 100]	-	-	-
max_depth	[3, 5, 10]	[3, 5, 10]	[5, 10, None]	[3, 5, 10]
learning_rate	[0.01, 0.05, 0.1]	[0.01, 0.05, 0.1]	-	-
n_estimators	[100, 300]	[100,300]	[300, 500]	-
min_child_samples	[20, 50, 100]	-	-	-
subsample	[0.6, 0.8, 1.0]	[0.4, 0.6, 0.8]	-	-
colsample_bytree	[0.6, 0.8, 1.0]	[0.6, 0.8, 1.0]	-	-
min_child_weight	-	[1, 5, 10]	-	-
gamma	-	[0.1, 0.2, 0.3]	-	-
min_samples_split	-	-	[2, 5, 10]	[2, 5, 10]
min_samples_leaf	-	-	[1, 2, 4]	[1, 2, 4]
bootstrap	-	-	True, False	-

Performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared (R^2 - Determination Coefficient) were used to evaluate model performance. Mean Squared Error (MSE), calculates the average of the squared differences between the actual and predicted values. RMSE is the square root of MSE and expresses errors in the original data units. MAE sums the absolute values of individual errors and computes their average. R^2 measures how much of the variance in the dependent variable is explained by the model. While $R^2=1$ indicates perfect fit, $R^2=0$ implies that the model has no explanatory power. MSE, RMSE, MAE and R^2 are defined in Equations 1, 2, 3 and 4, respectively. In the equations, n represents the total number of data points, y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and $\bar{y}$ represents the mean of the actual values.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

3. Results and Discussion

The results of daily total energy consumption predictions, using chargingDuration, vehicle_count, mean_connectionTime and median_connectionTime as input features, are summarized in Table 3 and Table 4. Table 3 presents the performance metrics of daily energy consumption predictions obtained with DT, RF, XGBoost, and LightGBM models using the raw dataset. Conversely, Table 4 provides the performance metrics after removing outliers from the input data via the z-score method.

As presented in Table 3 and Table 4, the performance metrics of the models demonstrate minimal differences before and after outlier removal using the z-score method. For instance, the RMSE for

LightGBM decreased only slightly from 39.01 to 37.35, and the MAE dropped marginally from 30.25 to 28.93. Similarly, R^2 values remained nearly unchanged, indicating consistently high explanatory power for all models. These results suggest that the impact of outlier removal on model performance is limited, likely due to the models' inherent ability to handle such data points or the relatively low number of outliers in the dataset.

Table 3. Performance metrics for raw dataset.

Model	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	R-Squared (R^2)
DT	42.46417	32.93693	0.96533
LightGBM	39.00986	30.24578	0.97103
XGBoost	38.89783	30.14571	0.97120
RF	39.36622	30.51473	0.97050

Table 4. Performance metrics for cleaned dataset with z-score.

Model	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	R-Squared (R^2)
DT	40.79563	31.67409	0.96718
LightGBM	37.35035	28.92905	0.97268
XGBoost	37.38780	28.97927	0.97262
RF	37.67076	29.18459	0.97221

Among the evaluated models, LightGBM demonstrated the highest performance, achieving the lowest RMSE of 37.35 and MAE of 28.93, alongside the highest R^2 value of 0.9727. These results indicate superior predictive accuracy and a strong correlation between the predicted and actual values. XGBoost closely followed with an RMSE of 37.39, MAE of 28.98, and an R^2 value of 0.9726, exhibiting slightly reduced precision compared to LightGBM while maintaining robust predictive performance. RF ranked third, with an RMSE of 37.67, MAE of 29.18, and an R^2 value of 0.9722, reflecting relatively less precise predictions than the other two models. The DT model yielded the least favorable results, with an RMSE of 40.80, MAE of 31.67, and an R^2 value of 0.9672, indicating a noticeable drop in prediction accuracy compared to the ensemble-based approaches. Despite these variations, all four models achieved R^2 values exceeding 0.96, highlighting their overall strong predictive capability. Overall, LightGBM emerged as the most effective model for this dataset due to its superior performance across all evaluation metrics.

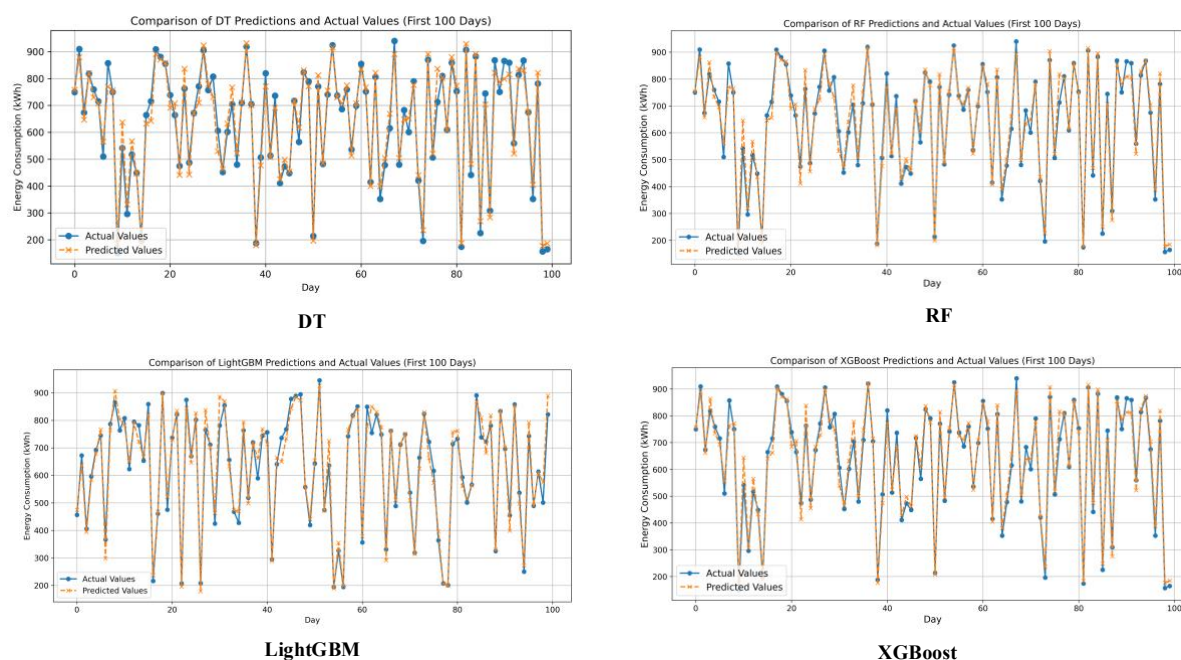
Table 5 summarizes the best hyperparameters for the DT, LightGBM, XGBoost and RF models, obtained through GridSearchCV with 5-fold cross-validation. The selected parameters reflect the models' ability to adapt to the characteristics of the dataset and ensure consistent performance across the evaluation metrics presented in previous tables.

Table 5. Best parameters for the models.

Model	Best Parameters
DT	'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10
LightGBM	'colsample_bytree': 0.8, 'learning_rate': 0.05, 'max_depth': 3, 'min_child_samples': 50, 'n_estimators': 300, 'num_leaves': 20, 'subsample': 0.6
XGBoost	'colsample_bytree': 0.8, 'gamma': 0.1, 'learning_rate': 0.05, 'max_depth': 3, 'min_child_weight': 10, 'n_estimators': 300, 'subsample': 0.6
RF	'bootstrap': True, 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 500

The comparison of actual and predicted daily energy consumption over the first 100 days, as shown in Figure 5, reveals a high degree of alignment across LightGBM, XGBoost, RF and DT. The graphs indicate that the predicted values for each model closely follow the actual values, with minimal visible deviations. These results demonstrate the capability of tree-based models to accurately forecast energy consumption in EV charging stations.

Figure 6 illustrates residual plot for the LightGBM model, visualizing the distribution of residuals relative to the predicted values. In an ideal scenario, residuals should exhibit a random scatter around the zero line, indicating no systematic bias in the model's predictions. In this plot, the residuals are relatively well-distributed around zero for lower and medium predicted values, suggesting consistent model performance in these ranges. However, a noticeable increase in residual dispersion is observed for higher predicted values, indicating that the model tends to make larger errors when predicting higher energy consumption. This pattern suggests that the model might slightly struggle with capturing the variability in high-value predictions, potentially leading to underestimation or overestimation in these cases. Despite this, the overall concentration of residuals near zero supports the accuracy and reliability of the model in most cases.

**Figure 5.** Comparison of models' predictions and actual values.

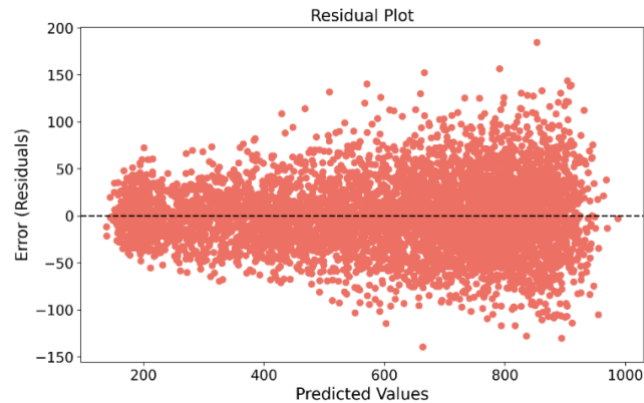


Figure 6. Residual plot for the LightGBM model.

The use of GridSearchCV in this study not only ensured hyperparameter optimization but also contributed significantly to the robustness of the final models. Its exhaustive and systematic nature, combined with cross-validation, allowed for the selection of parameter sets that generalized well across unseen data. This minimized overfitting and increased the credibility of the reported test performance. Given the high dimensionality and volume of the dataset, GridSearchCV proved to be an effective tool that balanced performance with reproducibility two key objectives in applied machine learning studies.

Despite the strengths of the optimization method used, the predictive capacity of the model is inherently constrained by the features available in the dataset. In this study, the model was developed using only basic attributes: vehicle count, connection time, charging duration. The dataset used does not include other potentially influential variables such as environmental or temporal factors (e.g., weather conditions, day information, traffic density). Therefore, the predictive potential of the model is limited to the available features. In future studies, the use of more comprehensive datasets and the integration of such additional variables may contribute to more accurate and dynamic modeling of electric vehicle charging behavior.

Furthermore, it should be noted that the dataset used in this study was obtained solely from the Caltech campus. Therefore, the observed charging behaviors may reflect geographic or demographic characteristics specific to that environment. This may limit the generalizability of the findings to other settings with different user profiles or infrastructure conditions.

4. Conclusion

This study aimed to predict the daily energy consumption of an EV charging station using a public dataset, emphasizing the role of data preprocessing and model selection. LightGBM, XGBoost, RF and DT were evaluated on both the original and Z-score-cleaned datasets. The results demonstrated that all models achieved R^2 values exceeding 0.96, showcasing their robustness in handling both raw and cleaned datasets. After the cleaning process, LightGBM exhibited the best overall performance, while all models maintained high accuracy in predicting energy consumption. In the study, hyperparameter optimization for the models was conducted using GridSearchCV, which played a critical role in realizing the potential performance of the models.

Ethical statement

The author declares that this document requires no ethical approval or special permission.

Conflict of interest

The author declares no conflict of interest.

Generative AI statement

During the preparation of this work the author used ChatGPT in order to improve language and readability, with caution. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- [1] Bebkiewicz, K., Chłopek, Z., Sar, H., Szczepański, K., “Comparison of pollutant emission associated with the operation of passenger cars with internal combustion engines and passenger cars with electric motors”, *International Journal of Energy and Environmental Engineering*, 12, 215-228, 2021.
- [2] Brown, A., “Green technologies and the mobility industry”, *SAE international*, 2010.
- [3] Barman, P., Dutta, L., Bordoloi, S., Kalita, A., Buragohain, P., Bharali, S., Azzopardi, B., “Renewable energy integration with electric vehicle technology: A review of the existing smart charging approaches”, *Renewable and Sustainable Energy Reviews*, 183, 113518, 2023.
- [4] Boglietti, S., Carra, M., Sotgiu, M., Barabino, B., Bonera, M., Maternini, G., “Planning for electric car charging: a review of technologies, criteria and methods”, *Electrifying Mobility: Realising a Sustainable Future for the Car*, 159-187, 2022.
- [5] Subashini, M., Sumathi, V., “Smart algorithms for power prediction in smart EV charging stations”, *Journal of Engineering Research*, 12(2), 124-134, 2024.
- [6] Aduama, P., Zhang, Z., Al-Sumaiti, A.S., “Multi-feature data fusion-based load forecasting of electric vehicle charging stations using a deep learning model”, *Energies*, 16(3), 1309, 2023.
- [7] Bai, X., Wang, Z., Zou, L., Liu, H., Sun, Q., Alsaadi, F.E., “Electric vehicle charging station planning with dynamic prediction of elastic charging demand: A hybrid particle swarm optimization algorithm”, *Complex & Intelligent Systems*, 1-12, 2022.
- [8] Sun, F., Diao, R., Zhou, B., Lan, T., Mao, T., Su, S., ... & Lu, S., “Prediction-based EV-PV coordination strategy for charging stations using reinforcement learning”, *IEEE Transactions on Industry Applications*, 2023.
- [9] Huang, J., Zhou, X., “Optimizing EV Charging Station Placement in New South Wales: A Soft Actor-Critic Reinforcement Learning Approach”, *In 2024 5th International Conference on Computer Engineering and Application (ICCEA)*, 2024, April, pp. 1790-1794, IEEE.
- [10] Zhao, Z., Lee, C.K., Huo, J., “EV charging station deployment on coupled transportation and power distribution networks via reinforcement learning”, *Energy*, 267, 126555, 2023.
- [11] Wang, S., Bi, S., Zhang, Y.A., “Reinforcement learning for real-time pricing and scheduling control in EV charging stations”, *IEEE Transactions on Industrial Informatics*, 17(2), 849-859, 2019.
- [12] Yin, W., Ji, J., “Research on EV charging load forecasting and orderly charging scheduling based on model fusion”, *Energy*, 290, 130126, 2024.

- [13] Kapoor, A., Patel, V.S., Sharma, A., Mohapatra, A., “Optimal Planning of Fast EV Charging Stations in a Coupled Transportation and Electrical Power Distribution Network”, *IEEE Transactions on Automation Science and Engineering*, 2023.
- [14] Lu, Y., Li, Y., Xie, D., Wei, E., Bao, X., Chen, H., Zhong, X., “The application of improved random forest algorithm on the prediction of electric vehicle charging load”, *Energies*, 11(11), 3207, 2018.
- [15] Usman, D., Abdul, K., Asim, D., “A Data-Driven Temporal Charge Profiling of Electric Vehicles”, *Arabian Journal for Science and Engineering*, 48(11), 15195-15206, 2023.
- [16] Gholizadeh, N., Musilek, P., “Daily electric vehicle charging dataset for training reinforcement learning algorithms”, *Data in Brief*, 110587, 2024.
- [17] Lee, Z.J., Li, T., Low, S.H., “ACN-data: Analysis and applications of an open EV charging dataset”, *In Proceedings of the tenth ACM international conference on future energy systems*, 2019, June, pp. 139-149.
- [18] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K., “Modeling tabular data using conditional gan”, *Advances in neural information processing systems*, 32, 2019.
- [19] Gholizadeh, N., Musilek, P., “Electric Vehicle Charging Dataset”, *Mendeley Data*. 2024. doi: 10.17632/5zrtmp7gwd.2.
- [20] Breiman, L., “Random forests”, *Machine learning*, 45, 5-32, 2001.
- [21] Chen, T., Guestrin, C., “Xgboost: A scalable tree boosting system”, *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, August, pp. 785-794.
- [22] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T.Y., “Lightgbm: A highly efficient gradient boosting decision tree”, *Advances in neural information processing systems*, 30, 2017.
- [23] Friedman, J.H., “Greedy function approximation: a gradient boosting machine”, *Annals of statistics*, 1189-1232, 2001.