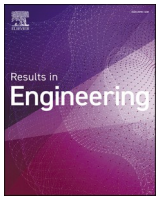
Results in Engineering 24 (2024) 103095

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Electric vehicle charging station demand prediction model deploying data slotting

A.V. Sreekumar , R.R. Lekshmi *

Department of Electrical and Electronics Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India

ARTICLE INFO

Keywords:
Energy demand prediction
Electric vehicle charging stations
Machine learning
Regression models
Variance

ABSTRACT

Accurate prediction of energy requirement at charging station is essential for optimizing infrastructure usage, ensuring grid stability, and minimizing operational cost. Literatures suggest deployment of machine learning techniques to forecast the station demand. One major challenge associated with development of machine learning models is the inherent uncertainty in electric vehicle charging behaviour that includes variations in charging patterns, user preferences, and vehicle types. The conventional pre-processing techniques fail to dislodge nonlinearities and highly random patterns that include very low or zero-charging. Employing such techniques affects the model's forecast accuracy. This article performs data-slotting during pre-processing stage and then selects the best among 1-h, 2-h, 3-h and 4-h slots, to frame the feature vectors. The 4-h data with minimum variance is suggested to frame the dataset. Four distinct datasets, comprising different combination of average and total demands as predictor and response respectively are considered. The created dataset is deployed in Random Forest, Categorical Boosting, Extreme Gradient Boosting and Light Gradient Boosting models. The article recommends Categorical Boosting Regression model with least mean absolute error, mean square error and root mean square error of 0.0726, 0.0112, and 0.1059 respectively. Furthermore, the use of feature vector comprising of aggregated load for prescribed slots and the response representing the aggregated demand is observed to provide the least prediction error by the suggested model. The suggested model fed by the proposed feature vector offers significant advantage to charging station operator by enhancing the operational efficiency while performing resource and cost management with strategic planning.

1. Introduction

Over the past 5 years, there has been a significant hike in the on-road mobile sources. The oil crisis, curbing of emissions related to air pollutants, greenhouse gas and exhaust trap that affects the climate change have spurred for the transition from Gasoline-powered conventional to Electric Vehicles (EVs). Vehicle electrification that addresses energy security concerns has been recognized as a pivoting part in alleviating global climate change and a key aspect to provide sustainable transport [1]. Since 1990s, the research and development of EV has significantly intensified, owing to the political pressure in several countries. The high efficiency, low noise, very low to zero exhaust emissions and flexibility in grid integration and operation are the key characteristics of EVs [2,3]. Nevertheless, the overall air quality benefits can be achieved with EV charging employing clean energy methods, such as wind, solar, and hydro. Integrating energy storage systems and renewable energy sources further enhances the reliability and sustainability of EV charging stations, promoting environmentally friendly practices and bolstering grid resilience. This assures full leverage of the environmental benefits. As the number of EVs on the roads rises, there is a critical need for Electric Vehicle Charging Stations (EVCS). The widespread adoption of EVs relies heavily on the availability of EV charging infrastructure. While EVs offer benefits in reducing fossil fuel dependency and environmental pollution, they present challenges to power systems due to their complex charging behaviors and high demand [4]. This calls for efficient energy management strategies [5] to optimize EV functionality. One commonly utilized approach involves load balancing, wherein charging stations allocate supply among connected vehicles intelligently to alleviate strain on the grid and ensure timely charging for all users. Additionally, the utilization of sophisticated scheduling algorithms helps staggering charging sessions, reducing peak demand, and thereby easing pressure on the electrical grid while potentially leading to cost savings on electricity bills. By implementing these energy management techniques, EVCS can effectively tackle the challenges associated with

* Corresponding author.
E-mail addresses: av_sreekumar@cb.students.amrita.edu (A.V. Sreekumar), rr_lekshmi@cb.amrita.edu (R.R. Lekshmi).

https://doi.org/10.1016/j.rineng.2024.103095
Received 23 August 2024; Received in revised form 3 October 2024; Accepted 7 October 2024
Available online 10 October 2024

Table 1
Summary of key studies in data pre-processing methods and EV charging demand prediction models.

Reference No.	Preprocessing Method	Prediction Method
[9]	• Removal of missing values • Data aggregation • Time feature manipulation	• ARMA • ARIMA • SARIMA
[12]	• Removal of outliers using isolation forest • Transformation and aggregation of weather and traffic data • Feature alignment.	• Random forest • Support vector machines • XGBoost • Deep ANN
[13,14]	• Data aggregation • Feature definition • Data structuring	• LSTM
[15]	• Data cleaning • Event extraction • Data aggregation	• LSTM • ARIMA • MLP
[16]	• Not mentioned	• LSTM and MCS
[17]	• Removal of duplicate and abnormal data • Missing data imputation • Min-max Normalization	• Improved backpropagation neural network
[18]	• Data aggregation • Cleaning • Min-max normalization	• Transformer model • LSTM • RNN • SARIMA • ARIMA
[19]	• Data cleaning • Outlier identification and replacement • Normalization	• CNN-BiLSTM
[20]	• Scaling • Normalization	• GCNN-LSTM • ARIMA • Support vector machines • FNN • CNN
[21]	• Not mentioned	• GA-BPNN • BPNN
[22]	• Down-sampling from hourly to daily energy consumption • Re-sampling • Z-score normalization	• Convolutional LSTM and Bidirectional convolutional LSTM
[23]	• Back casting method to fill missing values • Forward casting method to ensure accuracy	• DNN based on α^2-LSTM
[24]	• Data cleaning and handling missing values • Classification of operating hours into 3 categories • Classification of parking availability into 3 categories • Exclusion of low charging demand	• DL (seq2seq) • ARIMA • Prophet • XGBoost • LSTM
[25]	• Data collection and cleaning • Short data filtering • Data split into non-periodic and periodic pile group stations through Fourier Transform method	• T-LSTM-Enc combined T-LSTM-Ori-Time Features • ARIMA • Vanilla LSTM • LSTM-Enc (proximity) • Temporal LSTM-Enc • T-LSTM-Enc-Time Features • T-LSTM-Ori-Time Features
[26]	• Feature alignment • Handling missing value • Calculation of gap duration between two charging events • Sub-classification of gap duration into two	• GRNN • ANN • RNN • LSTM • BiLSTM • GRU • DNN
[27]	• Data cleaning and splitting • Feature extraction and labeling • Min-max normalization	• ANN • Linear regression • ARIMA • Random forest • Support vector machines • KNN
[28]	• Similar day selection using Gray correlation analysis • Data decomposition based on KL-VMD	• BP • LSSVM • WOA-LSSVM • IWOA-LSSVM • Kullback-Leibler Divergence -Variable Distance Metric -LSSVM • Kullback-Leibler Divergence -Variable Distance Metric and IWOA-LSSVM
[29]	• Data collection • Feature selection • Daily data aggregation • Weekly and monthly data segmentation • Data cleaning	• RNN • LSTM • Bi-LSTM • GRU • CNN • Transformer model
[30]	• The prediction resolution of datasets with huge data fluctuations are modified from the energy consumption per hour to the average energy consumption for three consecutive hours. • Feature engineering techniques to obtain the dataset's single attribute.	• RFAM-GRU • DeepDeff GRU • RNN • DNN • KNN
[31]	• Data collection and cleaning • Normalization • Splitting the data into 24 h slot	• ANN • RNN • Q-learning
[32]	• Data collection • Aggregation of charging demand • Point of interest data integration • Clustering using K means algorithm for region specific prediction • Construction of spatial and temporal graph • Normalization • Handling missing data	• Heterogeneous spatio-temporal graph convolutional network • History Average • Support Vector Regression • Random forest • GRU • Temporal Graph Convolutional Network • Spatio-temporal Graph Neural Controlled Differential Equation

power distribution and contribute to the development of a more sustainable transportation infrastructure. A crucial aspect of this management involves forecasting energy demand, which is vital for maintaining grid stability, optimizing charging infrastructure, and cutting operational costs. Thus, efficient management of energy plays a crucial role in optimizing the performance of EVCS, reducing operational expenses, and maintaining grid stability [6]. At the heart of effective energy management lies load forecasting, which involves predicting the future electricity demand of EVCS. Accurate load forecasting enables proactive resource allocation, implementation of demand response strategies, and strategic infrastructure planning [7]. This calls for the selection of a suitable prediction model that predicts the EVCS load demand. Intensive literature survey has been done related to EVCS demand predictor development to choose effective and efficient forecasting models. Various methodologies, from traditional statistical techniques to advanced Machine Learning (ML) algorithms, are employed for load forecasting [8]. Monte Carlo simulations (MCS) have been used in Ref. [9] to test the power consumption prediction using Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), and Seasonal ARIMA (SARIMA) models. The results show that SARIMA has performed better than the other models in minimizing prediction errors. ML techniques offer a flexible approach to capture complex interactions and leveraging domain knowledge to enhance the interpretability and effectiveness of the models. ML techniques, such as regression analysis, offer robust solutions for forecasting EV charging demand by harnessing historical data and relevant variables [10]. Regression-based approaches have garnered attention due to their

simplicity, efficiency, and effectiveness, particularly in short-term load forecasting [11,12]. These techniques entail modelling the relationship between input variables such as historical charging data, weather conditions, driving patterns and the output variable that include charging demand through regression analysis. Using a mixed-integer nonlinear programming optimization model and a neural network-based charging demand forecasting approach, the study in Ref. [13] produces greater prediction accuracy and performance. Authors in Ref. [14] have employed a Long Short-Term Memory (LSTM) recurrent neural network to forecast short-term load at EVCS. To handle the missing data in the data set and to improve the prediction accuracy, a novel imputation method is also proposed. The suggested model has been proved to outperform conventional methods. Additionally, an LSTM has been introduced in Ref. [15] for short-term prediction of EV charging demand, outperforming ARIMA and Multilayer Perception (MLP) models across different scenarios. In addition, a synergistic learning technique that combines MCS with LSTM neural network is set up to improve the forecast accuracy of EV charging demand in Ref. [16]. An improved backpropagation neural network has been utilized in Ref. [17] for load forecasting using power big data, targeting both station-specific and area-wide charging stations. The research in Ref. [18] utilizes the Transformer model for predicting EV charging demand, based on 7 days, 30 days, and 90 days time stamps to cover both short-term and long-term forecasts. The Transformer model outperforms other models in terms of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), demonstrating its effectiveness in dealing with time series difficulties and EV charging forecasts, as well as its ability to help in efficient energy grid management. Moreover, a hybrid transfer learning approach, incorporating a Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM), has been introduced to enhance EV charging profile forecasting accuracy with limited data in Ref. [19]. This approach has adopted to predict network voltage from EV data without power flow, has proved to exhibits superior accuracy over conventional and new models in predicting charging demand and voltage profiles. Furthermore [20], provides a hybrid model that combines a Graph Convolution Neural Network (GCNN) with a LSTM network to forecasts EV charging demand [21]. presents a Genetic Algorithm-Back Propagation Neural Network (GA-BPNN) based prediction model by combining genetic algorithms with back propagation neural network and optimizing BPNN parameters to increase short term prediction accuracy. For further improvement in energy demand prediction accuracy [22], identifies load-affecting factors using Convolutional LSTM and Bidirectional Convolutional LSTM models. A novel Deep Neural Network (DNN) based on α^2-LSTM to predict EV charging demand over a 15-min time frame is proposed in Ref. [23]. The authors have illustrated the significance of data fusion approach and multi source information integration. Recent studies shows that the efficient demand forecasting ensures enhanced reliability in short term grid operation and long-term infrastructure planning with resource allocation [24]. have proved the superior prediction accuracy of Deep Learning (DL) based Sequence to Sequence model to forecast the monthly charging demand of commercial charging stations over ARIMA, Prophet, XGBoost and LSTM models [25]. focuses on a novel DL based hybrid approach that comprises Temporal Encoder-Decoder + LSTM (T-LSTM-Enc) Concatenated with Temporal LSTM (T-LSTM-Or-i-TimeFeatures) for forecasting the charging demand of EVCS. The proposed method effectively addresses the influence of temporal factors and hidden relation between feature vectors and accurately forecast the energy demand of EVCS.

For accurate load demand prediction, a General Regression Neural Network (GRNN) is proposed in Ref. [26]. With the right inputs and delayed variables, the proposed model has performed better than Artificial Neural Network (ANN), Recurrent Neural Network (RNN), LSTM, Bi-LSTM, Gated Recurrent Unit (GRU), and DNN [27]. Provides a comparative analysis of probabilistic energy forecast models using a comprehensive dataset from Germany. The work proves the better performance of Ada Boosting and Random Forest naive benchmark model [28]. employs an Improved Whale Optimization Algorithm- Least Squares Support Vector Machine (IWOA-LSSVM) model with Improved Variational Mode Decomposition (IVMD) to predict load. The model takes into account the highest temperature, day type, and weather type, and has proved to surpass the performance of conventional ML models. Several ML methods, including RNN, LSTM, Bi-LSTM, GRU, CNN, and transformers, have been employed in Ref. [29] to forecast EV charging loads. Among these, the transformer model exhibits superior performance, showcasing its efficacy in time series forecasting for charging demand. The research detailed in Ref. [30] describes a novel short-term energy consumption prediction approach for electric car charging stations that employs attention feature engineering and a multi-sequence stacking GRU network (RFAM-GRU). The strategy enhances prediction accuracy by 27.2 % over existing algorithms such as RNN, stacked-GRU, and Deep Decomposition Forecasting Framework (DeepDeff) GRU. In order to improve the performance of supervised AI methods, Q-learning-based reinforcement learning is deployed for load forecasting in Ref. [31]. [32] has implemented a heterogeneous spartal-temporal graph convolutional network to analyze and forecast short-term EV charging demand at various resolutions. The model is then employed to control smart grids [33] and intelligent transportation systems. The comparison of literatures that contributes to the data pre-processing methods and development of EV charging demand prediction models is provided in Table 1.

Table 1 shows the summary of key studies deploying various data pre-processing methods and EV charging demand prediction models. From the literature survey it is observed that Random Forest Regressor (RFR), Categorical Boosting Regressor (CBR), Extreme Gradient Boosting Regressor (XGBR) provides a satisfactory prediction while employing conventionally pre-processed data. Furthermore, Light Gradient Boosting Machine Regressor (LGBMR) is proved to provide fast and accurate prediction compared to other Gradient boosting methods. This article thus suggests the development and analysis of RFR, CBR and XGBR and LGBMR for EVCS demand prediction. The explored real time EVCS charging data is associated with high nonlinearities that must be treated in the pre-processing stage. The existing articles prove the significance of various pre-processing methods to ensure accuracy of prediction models. The study shows the widespread use of data cleaning, feature extraction and normalization for pre-processing the data. However, these methods do not effectively take into account the data nonlinearities and its periodic characteristics. In this regard, this article suggests a pre-processing method to deal with the dynamic characteristics of EVCS charging dataset.

1.1. Research gap

Electric vehicle charging data sets are complicated, highly nonlinear, and exhibit zero value at many instants. Model training and validation are hindered by limited historical data availability and inconsistent data collection procedures. The implementation of demand response programs, charging regulations, and grid constraints by utilities and charging operators adds nonlinearities to charging datasets. The inherent unpredictability of EV charging behavior, encompassing variations in charging patterns, user preferences, and vehicle types, poses a notable hurdle. Moreover, vehicle attributes such as battery capacity, charging efficiency, and charging rate, contribute to the complexity. Additionally, the dynamic nature of external factors such as weather conditions, traffic congestion, journey distance, availability of on road charging stations, and grid load further complicates the prediction process. The existing literature has performed data pre-processing in a conventional manner that leaves multiple EVCS entries empty and highly random. Moreover, the procedures performed disturb the periodic characteristics of the data. Deployment of this data results in less accurate ML model. Researchers have suggested many powerful prediction models like RFR, CBR, XGBR and LGBMR that suits many real-

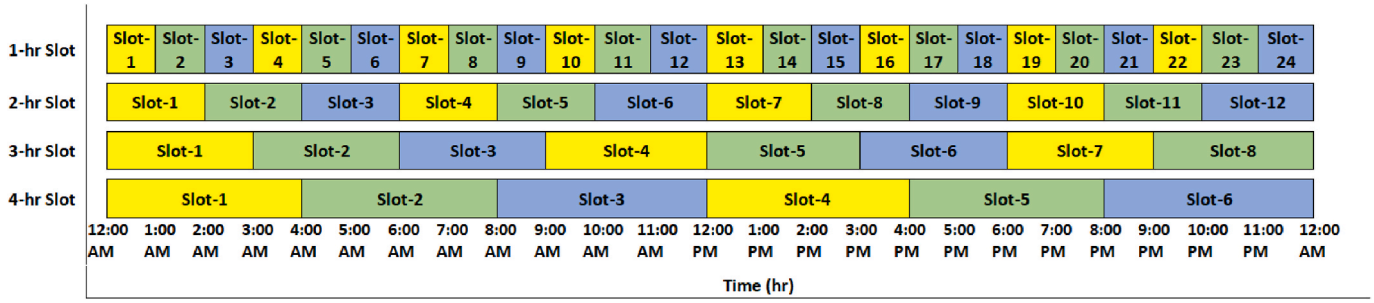


Fig. 1. Slot allocation.

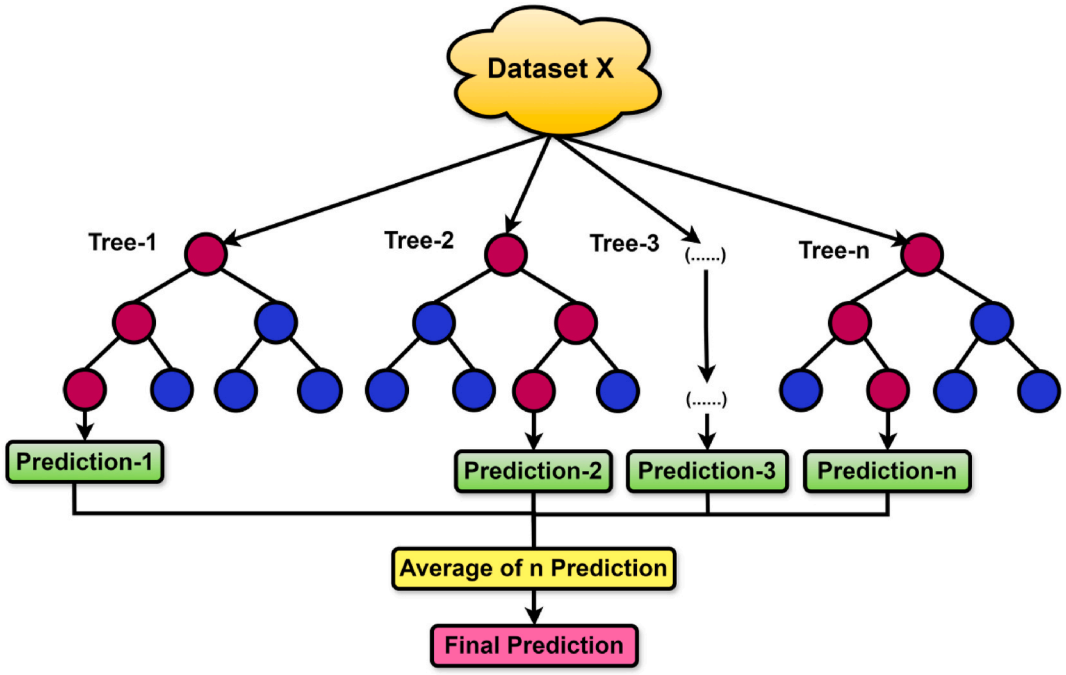


Fig. 2. Flow diagram of Random Forest regression model.

world applications. However, there exists no work that compares the performance of these models in predicting EVCS demand when provided with feature vectors containing slotted data.

1.2. Highlights of this article

Based on the identified research gap, the contributions of this article are highlighted as follows:

Develop a suitable ML model to forecast energy demand of EVCS based on a prescribed feature vector deploying the following procedures.

1. Data collection of EVCS sourced from Perth & Kinross Council's EV charging stations for the period spanning from September 01, 2017 to December 08, 2019.
2. Pre-processing of the collected data including slotting into various time frames to address the data non-linearity and determination of best slot based on the variance.
3. Preparation of 4 distinct following dataset using slotted data from stage 2:
   i) Date, day of the week, slot number, average demand at same slot on previous day, average demand at same slot on two days prior and average demand at previous slot on the previous day are predictor variable. The output variable represents the average energy demand for a prescribed slot and date.
   ii) Predictor variables are the date, day of the week, slot number, total demand at same slot for the previous day, 2 days prior total demand at same slot, previous day previous slot total demand. The output variable indicates the total energy demand for a prescribed time slot and date.
   iii) The date, day of the week, slot number, average demand at the same slot on the previous day, average demand at the same slot for two days ahead, and the average demand at the preceding slot on the previous day as predictor variables. The response represents the total energy demand for a specified slot and date.
   iv) Predictor variables feature the date, day of the week, slot number, total demand at the same slot on the previous day, total demand at the same slot two days prior, and the total demand at the preceding slot on the previous day, whereas the output variable represents the average energy demand for a specified slot and date.
4. Development of ML models such as RFR, CBR, XGBR and LGBMR algorithms deploying various dataset created in 3.
5. Performance analysis of each ML model under different case studies based on the performance indices, MAE, mean square error (MSE) and RMSE.

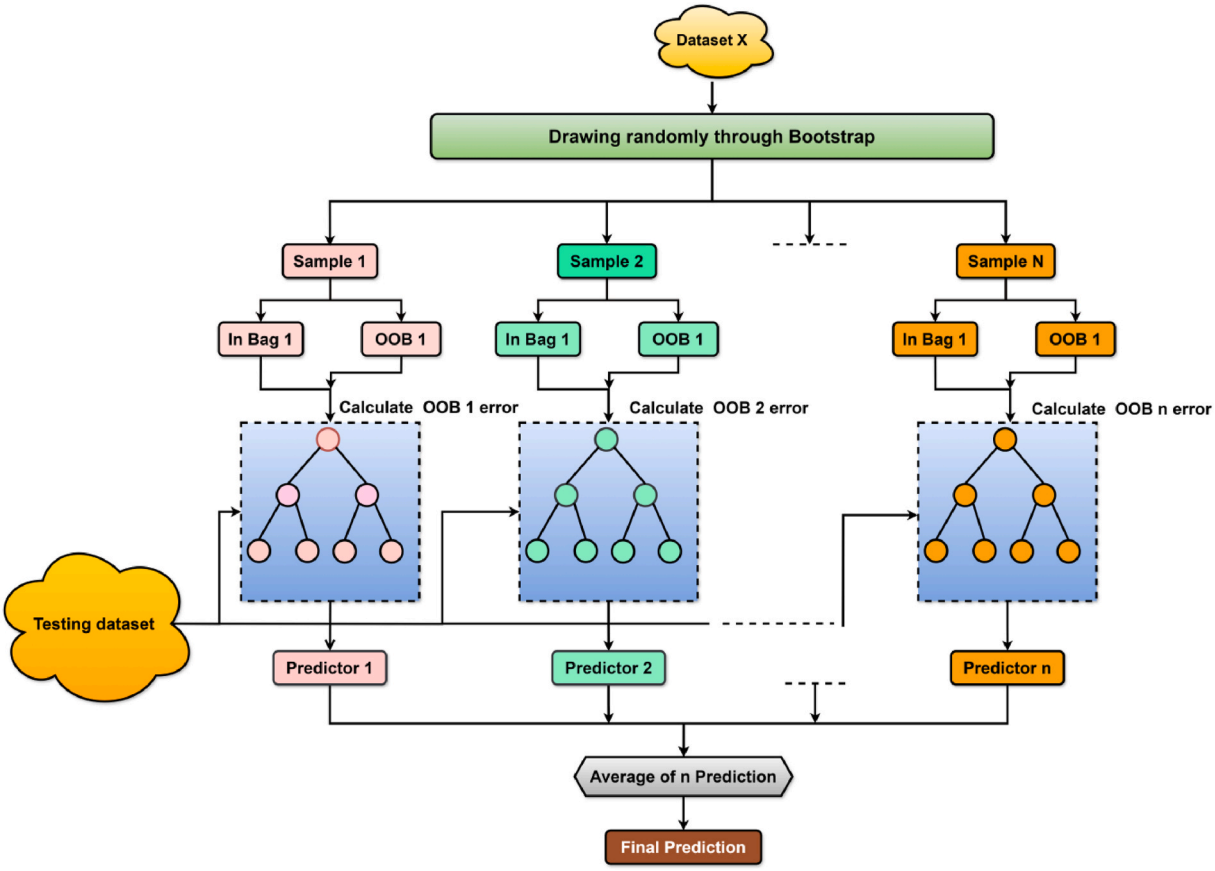


Fig. 3. Flow diagram of CatBoost regression model.

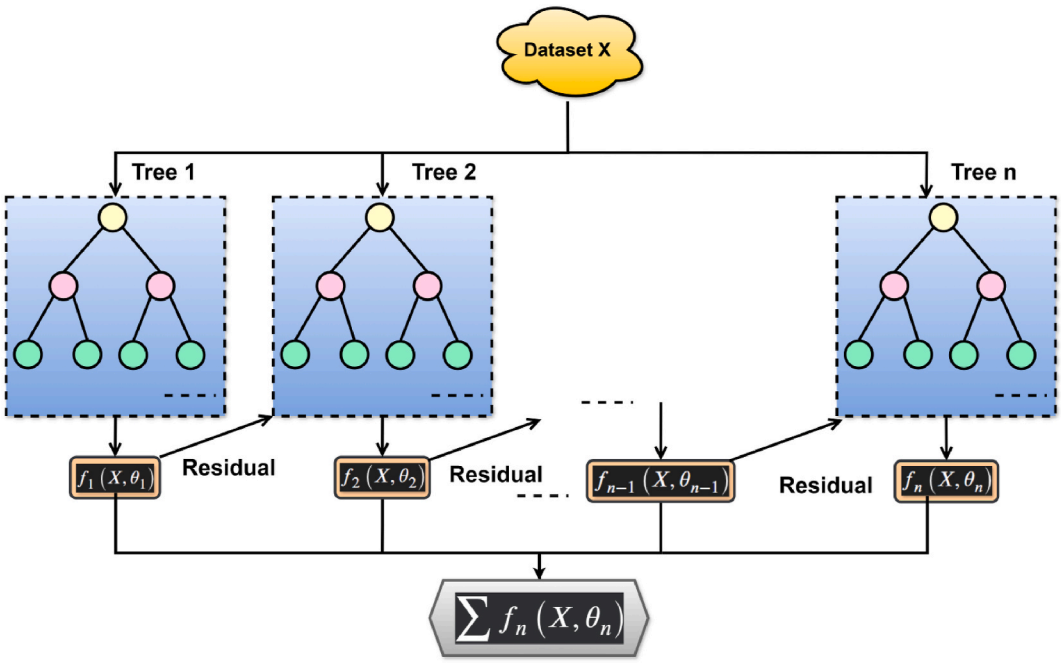


Fig. 4. Flow diagram of XGBoost regression model.

6. Selection of the best EVCS energy forecast model among RFR, CBR, XGBR, and LGBMR algorithms and the most suitable dataset that provides accurate result. The selected model is expected to enhance the operational efficiency of charging station.

2. Dataset Description

This study aims to develop a data-driven energy demand forecasting models of EVCS deploying RFR, CBR, XGBR, and LGBMR algorithms.

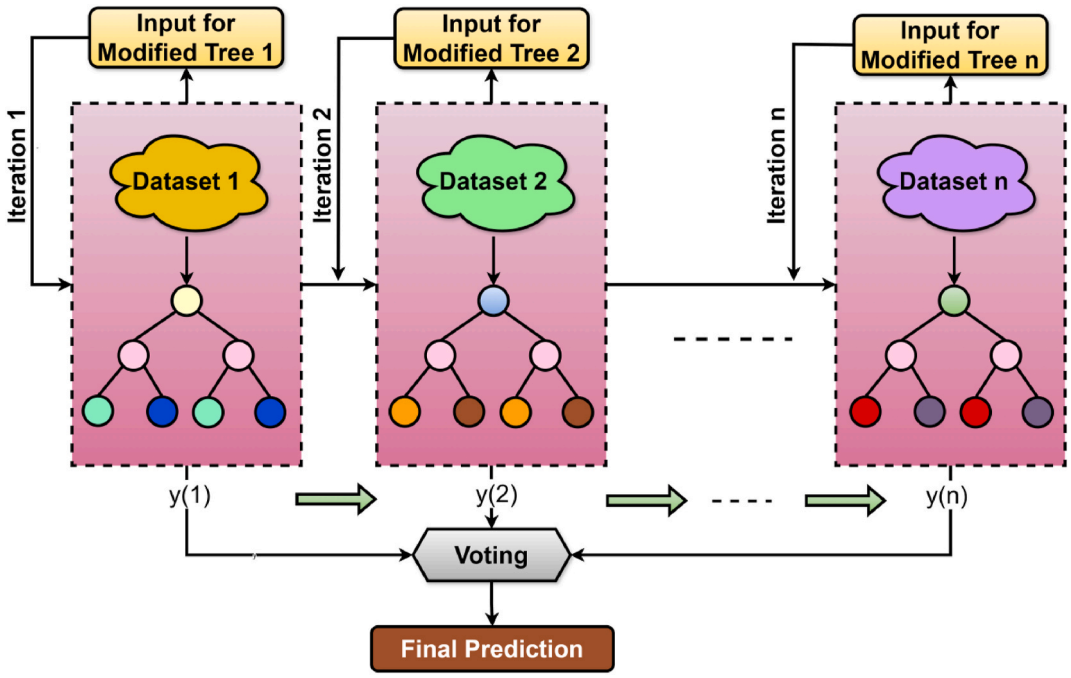


Fig. 5. Flow diagram of Light GBM regression model.

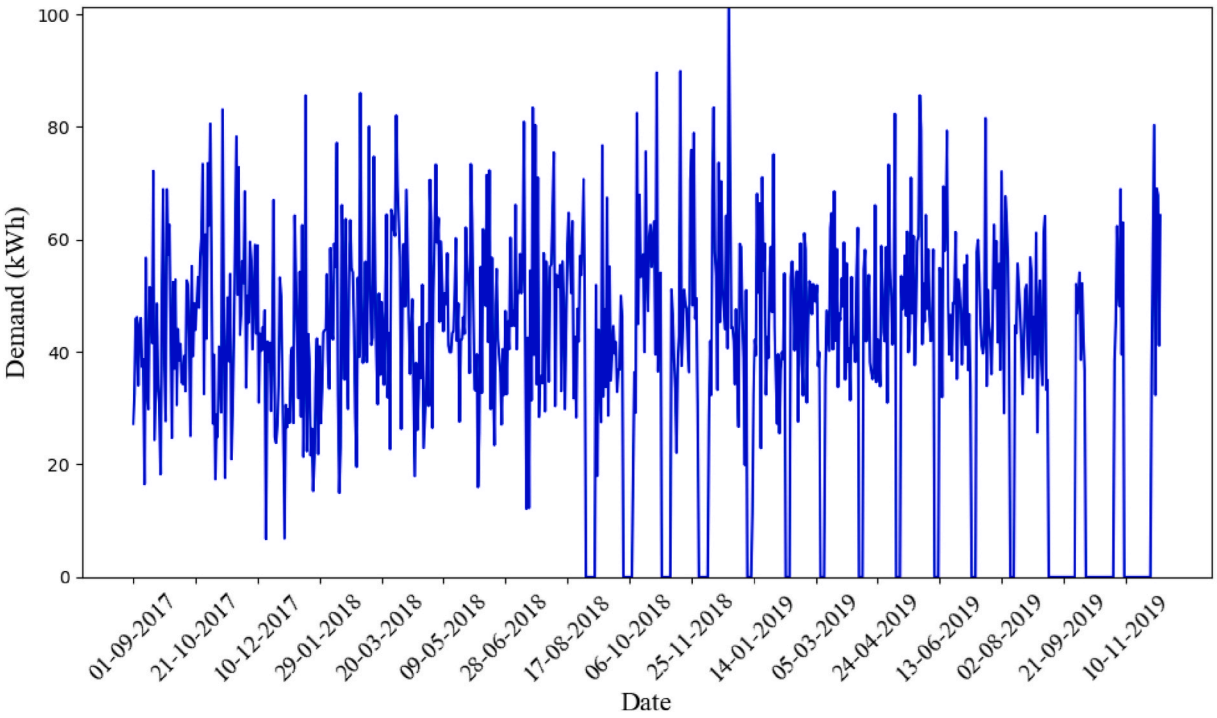


Fig. 6. Daily charging demand of the data set.

The models are developed employing the dataset [34] sourced from Perth & Kinross Council's EVCS, located in Perth, UK for the period spanning from September 01, 2017 to December 08, 2019. The dataset provides details from 14 charging stations that include Kinross Park and Ride Kinross, Atholl Street Car Park Dunkeld, Moness Terrace Car Park Aberfeldy, Mill Street Perth, Leslie Street Car Park Blairgowrie, Broxden Park & Ride Perth, Rie-Achan Road Car Park Pitlochry, South Inch Car Park Perth, Crown Inn Wynd Car Park Auchterarder, Market Square Alyth, Friarton Depot Perth, King Street Car Park Crieff, Friarton Depot Perth and Canal Street Car Park 3rd floor Perth. The dataset comprises of information such as charger identities, session start and end times, energy consumption, charger power, and locations.

An efficient EVCS energy management and planning requires a reliable prediction model that needs an adequate quantity of data to guarantee accurate estimation of energy demand. Hence, the Broxden Park & Ride station, with comparably substantial volume of 11,515 charging event entries is chosen for the model development.

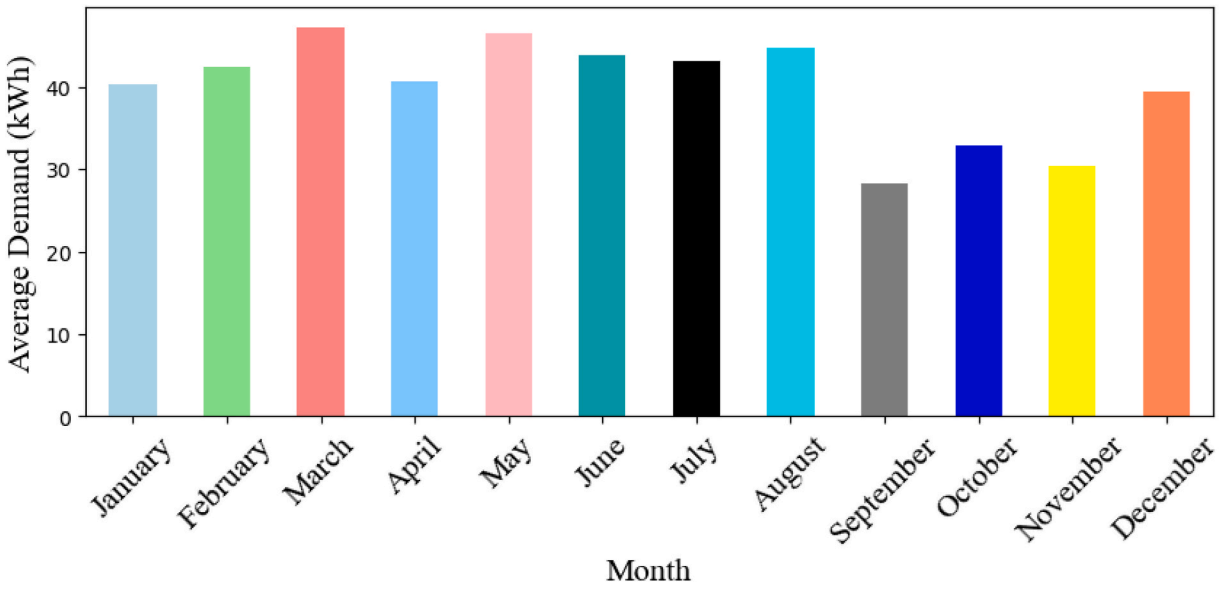


Fig. 7. Monthly average demand of the two-year data set.

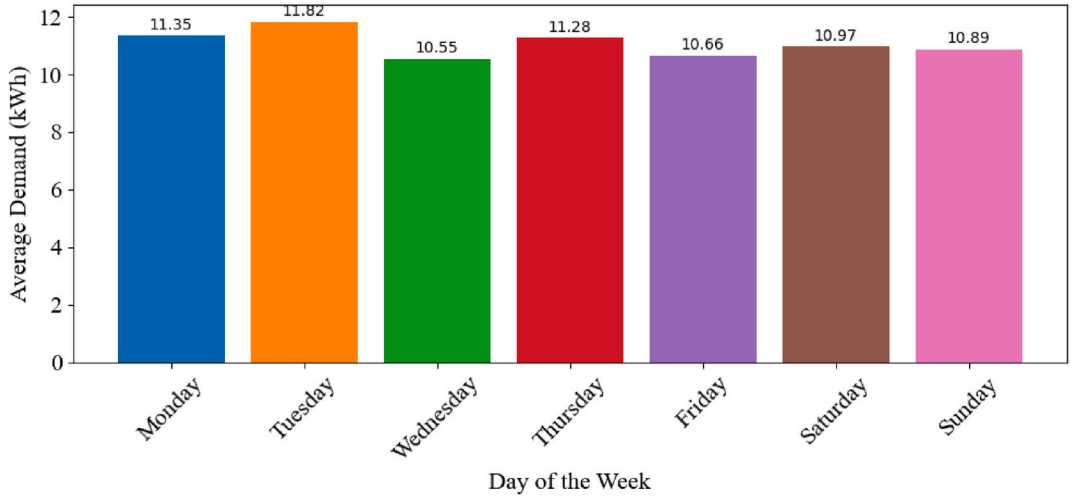


Fig. 8. Average demand of each day in a week.

3. Data pre-processing

The EV charging data distributions are complex and highly random that are influenced by factors such as the EV model, user charging habits, charging station availability, type of charging infrastructure, and battery capacity. Prediction of accurate charging demand is particularly challenging due to the nonlinear nature of demand patterns across different charging stations. Within the realm of preparing data for regression analysis, scholars have explored various techniques aimed at improving the quality and reliability of input data. These methodologies address challenges such as missing data, outliers, and feature scaling, with the goal of enhancing the performance of regression models [35]. Common strategies include feeding missing values, detecting and eliminating outliers, normalizing, or standardizing features for consistency, and selecting pertinent features to reduce dimensionality and improve model interpretability. A multitude of studies have examined the effectiveness of these techniques across different regression contexts, offering valuable insights into their advantages and limitations. In order to provide accurate and reliable regression analysis, careful data preparation is essential. Research in Ref. [36] demonstrates this by analyzing how various data preparation methods affect the performance of regression models. The comprehensive review in Ref. [37] provides practical guidance tailored to regression tasks, assisting researchers and practitioners in effectively navigating the landscape of data preparation. The research contributions in Refs. [38–40] delve into specific aspects of data preparation for regression analysis, enriching the collective understanding of this critical domain.

This paper deploys pre-processing steps to remove erroneous and incomplete entries from the dataset. Subsequently, with less frequent changing events, the data include numerous zero entries. These entries affect the accurate model prediction. This can be dealt with slotting described by equation (1).

$$No.ofslots \in n_h our' slot = \frac{24}{n} \tag{1}$$

As seen in equation (1), an n_hour slotting involves segmentation of the 24-h into 24/n slots. The data is categorized into slots, where the energy demand corresponds to the average of all entries in the prescribed slot period. The paper considers 1-h, 2-h, 3-h and 4-h slots. The period corresponding to each slot for a day is tabulated in Fig. 1.

The 4-h slot in Fig. 1 involves categorizing 24 h duration into 6 time slots, each spanning 4 h from 12 a.m. to 12 p.m. Slot 1 corresponds to the

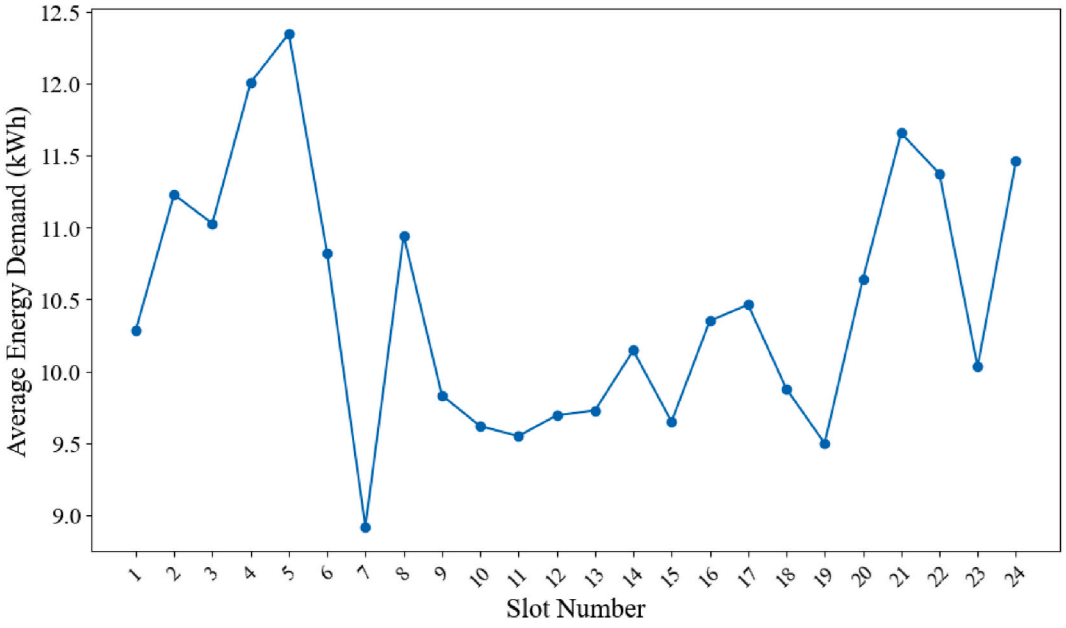


Fig. 9. Average energy demand for 1-h slot.

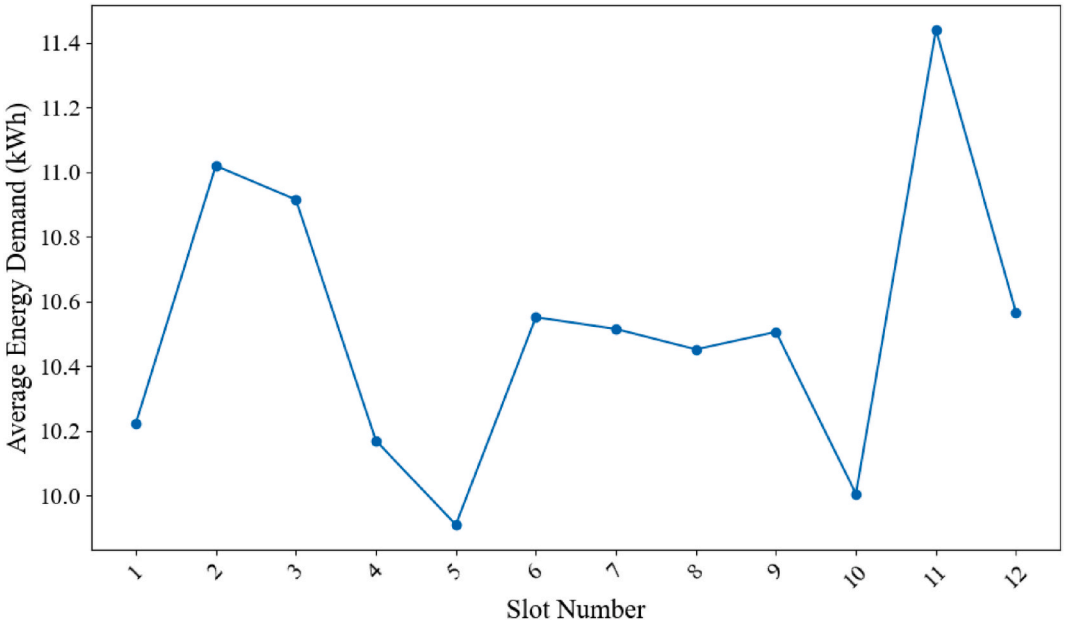


Fig. 10. Average energy demand for 2-h slot.

time from 12 a.m. to 4 a.m., Slot 2 ranges from 4 a.m. to 8 a.m., 8 a.m. to 12 p.m. is the time frame of Slot 3, Slot 4 spans 12 p.m. to 4 p.m. Slot 5 ranges from 4 p.m. to 8 p.m. and 8 p.m. to 12 a.m. is the duration of Slot 6. Correspondingly the variance is calculated to select optimum slotting for energy demand prediction.

4. Importance of variance in data selection for prediction

Variance plays a crucial role in data selection for prediction tasks as it directly affects the predictive models' generalizability and dependability. Variance in statistics and data analysis refers to the distribution or dispersion of data points around the mean. It is represented using equation (2).

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N} \tag{2}$$

Where x_i represents the sample value, $\overline{x}$ is the mean of all sample values, N is the number of samples.

A low variance denotes data points that are closely grouped around the mean, whereas a large variance denotes more dispersion. Comprehending the importance of variance facilitates the evaluation of the representativeness of chosen datasets, guaranteeing that the variability observed in the population is correctly captured. Preventing biased or erroneous predictions is essential, particularly when working with low variance datasets that may not adequately capture population variation. High dataset variance also improves model generalizability by exposing

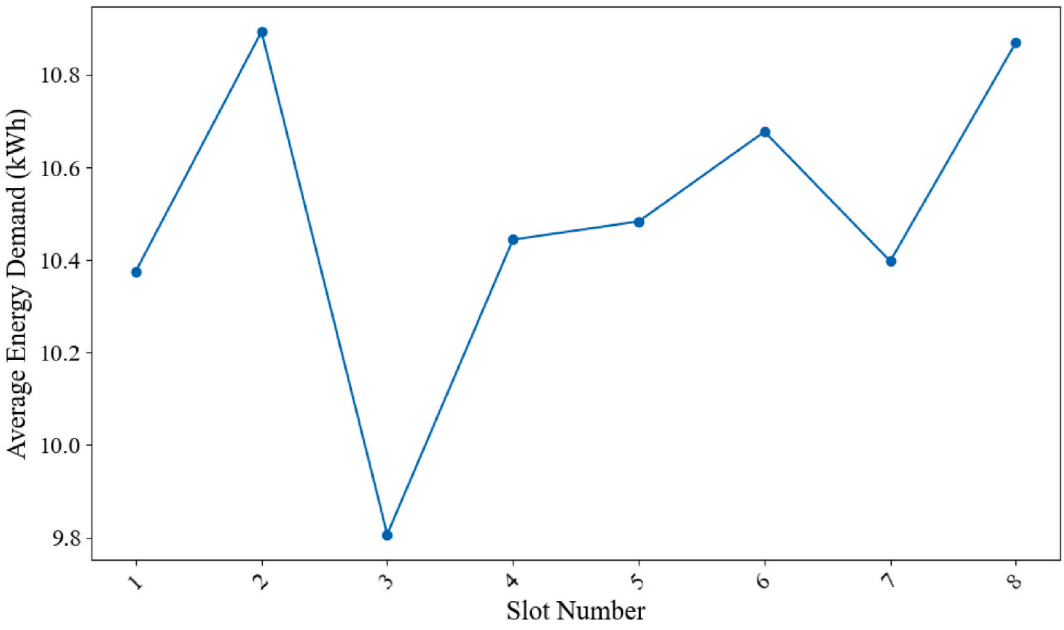


Fig. 11. Average energy demand for 3-h slot.

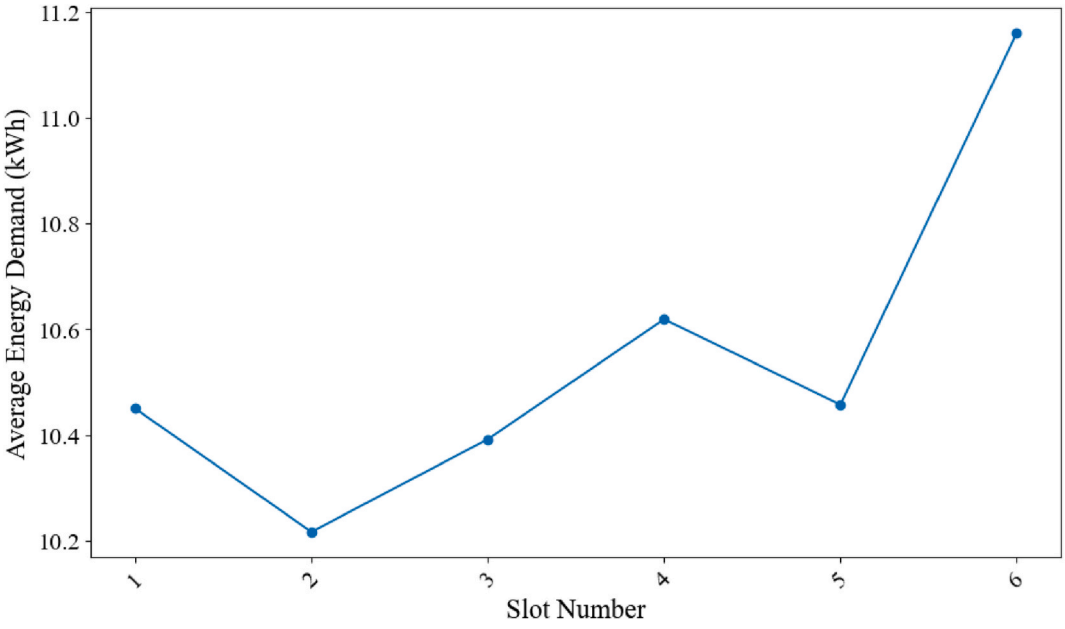


Fig. 12. Average energy demand for 4-h slot.

model to a greater variety of training sets, which helps them generalize effectively to new data. On the other hand, overfitting, in which models capture noise rather than underlying patterns, is another concern associated with large variance. Thus, choosing datasets with a suitable degree of variation aids in reducing overfitting and creating strong prediction models. Variance affects performance evaluation for predictive models as well, which highlights the significance of choosing evaluation datasets with enough variance for assessments that are more accurate. In order to prevent adding noise or bias, it is crucial to strike a balance between higher and lower diversity when choosing datasets. The choice of pertinent characteristics is guided by domain knowledge, which also establishes the allowed degree of variability depending on the complexity of the problem and the intended model performance. Cross-validation methods, such as k-fold cross-validation, help determine possible causes of bias or overfitting as well as evaluate how resistant a model is to perturbations in the data. In summary, developing trustworthy and efficient predictive models for a range of applications requires an awareness of variance and the integration of it into the data selection process. This article computes the variance of entries in each slot for 1-h, 2-h, 3-h and 4-h. The best slot with minimum variance is selected to provide as observations to the ML model.

5. Data driven methods

The paper considers development of RFR, CBR, XGBR, and LGBMR models to predict the EVCS energy demand. Each of these approaches uses unique algorithms and strategies to guarantee a precise solution. RFR involves a large number of decision trees that are trained by ensemble learning to yield the mean forecast for each tree. This method captures the nonlinear interactions between predictors and the target

Table 2
Combination of feature vectors under each case study.

	Case study I	Case study II	Case study III	Case study IV
Feature 1	Date	date	date	date
Feature 2	day of the week	day of the week	day of the week	day of the week
Feature 3	n_d	n_d	n_d	n_d
Feature 4	Average energy demand at n_{d-1}	Sum energy demand at n_{d-1}	Average energy demand at n_{d-1}	Sum energy demand at n_{d-1}
Feature 5	Average energy demand at n_{d-2}	Sum energy demand at n_{d-2}	Average energy demand at n_{d-2}	Sum energy demand at n_{d-2}
Feature 6	Average energy demand at n-1_{d-1}	Sum energy demand at n-1_{d-1}	Average energy demand at n-1_{d-1}	Sum energy demand at n-1_{d-1}
Target	Average energy demand at n_d	Sum energy demand at n_d	Sum energy demand at n_d	Average energy demand at n_d

Table 3
Sample Observation under each case study.

Feature vector type	Input	Output
1	[21-06-2019, Friday, Slot 3, 5.29, 13.64, 27.43]	3.4
2	[11-03-2018, Sunday, Slot 6, 118.92, 18.55, 30.05]	18.42
3	[28-01-2018, Sunday, Slot 5, 10.55, 8.91, 13.82]	39.43
4	[15-10-2018, Monday, Slot 6, 67.89, 19.87, 6.43]	9.56

Table 4
Hyperparameter setting (Case study I).

RFR	
n_estimators	10
random_state	0
oob_score	True
CBR	
iterations	20000
learning_rate	0.09
depth	6
XGBR	
iterations	10000
learning_rate	0.85
depth	0.8
LGBMR	
random_state	1
max_iter	5000

Table 5
Comparison of Performance indices (Case study I).

Sl. No.	Model Name	MAE	MSE	RMSE
1	RFR	0.2193	1.34	1.1576
2	CBR	0.248	1.9017	1.379
3	XGBR	0.1476	0.0448	0.2117
4	LGBMR	0.2782	2.0541	1.4332

variable and performs very well with huge datasets that contain a diverse variety of input feature sets [41]. It works especially well in complicated scenarios where a vehicle's energy consumption is influenced by its kind, the weather, and the time of day. RFR minimizes overfitting and produces accurate forecasts by averaging many decision trees. The decision trees in RFR models are considered standalone that divide data into smaller subsets according to feature values, forecast outcomes based on these divisions. Numerous decision trees are trained using arbitrary features and data subsets in order to predict the load forecast. The results are then merged to provide an output. They provide feature significance ratings and improve prediction accuracy by lowering bias and variance after being trained on random selections of features and data. The model combines tree predictions to provide a final load projection. The flow diagram of RFR model is shown in Fig. 2.

Fig. 2 illustrates the construction of multiple decision trees such as Tree 1, Tree 2, and Tree n for a dataset. The pre-processed and normalized energy demand dataset of EVCS is used to train each tree, resulting in a variety of independent predictions, including Prediction-1, Prediction-2, and Prediction-n. The final projection is then calculated by averaging these individual forecasts. This procedure improves the final prediction's accuracy and resilience by reducing overfitting when compared to individual decision trees.

CBR leverages gradient boosting to effectively handle categorical variables with little pre-processing. CBR is able to estimate energy use correctly by using creative oblivious trees in conjunction with ordered boosting, especially in cases when there are a lot of categorical factors impacting the consumption. Its resilience to overfitting and automated handling of missing data and categorical variables provides accurate predictions with little adjustment of the hyperparameters [42]. The process flow CBR model is demonstrated in Fig. 3.

Fig. 3 illustrates the process of random multiple bootstrap samples through the bootstrap method for the energy dataset. Each sample is then split into in-bag data to train the individual trees and out of box (OOB) data to validate and calculate OOB errors for each tree as shown Fig. 3. This process helps to avoid over fitting. Subsequently, the testing dataset is predicted by each of the individual trained trees, and the final prediction is generated by averaging their predictions. High accuracy, robustness, and user-friendliness are guaranteed by CatBoost's special order-aware boosting and effective categorical feature management [43].

Regularized model formalization is used by the well-known and quick XGBR model to avoid overfitting. Due to the high degree of flexibility, predictions are optimized through hyperparameter adjustments [44,45]. Because of its ability to handle large datasets and capture intricate correlations, XGBR is a good choice for predicting energy demands in EVCS. This is further supported by skills like feature significance analysis and management of missing values. Fig. 4 displays XGBR model process flow.

Fig. 4 illustrates XGBR model process flow that receives various input features of EVCS charging demand and minimizes prediction error by building an ensemble of trees (Tree 1 to Tree n) consecutively. XGBR employs a sequence of decision trees, each constructed using a subset of attributes and data. The first tree (Tree 1) is trained to predict the output values starting with the energy demand dataset, which is represented as dataset X. This produces a set of residuals. The residual serves as the objective for Tree 2, the subsequent tree that is taught to fix the errors caused by Tree 1. Each successive tree is trained to anticipate the result. $f_1(X,\theta_1)$, $f_2(X,\theta_2)$, ….., $f_n(X,\theta_n)$ in Fig. 4 represents predictions of each tree in the sequence, where θ stands for the parameter of each tree. By lowering the error from the preceding tree, each tree enhances the forecast. The total of all the trees' forecasts makes up the final forecast. The iterative residual correction method guarantees strong prediction performance for XGBR.

With the aims to optimize the speed and efficiency, LGBMR employs leaf-wise data partitioning and gradient-based tree building [46]. This architecture is perfect for accurate energy consumption calculations in EV charging scenarios since it performs incredibly well with large datasets and high-dimensional feature spaces. A variety of regularization and complexity management strategies help users improve the model's performance while retaining its capacity to identify intricate patterns and connections. Fig. 5 displays the process flow in association with LGBMR model.

The process of LGBMR shown in Fig. 5 starts with data input. The decision trees (Tree 1 to Tree n) are then literally trained by the model. Each tree picks up the knowledge from the residuals of the preceding

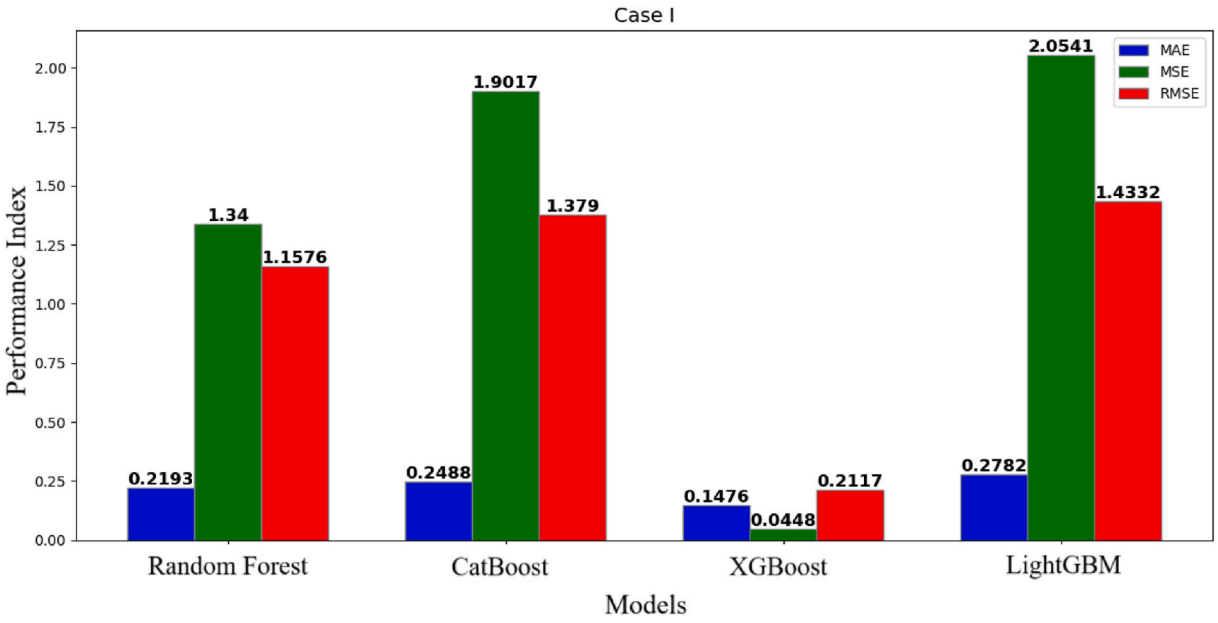


Fig. 13. Performance comparison of case study I.

Table 6
Hyperparameter setting (Case study II).

RFR	
n_estimators	50
random_state	1
oob_score	True
CBR	
Default values	
XGBR	
Default values	
LGBM	
Default values	

Table 7
Comparison of Performance indices (Case study II).

Sl. No.	Model Name	MAE	MSE	RMSE
1	RFR	0.0788	0.0125	0.1118
2	CBR	0.0726	0.0112	0.1059
3	XGBR	0.0942	0.0186	0.1365
4	LGBMR	0.0797	0.0131	0.1145

tree until it eventually outputs the anticipated values through a voting mechanism. This mechanism involves multiple trees votes to determine the final prediction. This voting process improves the model's accuracy and robustness, allowing LGBMR to produce accurate and interpretable results.

By utilizing several decision trees to improve accuracy and resilience, ensemble learning models like RFR, CBR, XGBR and LGBMR represent significant advances in predictive modeling. By averaging the outputs of various trees trained on various data subsets RFR increases the dependability of predictions. With minimum pre-processing CBR achieves great accuracy and user friendliness by introducing order-aware boosting and efficient handling of categorical data. Gradient boosting is optimized by XGBR which iteratively corrects residuals to reduce prediction errors and improve performance. Similarly, LGBMR emphasizes efficiency and speed in streamlining this process which makes it especially well suited for big datasets. They perform particularly well with complicated datasets while minimizing over fitting, enhancing accuracy, and guaranteeing efficient computation. Among different prediction model, the best model can be selected by analyzing its performance considering various indices, with unseen data.

6. Performance indices

In the realm of ML model-based prediction, assessing model performance using performance indices is of paramount importance. These metrics serve as tools for evaluating the accuracy, reliability, and generalization capabilities of ML models across diverse domains. Several performance indicators have been utilized in this research in order to assess the precision and dependability of the demand forecasting model for EVCS. This article deploys MAE, MSE, and RMSE to assess the model prediction accuracy.

6.1. Mean absolute error

The simplest approach to quantify average prediction error in EVCS demand forecasting models is MAE. It is a measure of the average magnitude of the errors in a set of predictions, without considering their direction. With respect to outlier resistance and interpretability, MAE provides a simple way to quantify the error between the actual and predicted values. MAE is computed using equation (3).

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

Where,

n is the number of observations

y_i is the actual value.

$\hat{y}_i$ is the predicted value

MAE shows how close the model's predictions are with respect to the actual results. A lower MAE indicates a model with better fit to the data, which implies more closeness of predictions and the actual values.

6.2. Mean squared error

MSE is another index deployed to assess the precision and dependability of various forecasting models in the context of predicting the demand for EV charging facilities. The average of the squares of the errors is measured by the MSE. MSE is helpful in determining the variance of the errors and the average deviation between the forecasts and

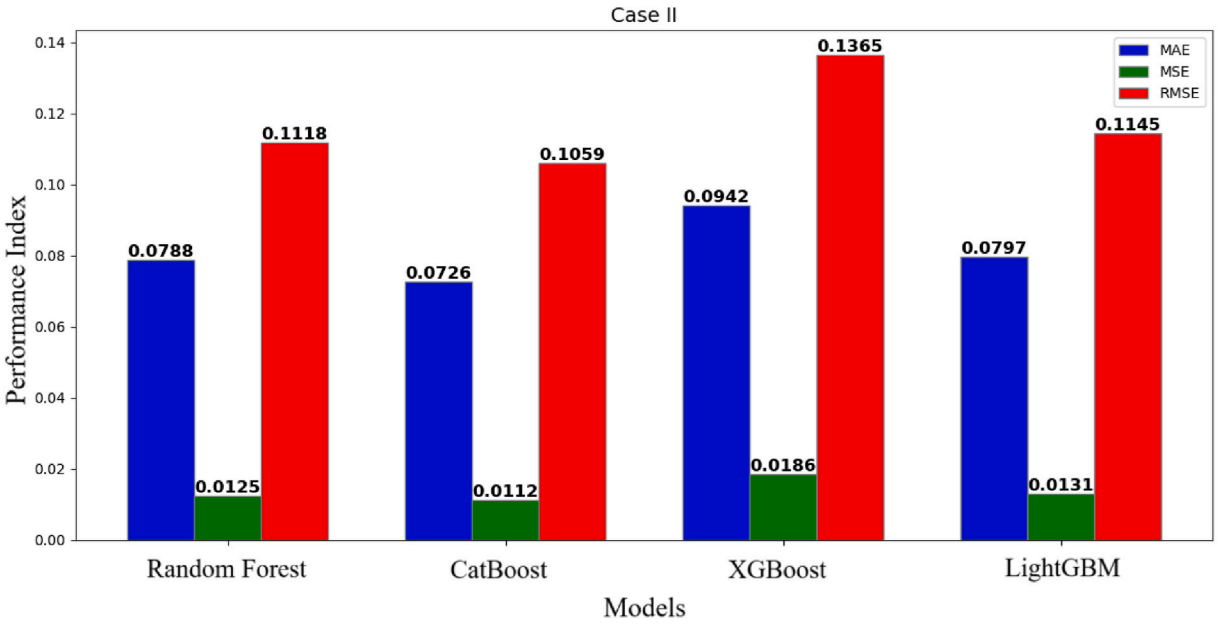


Fig. 14. Performance comparison of case study II.

Table 8
Hyperparameter setting (Case study III).

RFR	
n_estimators	10
random_state	1
oob_score	True
CBR	
iterations	2000
learning_rate	0.09
depth	0.78
XGBR	
iterations	10000
learning_rate	0.85
depth	0.8
LGBMR	
Default values	

Table 9
Comparison of Performance indices (Case study-III).

Sl. No.	Model Name	MAE	MSE	RMSE
1	RFR	0.0953	0.0176	0.1327
2	CBR	0.092	0.0166	0.129
3	XGBR	0.1136	0.0244	0.1561
4	LGBMR	0.0878	0.0155	0.1246

actual values. The MSE is computed using equation (4).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y|i - \widehat{y}_i)^2 \tag{4}$$

A model that fits the observed data more closely is indicated by a lower MSE. An optimal system holds least MSE value.

6.3. Root mean squared error

RMSE is the square root of MSE and provides an error metric in the same units as the original data, making it more interpretable in practical terms. RMSE is better appropriate for situations where accurate estimate of greater mistakes is crucial since it penalizes larger prediction errors more harshly. RMSE is defined using equation (5).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y|i - \widehat{y}_i)^2} \tag{5}$$

Least value of RMSE corresponds to the model that is more closely fitted to the data, resulting in predictions that are more accurate.

By leveraging these performance indices, different forecasting models can be analyzed for its performance, thereby facilitating the best model to enhance the operation and planning of EV charging infrastructure. This rigorous evaluation ensures that the forecasting model meets the necessary standards of accuracy and reliability required for practical implementation.

7. Results and discussion

This paper focuses on the design and development of ML model to predict the EVCS energy demand for a specified slot. The different models described in section 5 are analyzed for its performance and the best model is suggested. The data analysis and the models' development are done using Python under Visual studio code editor, version 1.93. The quality of ML model can be ensured with the help of large data set with diverse number of samples. The paper considers the Broxden Park & Ride station dataset sourced from Perth & Kinross Council's EVCS for the period spanning from September 2017 to August 2019, as mentioned in section 2.

The total energy demand per day for the period spanning from September 2017 to December 2019 is depicted in Fig. 6.

The fluctuating curve shown in Fig. 6 portrays the variations in daily demand, reaching a peak of 101.35 kWh on December 25, 2018. The plot shows extreme variability in demand over the given period. Several instants show zero or very low energy demand, making it very difficult to formulate a clear trend from this particular energy demand curve. A pictorial representation of the monthly average charging demand over a two-year is shown in Fig. 7.

Fig. 7 shows comparatively high energy demand for the month of March and May. Meanwhile, the consumption is found to be low during the month of September. The average daily demand bar chart representation corresponding to the fourth week of May 2019 is shown in Fig. 8.

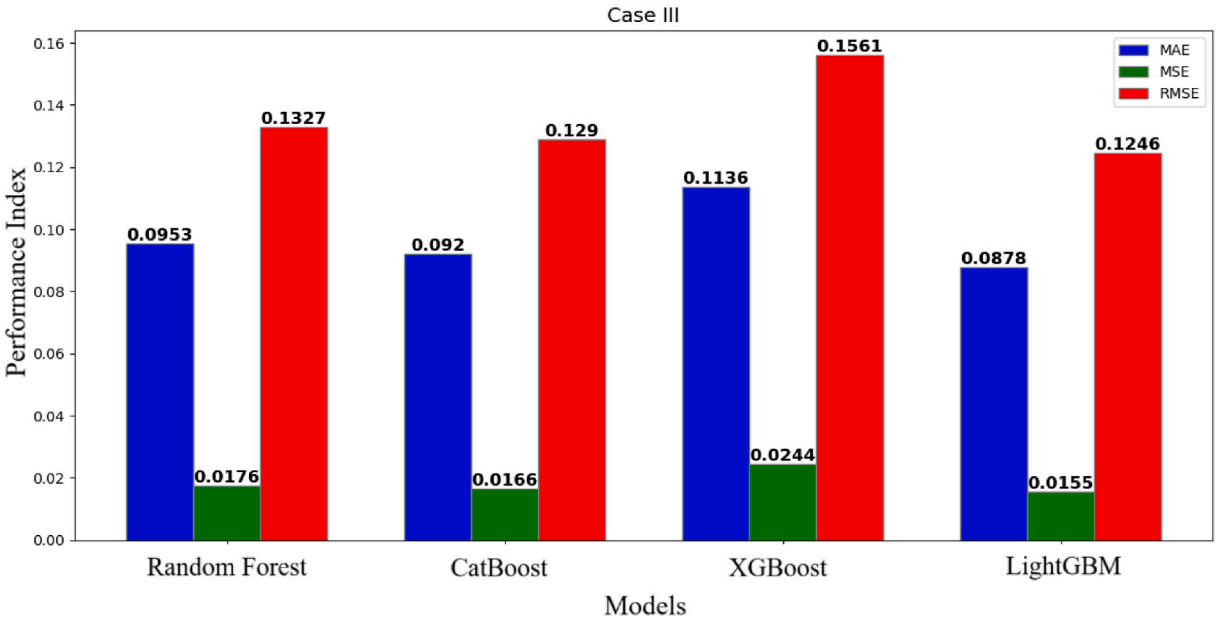


Fig. 15. Performance comparison of case study III.

Table 10
Hyperparameter setting (Case study IV).

RFR	
n_estimators	100
random_state	1
oob_score	True
CBR	
iterations	2100
learning_rate	0.09
depth	0.8
XGBR	
iterations	11000
learning_rate	0.85
depth	0.8
LGBMR	
Default values	

Table 11
Comparison of Performance indices (Case study IV).

Sl. No.	Model Name	MAE	MSE	RMSE
1	RFR	0.1004	0.0173	0.1314
2	CBR	0.1031	0.0178	0.1334
3	XGBR	0.1212	0.0236	0.1536
4	LGBMR	0.0965	0.016	0.1266

Fig. 8 shows that the average energy demand per day in the 4th week of May 2019 that ranges from 10.55 kWh to 11.82 kWh. The customer plug-in behaviour, travel pattern, access to infrastructure and preferences, battery capacity and on road vehicles decides the frequency of EV charging. The charging pattern is observed to be highly wobbled. In this regard long term prediction of EVCS energy demand is highly challenging. Hence this article employs short term prediction of energy demand. With the facility to plug-in at home or worksite, it is often found that EVCS are left unutilized. This results in highly dynamic nature in charging data set. Employing such type of dataset while developing the ML model can cause a significant impact on its accuracy. Thus, the available data is slotted into 1-h, 2-h, 3-h, and 4-h slots. The average demand for 1-h, 2-h, 3-h, and 4-h slots are depicted in Figs. 9–11 and 12 respectively.

Fig. 9 represents the average energy consumption (kWh) at different slots, where x-axis represents the slot number from 1 to 24 and y-axis represents the average energy consumed. Interestingly, the energy demand reaches a peak of 12.35 kWh at slot 4 before falling sharply down to 8.92 kWh at slot 7. The power consumption varies considerably between slots 10 and 20, with numerous crests and troughs. A sudden spike happens at around slot 21 to reach 11.5 kWh before falling again toward slot 23. This shows high nonlinearity of energy demand at various slots. The average energy demand is sketched for 2-h slot and is depicted in Fig. 10.

Fig. 10 shows average energy consumption in kWh per slot across 12 slots with each slot having 2-h duration. Slot 2 and slot 11 exhibit peak consumption, reaching around 11.02 kWh and 11.44 kWh respectively. The lowest energy consumption occurs at slot 5, where the demand usage drops to 9.91 kWh. The curve follows a trend with moderate rise in consumption between slot 6 and slot 9. The obtained demand profile shows a decreased volatility in energy usage across different slots. The average energy consumption over 3-h slot is displayed in Fig. 11.

It is observed from Fig. 11 that slots 2 and 8 are occupied with highest energy demand of 10.89 kWh. A steep decline in the profile to 9.81 kWh at slot 3, followed by a gradually escalated consumption at slot 6. From slot 4 to slot 6 the energy demand is observed to be almost uniform. This variability suggests dynamic charging patterns within this time frame. The average energy demand of 4-h slot is featured in Fig. 12.

It is clear from Fig. 12 that the peak consumption of 11.16 kWh is observed at slot 6 and the least consumption of 10.21 kWh occurs at slot 2. Slots 1, 3, 4 and 5 exhibit considerably similar consumption pattern. Among the four consumption curves depicted from Figs. 9–12, the 4-h slot demonstrates the least variation in average energy consumption. Due to the inherently nonlinear nature of electrical car charging station data, prediction algorithms cannot guarantee 100 % accuracy. This article takes into account the dataset with less non-linearity. This is achieved by computing the variance of the data set using equation (2). The variance against 1-h, 2-h, 3-h, and 4-h slots are found to be 0.7905, 0.1917, 0.1189 and 0.1064 respectively. The 4-h slot data provides the comparative minimum variance of 0.1064 and hence is chosen to develop ML model to predict EVCS energy demand. The features available in the dataset are required to be considered with utmost importance as it provides an impact on the prediction accuracy of ML model. The

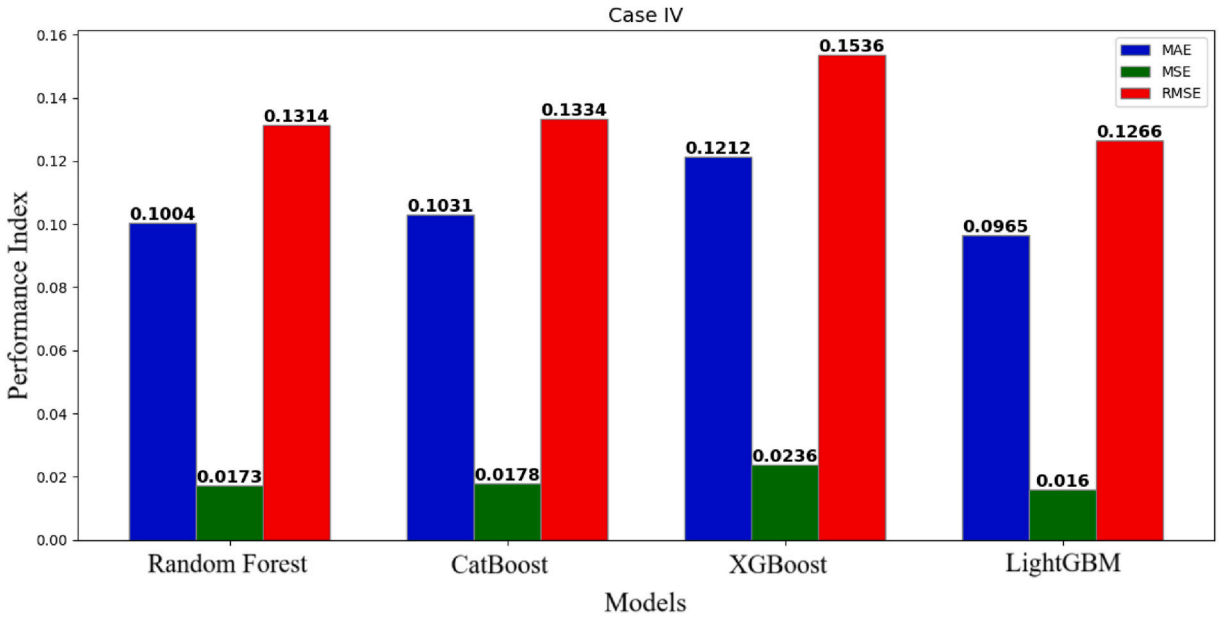


Fig. 16. Performance comparison of case study IV.

article considers 4 different combinations of features to develop the prediction model. To predict demand at slot number 'n' on day 'd', the model development is done with multiple observations having features that include date, day of the week, n_d, average or total demand at n_{d-1}, average or total demand at n_{d-2}, average or total demand at n-1_{d-1}. The article considers the combinations of features as Tabulated in Table 2.

Table 2 show the consideration of past average energy demand for case study I and III, while case study II and IV considers past total energy demand as one of the input features. Meantime, case study I and IV considers average energy demand as the response, while case study II and III considers total of energy demand as the response. A sample of observation from each scenario is presented in Table 3.

This dataset is normalized as mentioned in section 3. The data is then split into train and test categories in the ratio of 80:20. Each combination of feature is then applied to the ML model to perform train and test as follows:

A total dataset of 11,515 charging events are utilized for this work spanning from September 01, 2017 to December 08, 2019. The raw data are pre-processed and normalized, the required dataset with considered features are created. This is followed by splitting dataset into training and testing set. Post training, the RFR, CBR, XGBR and LGBMR models are tested. The accuracy of the prediction is then analyzed employing the test data in terms of MAE, MSE and RMSE. Finally, the predicted value versus actual value curve is plotted under each ML model for each feature combinations for better understanding of the effectiveness of the prediction. The following sections deals with the RFR, CBR, XGBR and LGBMR model development for various combinations of features as mentioned in Table 2.

7.1. Case study I

The raw data is pre-processed to represent daily average energy demand, is segmented into six 4-h slots. Predictor variables include date, day of the week, slot number, average demand at same slot on previous day, average demand at same slot on two days prior, and average demand at previous slot on the previous day. The output variable represents the average energy demand for a prescribed slot and date. The train dataset is provided to the RFR model having the Hyperparameter setting as furnished in Table 4.

The performance of the model is then analyzed employing the test data in terms of MAE, MSE and RMSE. The values are observed to be 0.2193, 1.340, and 1.1576 respectively. Later, CBR model is provided with train dataset while the model is defined with hyperparameter settings listed in Table 4 and the model is trained for 2000 iterations. Post completion of training, the model is subjected to 20 % test data. The obtained performance metrics are MAE of 0.248, MSE of 1.9017, and RMSE of 1.379. In order to maximize the accuracy and efficiency, the hyperparameter values of the XGBR model are fine-tuned. The model is trained for a predetermined number of rounds of boosting with learning rate and depth as presented in Table 4. The performance metrics that are obtained include RMSE of 0.2117, MSE of 0.0448, and MAE of 0.1476. The hyperparameter settings of LGBMR are detailed in Table 4. A predetermined number of boosting iterations are used to train the model. This is followed by the use of the remaining 20 % of the dataset for testing. The achieved performance metrics include RMSE value of 1.4332, MSE of 2.0541, and MAE of 0.2782. The observed performance indices of RFR, CBR, XGBR and LGBMR are summarized in Table 5.

Table 5 presents the performance metrics (MAE, MSE, and RMSE) for four distinct regression models that include RFR, CBR, XGBR, and LGBMR. The results indicate that RFR achieves MAE, MSE, and RMSE values of 0.2193, 1.34, and 1.1576, respectively. Meanwhile, CBR exhibited slightly higher errors compared to RFR, with MAE, MSE, and RMSE values of 0.248, 1.9017, and 1.3790, respectively. XGBR outputs the most favorable performance among the models, with MAE, MSE, and RMSE values of 0.1476, 0.0448, and 0.2117, respectively. Conversely, LGBMR deliver the least accurate prediction, with MAE, MSE, and RMSE values of 0.2782, 2.0541, and 1.4332, respectively. The obtained indices are represented as bar chart in Fig. 13.

As depicted in Fig. 13, the metrics corresponding to RFR indicate that the model performs adequately in predicting the target variable with moderate accuracy. The metrics obtained from CBR indicate reasonable predictive performance that is comparable to RFR. The measures shown by XGBR display excellent relatively better accuracy, which suggests that the fundamental patterns in the data are effectively captured, as visible in Fig. 13. The comparison highlights XGBR superior predictive performance in this context for the feature vector type 1.

7.2. Case study II

In this case study, the considered models are provided with predictor variables that include the date, day of the week, slot number, total demand at same slot for the previous day, 2 days prior total demand at

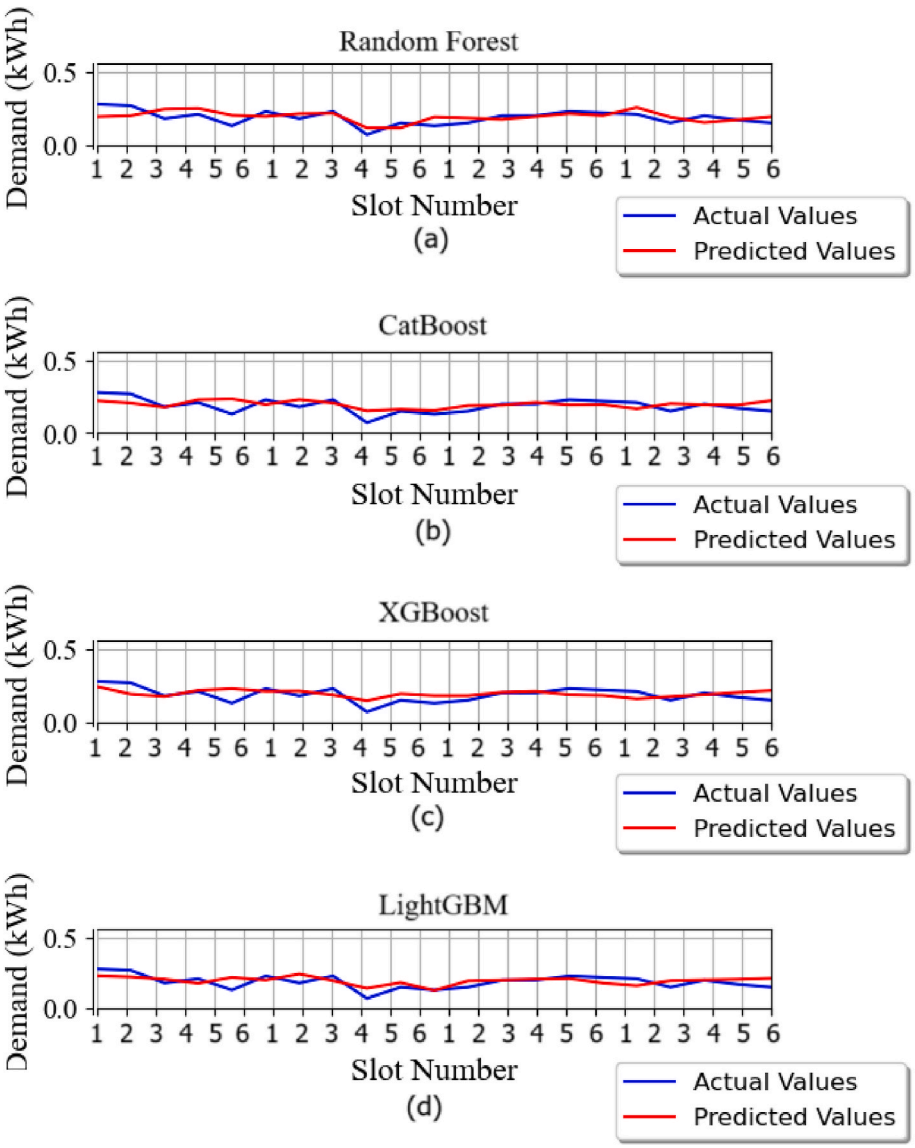


Fig. 17. Actual-Predicted demand curve using (a) Random Forest (b) CatBoost (c) XGBoost (d) LightGBM regression models (Case study I).

same slot, previous day previous slot total demand. The output variable indicates the total energy demand for a prescribed time slot and date. The normalized data with the feature vectors mentioned in Table 2 are split into train and test data sets. The train data is applied to the various regression models such as RFR, CBR, XGBR and LGBMR. The models are fine-tuned to obtain much better result with the hyperparameter setting shown in Table 6.

Post training, the models are tested with unseen dataset. Table 7 provides an overview of the observed performance indices for the RFR, CBR, XGBR, and LGBMR.

Table 7 shows that the performance indices of RFR are 0.0788 for MAE, 0.0125 for MSE, and 0.1118 for RMSE. Conversely, CBR demonstrates the most favorable results, with an RMSE of 0.1059, MSE of 0.0112, and MAE of 0.0726. XGBR exhibits the least optimal outcome among the four models, with MAE of 0.0942, MSE of 0.0186, and RMSE of 0.1365. Meanwhile, LGBMR displays MAE, MSE, and RMSE values of 0.0797, 0.0131, and 0.1145, respectively. The performance indices obtained from different models are represented in the form of bar chart and is depicted in Fig. 14.

From Fig. 14, it is evident that the MAE produced by LGBMR and RFR models are nearly equal. The XGBR model is found to produce highest MAE, MSE and RMSE. The CBR model provides the best prediction accuracy among the four models.

7.3. Case study III

The feature vector in this study includes the date, day of the week, slot number, average demand at the same slot on the previous day, average demand at the same slot for two days ahead, and the average demand at the preceding slot on the previous day while the response indicates the total energy demand for a specified slot and date. The preprocessed, normalized train data is given to the RFR, CBR, XGBR and LGBMR model with the parameter setting as given in Table 8.

The trained RFR, CBR, XGBR and LGBMR models are analyzed for the performance accuracy with preprocessed and normalized test data. The hyperparameter settings for each model are provided in Table 8. The RFR model is trained on the training data and then tested on the test data. The derived performance matrices have an MAE of 0.0953, MSE of 0.0176, and RMSE of 0.1327. Later, the CBR model is trained for 2000 iterations using the same data set. The testing is then completed, with MAE, MSE, and RMSE values of 0.092, 0.0166, and 0.129, respectively. The XGBR model also trained using the same dataset for 10,000 iterations. The MAE, MSE, and RMSE values are acquired throughout the testing process. The results obtained are MAE 0.1136, MSE 0.0244, and

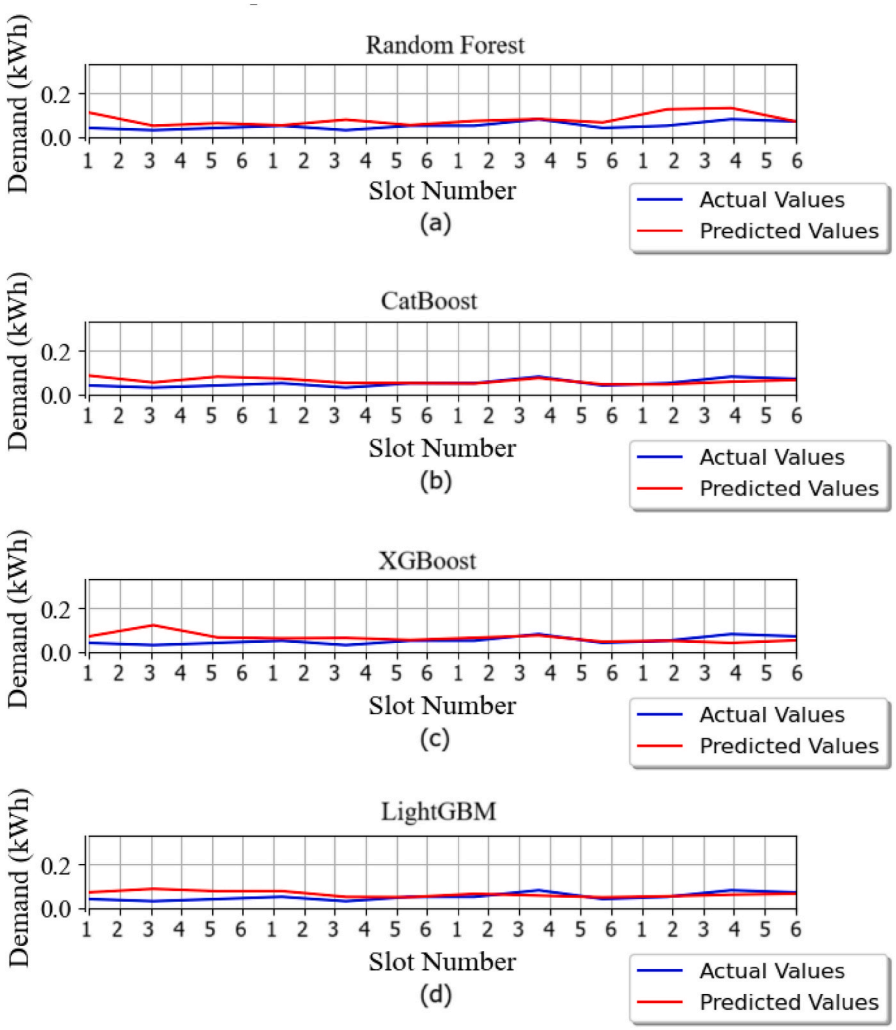


Fig. 18. Actual-Predicted demand curve using (a) Random forest (b) CatBoost (c) XGBoost (d) LightGBM regression models (Case study II).

RMSE 0.1561. After training with default hyperparameter setting and testing on the same set of data, the LGBMR model produces an RMSE value of 0.1246, an MAE of 0.0878, and an MSE of 0.0155. The observed performance indices from different models are tabulated in Table 9.

In the performance indices shown in Table 9, corresponding to case study III is found to be 0.0953, 0.0176, and 0.1327 against MAE, MSE, and RMSE respectively for RFR model. The CBR model slightly outperforms RFR with MAE of 0.0920, MSE of 0.0166, and RMSE of 0.1290. The least successful prediction comes from XGBR with MSE value of 0.0244 and RMSE of 0.1561, alongside an MAE of 0.1136. Conversely, LGBMR demonstrates the lowest prediction error, with an RMSE of 0.1246, MAE of 0.0878, and MSE of 0.0155. The MAE, MSE and RMSE for the four ML models are depicted as bar chart in Fig. 15.

Fig. 15 shows that the MSE value of RFR, CBR and LGBMR models are almost identical. The RMSE values of CBR and LGBMR are nearly equal. LGBMR showcases comparatively less MAE, MSE and RMSE values compared to RFR, CBR and XGBR and hence it is expected to provide an accurate prediction of EVCS energy demand.

7.4. Case study IV

The input feature vector employed in this case study includes the date, day of the week, slot number, total demand at the same slot on the previous day, total demand at the same slot two days prior, and the total demand at the preceding slot for the day before, where as the output is the average energy demand for a specified slot and date. The 80 % pre-processed and normalized dataset is given to the four regression models, which are defined using the parameter setting as provided in Table 10.

The RFR model defined using the parameter setting given in Tables 10 and 100 decision trees, is trained with 80 % data. The model is tested for its performance and the MAE, MSE and RMSE values obtained are 0.1004, 0.0173 and 0.1314 respectively. The CBR model is trained with the same data for 2100 iterations. The MAE, MSE and RMSE values are obtained during the testing. The values are obtained to be MAE of 0.1031 MSE, 0.0178 and the obtained RMSE value is of 0.1334. XGBR Regression model also provided with the same data and the model is trained for 11000 iterations. The testing is then performed and the MAE, MSE and RMSE values obtained to be 0.1212, 0.0236 and 0.1536 respectively. Finally the LGBMR model is trained with default hyperparameter setting and the testing results shows great prediction accuracy with least MAE, MSE and RMSE among the four models. The observed performance indices are tabulated in Table 11.

In this context, the RFR model demonstrates a satisfactory predictive accuracy, achieving 0.1004 MAE, MSE of 0.0173, and RMSE of 0.1314. CBR closely follows with performance indices of 0.1031 for MAE, 0.0178 for MSE, and 0.1334 for RMSE. Among the four models, XGBR yields the least accurate results, with an MAE of 0.1212, MSE of 0.0236, and RMSE of 0.1536 as shown in Table 11. Conversely, the LGBMR model once again outperforms the others, displaying the lowest MSE, RMSE, and MAE values-0.0160, 0.1266, and 0.0965, respectively. The bar chart representation of MAE, MSE and RMSE observed from RFR, CBR, XGBR and LGBMR predictions models while testing is presented in

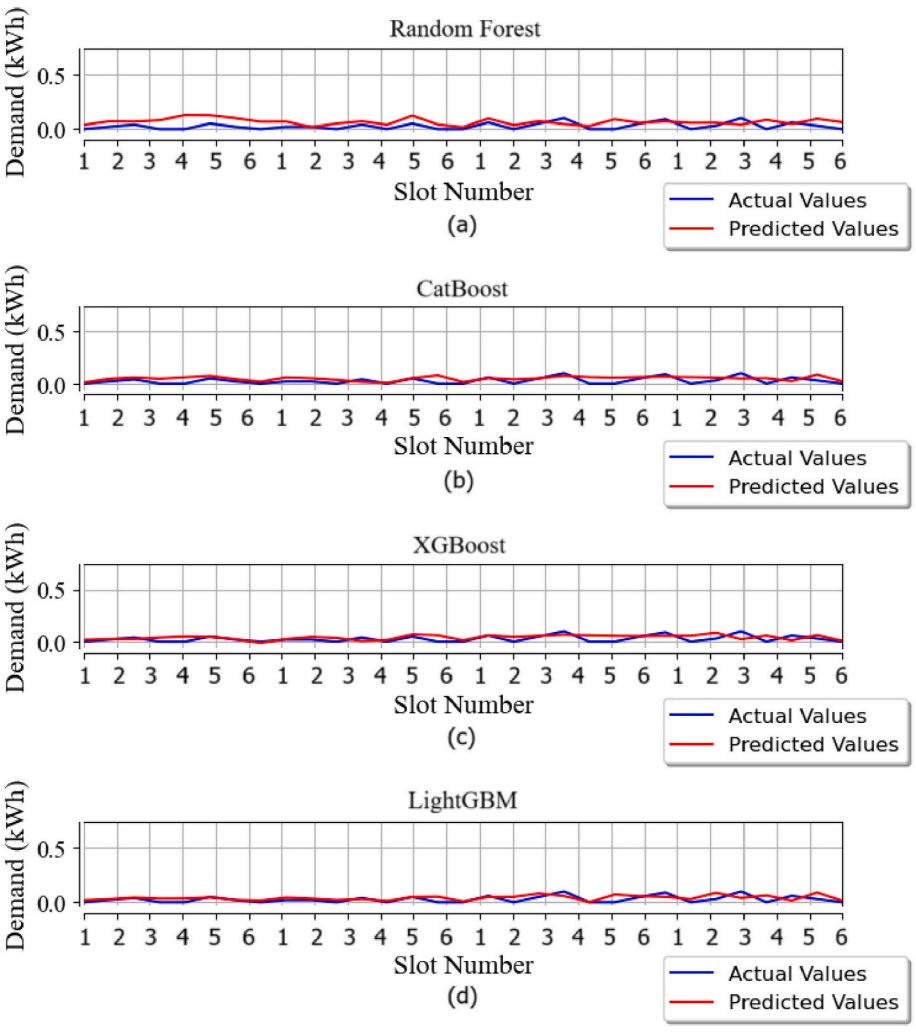


Fig. 19. Actual-Predicted demand curve using (a) Random forest (b) CatBoost (c) XGBoost (d) LightGBM regression models (Case study III).

Fig. 16.

As illustrated by Fig. 16 that RFR model closely follows CBR in terms of MAE, MSE and RMSE. LGBMR is expected to outperform with the lowest MAE, MSE, and RMSE values compared to other models. Once again XGBR model showcases lowest prediction accuracy with highest values of MAE, MSE and RMSE. It is now required to validate the prediction performance of RFR, CBR, XGBR and LGBMR employing the dataset September 01 2017 to December 08, 2019. The actual versus predicted demand curves are plotted under different case studies using the suggested ML models. The predicted and actual graphs are plotted as shown in Fig. 17.

Fig. 17(a), (b), 17(c) and 17(d) illustrate the actual and predicted demand curves of RFR, CBR, XGBR and LGBMR models. The X-axis shows the slot number that is from slot 1 to slot 6 for four days and Y-axis represents the average energy demand of various ML models. The graphs show the actual demand fluctuates significantly. RFR model is found to provide the prediction that matches with the actual for the days between slot 1 to slot 3 on day 2 and slot 2 to 6 on day 3. CBR model provides a prediction similar to actual from slot 3 to slot 4 on day 1 and slot 3 to slot 4 on day 3 and at slot 4 on day 4. XGBR model seems to follow actual values between slot 3 and slot 4 on day 1, slot 3 and slot 4 on day 3 and slot 2 and slot 4 on day 4. LGBMR model provides a prediction nearly equal to actual around slot 6 on day 2 and slot 3 to slot 4 on day 3. The models are verified for feature vector that include date, day of the week, slot number, total demand at same slot on previous day, total demand at same slot on two days prior, and total demand at previous slot on the previous day and total energy demand for a prescribed slot and date as the output variable. The obtained results are shown in Fig. 18.

In Fig. 18(a)–(b), Fig. 18(c) and (d) shows the actual and predicted demand curves of EVCS produced by RFR, CBR, XGBR and LGBMR. RFR model is found to provide the prediction that matches with the actual at slot 1 on day 2, between slot 5 to slot 6 on day 2 and from slot 3 to slot 5 on day 3. CBR model provides a prediction similar to actual between slot 5 to slot 6 on day 2 and between slot 1 to slot 6 on day 3. XGBR model is seen to predict values similar to actual values from slot 2 to slot 6 on day 3 and slot 1 to slot 2 on day 4. LGBMR model provides a prediction nearly equal to actual between slot 5 to slot 6 on day 2, around slot 2 on day 3 and from slot 2 to slot 3 on day 4. The models are verified for feature vector that include date, day of the week, slot number, average demand on the same slot on previous day, average demand at same slot two days prior, and average demand at previous slot on the previous day and the total energy demand for a given slot and date as the output variable. The results are presented in Fig. 19.

Fig. 19(a) shows that RFR model is found to provide the prediction that matches with the actual at slot 2 on day 2, slot 4 on day 3 and slot 6 on day 3. Fig. 19(b) clearly displays that CBR model provides a prediction similar to actual from slot 1 to slot 3 on day 1, slot 5 to slot 6 on day 1 and slot 4 to slot 5 on day 2. XGBR model is seen to predict values similar to actual values for slot values between 1 and 3 and slot 5 to 6 on day 1, slot 5 on day 2 and slot 1 on day 3. This is evident in Fig. 19(c). LGBMR model provides a prediction nearly equal to actual between slot 1 to slot 3 and slot 5 to slot 6 on day 1, slot 3 to slot 5 on day 2 and at slot

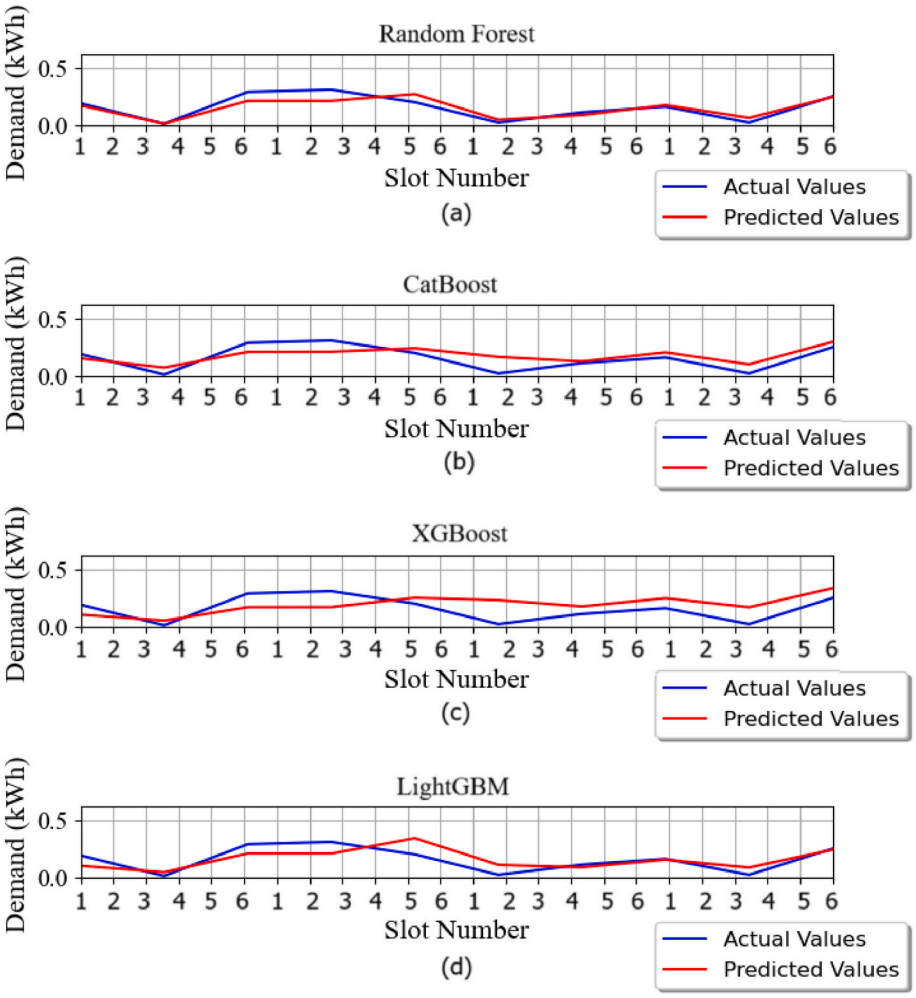


Fig. 20. Actual-Predicted demand curve using (a) Random forest (b) CatBoost (c) XGBoost (d) LightGBM regression models (Case study IV).

6 on day 3, as seen in Fig. 19(d). The models are verified for feature vector that include date, day of the week, slot number, total demand at same slot on previous day, total demand at same slot on two days prior, and total demand at previous slot on the previous day and average energy demand for a prescribed slot and date as the output variable. The obtained results are shown in Fig. 20.

In Fig. 20(a)–20(d) the actual and predicted curves of EVCS are shown, generated by the RFR, CBR, XGBR and LGBMR respectively. The RFR model closely aligns with the actual demand for the index between slot 1 to slot 3 on day 1 and slot 2 to slot 6 on third day. The CBR model provides accurate prediction corresponds to slot 2 on and slot 5 of day 1 and slot 4 to slot 6 on day 3. The XGBR models predictions match with the actual values at slot 3 and slot 4 on day 1. Finally, the LGBMR model predictions closely match the actual values under at slot 3 and slot 4 on day 1, slot 4 to slot 6 on day 3, between slot 1 and slot 2 and slot 6 on day 4. The comparative analysis of prediction accuracy of regression models RFR, CBR, XGBR and LGBMR under various cases has been depicted in Fig. 21.

The bar chart shown in Fig. 21 illustrates the performance of RFR, CBR, XGBR and LGBMR models that are subjected to four distinct case studies. The performances of these models are evaluated using MAE, MSE and RMSE from Fig. 21(a) and 21 (b) and Fig. 21 (c) respectively. In case study I, XGBR standout with lowest MAE, MSE and RMSE values of 0.1476, 0.0448 and 0.2117 respectively. On the other hand, LGBMR yield the lowest prediction accuracy with 0.2782 MAE, MSE of 2.0541 and 1.4332 RMSE value. In case study II the CBR demonstrate superior performance with lowest MAE of 0.0726, MSE of 0.0112 and RMSE of 0.1059, while RFR closely follows with 0.0788 MAE, 0.0125 MSE and 0.1118 RMSE. The XGBR produces least prediction accuracy among the four models. In the case study III LGBMR emerges as a winner in terms of prediction accuracy. The predicted demand that is observed from LGBMR closely follows the actual values with 0.0878 MAE, while MSE of 0.0155 and 0.1246 RMSE. The CBR showcases fine predictions with MAE, MSE and RMSE of 0.092, 0.0166 and 0.129 respectively. The RFR has produced good prediction accuracy. Meanwhile, XGBR showcases highest error with 0.1136 MAE, 0.0244 MSE and 0.1561 RMSE. Finally, the in case study IV once again LGBMR exhibits highest prediction accuracy with a MAE of 0.0965, MSE of 0.016 and RMSE of 0.126. Other models, including RFR, CBR and XGBR display higher prediction error.

The comparison of four regression models – RFR, CBR, XGBR and LGBMR across four different case studies indicate that CBR generally outperforms the other prediction models in terms of MAE, MSE and RMSE. XGBR shows highest error metrics while LGBMR shows better prediction results in case study III and IV. The RFR model produces satisfactory prediction accuracy throughout case study I to IV. Considering all cases and models, case study II stands out in terms of formation of dataset with input feature vectors that include the date, day of the week, slot number, total demand at same slot for the previous day, 2 days prior total demand at same slot, previous day previous slot total demand, and the average energy demand as the output. The chosen dataset is shown to provide a superior performance compared to all the feature vector combinations mentioned in section 7.1, 7.3 and 7.4. With this suggested feature vector it is observed that CBR model provides prediction of total energy demand with high accuracy when compared

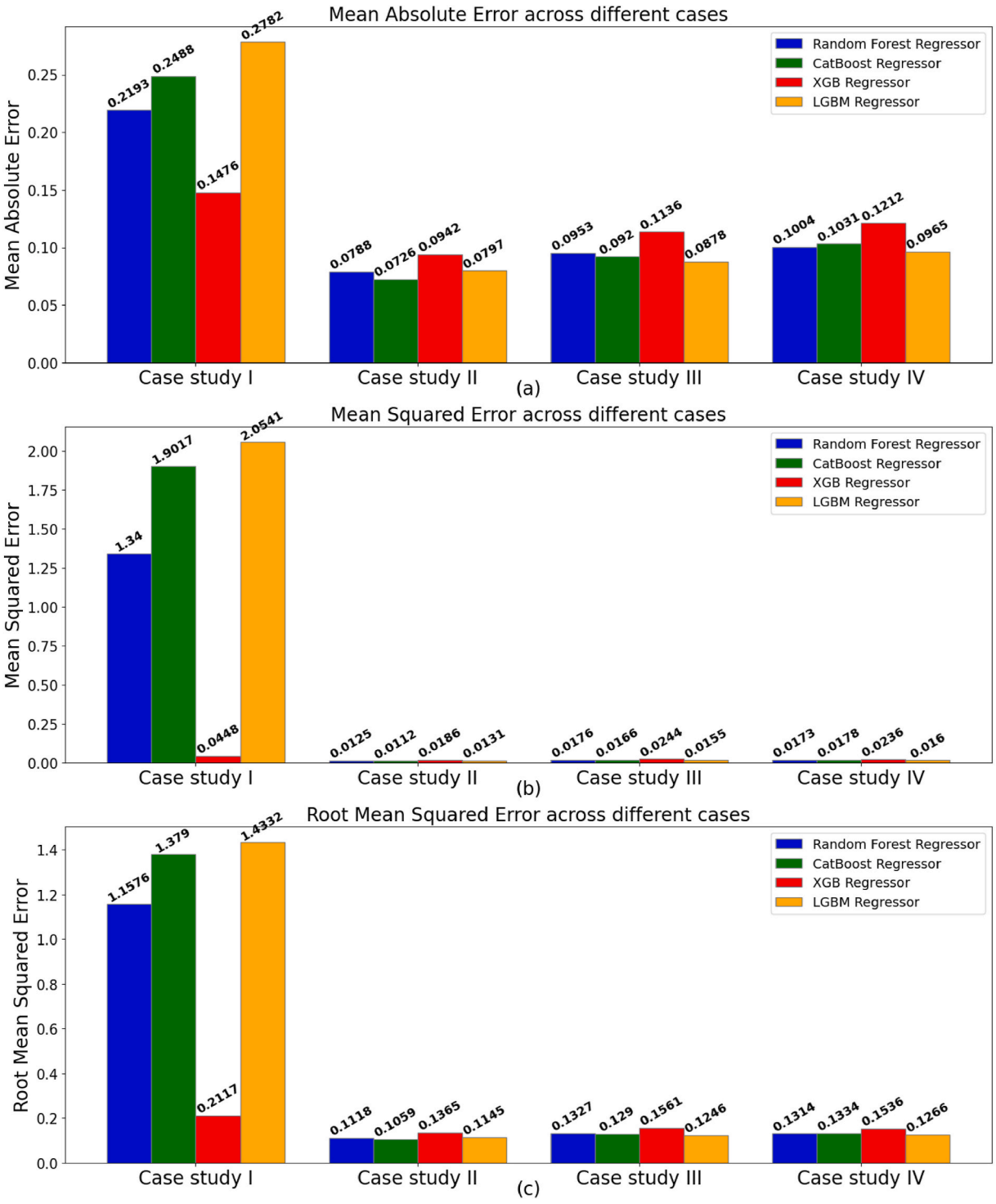


Fig. 21. Performance analysis of Random Forest, CatBoost, XGBoost and LightGBM regression models under case study I - IV using (a) MAE (b) MSE (c) RMSE.

to RFR, XGBR and LGBMR models. It can be concluded that CBR model can be deployed as the optimal model to predict the total energy demand of EVCS, while employing feature vector that include the date, day of the week, slot number, total demand at same slot for the previous day, 2 days prior total demand at same slot, previous day previous slot total demand.

8. Conclusions

Precise Electric Vehicle Charging Station demand forecast is essential for efficient resource management, optimization of operating cost, capacity planning, customer satisfaction while maintaining grid stability and reliability. This paper identifies a most suitable data driven model to forecast short term demand among Random Forest, Categorical Boosting, Extreme Gradient Boosting and Light Gradient Boosting Machine. The models are developed while employing the dataset sourced from Perth & Kinross Council's EV charging stations located in Perth from September 01, 2017 to December 08, 2019. The data is subjected to preprocessing to obtain average demand for 1-h, 2-h, 3-h, 4-h slots. The analysis based on the variance shows that 4-h slot data provides

minimum variance of 0.1064. The chosen 4-h slot is then utilized to frame the 4 distinct feature vectors that are employed in developing the considered machine learning models. The performance analysis of these models shows that, Categorical Boosting model provides a consistent performance with MAE, MSE, RMSE of 0.0726, 0.0112 and 0.1059 under case II; 0.092, 0.0166 and 0.129 under case study III; and 0.1031, 0.0178 and 0.1334 under case study IV. The analysis also proves that the identified model responds with appreciable performance for feature vector that includes the date, day of the week, slot number, total demand at the same slot on the previous day, total demand at the same slot two days prior, and the total demand at the preceding slot on the previous day, while the output variable represents the total energy demand for a specified slot and date as the best scenario. The selected feature vector is found to provide minimum MAE, MSE and RMSE of 0.0726, 0.0112 and 0.1059 respectively. The suggested data set along with the identified Categorical Boosting Regression model proves to provide promising advantages to the operator in optimizing the operational efficiency, cost management and strategic planning for future expansion.

CRediT authorship contribution statement

A.V. Sreekumar: Writing – original draft, Visualization, Methodology, Investigation, Data curation. **R.R. Lekshmi:** Writing – review & editing, Validation, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] A. Desreveaux, A. Bouscayrol, R. Trigui, E. Hittinger, E. Castex, G.M. Sirbu, Accurate energy consumption for comparison of climate change impact of thermal and electric vehicles, Energy 268 (2023) 126637, https://doi.org/10.1016/j.energy.2023.126637.

[2] K. Petrauskienė, M. Skvarnavičiūtė, J. Dvarionienė, Comparative environmental life cycle assessment of electric and conventional vehicles in Lithuania, J. Clean. Prod. 246 (2020) 119042, https://doi.org/10.1016/j.jclepro.2019.119042.

[3] R.S. Rani, R. Jayapragash, Review on electric mobility: trends, challenges and opportunities, Results in Engineering 23 (2024) 102631, https://doi.org/10.1016/j.rineng.2024.102631.

[4] M.O. Khan, S. Kirmani, M. Rihan, Impact assessment of electric vehicle charging on distribution networks, Renewable Energy Focus 50 (2024) 100599, https://doi.org/10.1016/j.ref.2024.100599.

[5] Tong Yang, Derek Clements-Croome, Matthew Marson, Building energy management systems, in: A. Martin (Ed.), Abraham, Encyclopedia of Sustainable Technologies, second ed., Elsevier, 2024, pp. 727–749, https://doi.org/10.1016/B978-0-323-90386-8.00025-5. ISBN 9780443222870.

[6] M. Bharathidasan, V. Indragandhi, V. Suresh, M. Jasiński, Z. Leonowicz, A review on electric vehicle: technologies, energy trading, and cyber security, Energy Rep. 8 (2022) 9662–9685, https://doi.org/10.1016/j.egyr.2022.07.145.

[7] F. Marzbani, A.H. Osman, M.S. Hassan, Electric vehicle energy demand prediction techniques: an in-depth and critical systematic review, IEEE Access 11 (2023) 96242–96255, https://doi.org/10.1109/ACCESS.2023.3308928.

[8] H. Li, J. Zhu, Y. Zhou, D. Feng, K. Zhang, B. Shen, Review of load forecasting methods for electric vehicle charging station, in: Proceedings of the 2022 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia), 2022, pp. 1833–1837, https://doi.org/10.1109/ICPSAsia55496.2022.9949707. Shanghai, China.

[9] K.C. Akshay, G.H. Grace, K. Gunasekaran, et al., Power consumption prediction for electric vehicle charging stations and forecasting income, Sci. Rep. 14 (2024) 6497, https://doi.org/10.1038/s41598-024-56507-2.

[10] S. Shahriar, A.R. Al-Ali, A.H. Osman, S. Dhou, M. Nijim, Machine learning approaches for EV charging behavior: a review, IEEE Access 8 (2020) 168980–168993, https://doi.org/10.1109/ACCESS.2020.3023388.

[11] S. Shahriar, A.R. Al-Ali, A.H. Osman, S. Dhou, M. Nijim, Prediction of EV charging behavior using machine learning, IEEE Access 9 (2021) 111576–111586, https://doi.org/10.1109/ACCESS.2021.3103119.

[12] C.S. Gujjarlapudi, D. Sarkar, S.K. Gunturi, Data driven machine learning models for short term load forecasting considering electrical vehicle load, Energy Storage (2024) e467, https://doi.org/10.1002/est2.467.

[13] J. Liu, G. Lin, C. Rehtanz, S. Huang, Y. Zhou, Y. Li, Data-driven intelligent EV charging operating with limited chargers considering the charging demand forecasting, Int. J. Electr. Power Energy Syst. 141 (2022) 108218, https://doi.org/10.1016/j.ijepes.2022.108218.

[14] B.-S. Lee, H. Lee, H. Ahn, Improving load forecasting of electric vehicle charging stations through missing data imputation, Energies 13 (2020) 4893, https://doi.org/10.3390/en13184893.

[15] S. Wang, C. Zhuge, C. Shao, P. Wang, X. Yang, S. Wang, Short-term electric vehicle charging demand prediction: a deep learning approach, Appl. Energy 340 (2023) 121032, https://doi.org/10.1016/j.apenergy.2023.121032.

[16] A. Garrison, M. Rashid, N. Chen, A synergistic learning based electric vehicle charging demand prediction scheme, in: Proc. SoutheastCon 2023, 2023, pp. 5–10, https://doi.org/10.1109/SoutheastCon51012.2023.10115078. Orlando, FL, USA.

[17] H. Sun, S. Wang, C. Liu, Load forecasting of electric vehicle charging station based on power big data and improved BP neural network, Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery 153 (2023) 425–435, https://doi.org/10.1007/978-3-031-20738-9_47.

[18] S. Koohfar, W. Woldemariam, A. Kumar, Prediction of electric vehicles charging demand: a transformer-based deep learning approach, Sustainability 15 (2023) 2105, https://doi.org/10.3390/su15032105.

[19] P. Banda, M.A. Bhuiyan, K.N. Hasan, K. Zhang, Assessment of hybrid transfer learning method for forecasting EV profile and system voltage using limited EV charging data, Sustainable Energy, Grids and Networks 36 (2023) 101191, https://doi.org/10.1016/j.segan.2023.101191.

[20] S.R. Fahim, R. Atat, C. Kececi, A. Takiddin, M. Ismail, K.R. Davis, E. Serpedin, Forecasting EV charging demand: a graph convolutional neural network-based approach, in: Proceedings of the 2024 4th International Conference on Smart Grid and Renewable Energy (SGRE), April 2024, pp. 1–6, https://doi.org/10.1109/SGRE59715.2024.10428726.

[21] Q. Li, H. Xiong, Short-term load prediction of electric vehicle charging stations built on GA-BPNN model, in: Proceedings of the 2023 International Conference on Journal of Physics: Conference Series, April 2023 012023, https://doi.org/10.1088/1742-6596/2474/1/012023 vol. 2474, no. 1.

[22] F. Mohammad, D.-K. Kang, M.A. Ahmed, Y.-C. Kim, Energy demand load forecasting for electric vehicle charging stations network based on ConvLSTM and BiConvLSTM architectures, IEEE Access 11 (2023) 67350–67369, https://doi.org/10.1109/ACCESS.2023.3274657.

[23] M. Bharat, R. Dash, K.J. Reddy, A.S.R. Murty, C. Dhanamjayulu, S.M. Muyeen, Secure and efficient prediction of electric vehicle charging demand using α2-LSTM and AES-128 cryptography, Energy and AI 16 (2024) 100307, https://doi.org/10.1016/j.egyai.2023.100307.

[24] Z. Yi, X.C. Liu, R. Wei, X. Chen, J. Dai, Electric vehicle charging demand forecasting using deep learning model, Journal of Intelligent Transportation Systems 26 (2021) 690–703, https://doi.org/10.1080/15472450.2021.1966627.

[25] M.D. Eddine, Y. Shen, A deep learning based approach for predicting the demand of electric vehicle charge, J. Supercomput. 78 (2022) 14072–14095, https://doi.org/10.1007/s11227-022-04428-0.

[26] M.A. Zamee, D. Han, H. Cha, D. Won, Self-supervised online learning algorithm for electric vehicle charging station demand and event prediction, J. Energy Storage 71 (2023) 108189, https://doi.org/10.1016/j.est.2023.108189.

[27] A. Orzechowski, L. Lugosch, H. Shu, R. Yang, W. Li, B.H. Meyer, A data-driven framework for medium-term electric vehicle charging demand forecasting, Energy and AI 14 (2023) 100267, https://doi.org/10.1016/j.egyai.2023.100267.

[28] F. Ru, X. Yang, H. Zou, L. Zhang, X. Xu, Electric vehicle charging station load prediction based on IWOA-LSSVM combined model with improved variational Mode decomposition, Proceedings of the Journal of Physics: Conference Series 2320 (2022) 012007, https://doi.org/10.1088/1742-6596/2320/1/012007.

[29] S. Koohfar, W. Woldemariam, A. Kumar, Performance comparison of deep learning approaches in predicting EV charging demand, Sustainability 15 (2023) 4258, https://doi.org/10.3390/su15054258.

[30] J. Zheng, J. Zhu, H. Xi, Short-term energy consumption prediction of electric vehicle charging station using attentional feature engineering and multi-sequence stacked gated recurrent unit, Comput. Electr. Eng. 108 (2023) 108694, https://doi.org/10.1016/j.compeleceng.2023.108694.

[31] M. Dabbaghjamanesh, A. Moeini, A. Kavousi-Fard, Reinforcement learning-based load forecasting of electric vehicle charging station using Q-learning technique, IEEE Trans. Ind. Inform. 17 (2021) 4229–4237, https://doi.org/10.1109/tii.2020.2990397.

[32] S. Wang, A. Chen, P. Wang, C. Zhuge, Predicting electric vehicle charging demand using a heterogeneous spatio-temporal graph convolutional network, Transport. Res. Part C Emerg, Technol. 153 (2023) 104205, https://doi.org/10.1016/j.trc.2023.104205.

[33] V. Jishnu Sankar, A. Hareendran, M. Nair, Enhanced smart grid resilience using autonomous EV charging station, in: P. Siano, S. Williamson, S. Beevi (Eds.), Intelligent Solutions for Smart Grids and Smart Cities, IPECS 2022, Lecture Notes in Electrical Engineering, vol. 1022, Springer, Singapore, 2023, pp. 153–166, https://doi.org/10.1007/978-981-99-0915-5_10.

[34] Perth and Kinross Open Data, EV Charging Data (2024). https://data.pkc.gov.uk/ (accessed January 15, 2024).

[35] R. Muthukrishnan, R. Rohini, LASSO: a feature selection technique in predictive modeling for machine learning, in: 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 18–20, https://doi.org/10.1109/ICACA.2016.7887916.

[36] A. Amato, V. Di Lecce, Data preprocessing impact on machine learning algorithm performance, Open Computer Science 13 (2023) 20220278, https://doi.org/10.1515/comp-2022-0278.
[37] V. Çetin, O. Yıldız, A comprehensive review on data preprocessing techniques in data analysis, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi 28 (2022) 299–312.
[38] K. Maharana, S. Mondal, B. Nemade, A review: data pre-processing and data augmentation techniques, Global Transitions Proceedings 3 (2022) 91–99, https://doi.org/10.1016/j.gltp.2022.04.020.
[39] P.A. Kowalski, M. Walczak, Feature selection for regression tasks based on explainable artificial intelligence procedures, in: Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 2023, pp. 1–8, https://doi.org/10.1109/IJCNN54540.2023.10191064.
[40] T.A. Alghamdi, N. Javaid, A survey of preprocessing methods used for analysis of big data originated from smart grids, IEEE Access 10 (2022) 46372–46388, https://doi.org/10.1109/ACCESS.2022.3157941.
[41] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, 1984.
[42] E. Al Daoud, Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset, Int. J. Comput. Inf. Eng. 13 (2019) 6–10, https://doi.org/10.5281/zenodo.3607805.
[43] A.V. Sreekumar, R.R. Lekshmi, Comparative study of data-driven methods for state of charge estimation of Li-ion battery, in: Proceedings of the 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS), 2023, pp. 1–6, https://doi.org/10.1109/PCEMS58491.2023.10136079.
[44] D. Sathyan, D. Govind, C.B. Rajesh, K. Gopikrishnan, G.A. Kannan, J. Mahadevan, Modelling the shear flow behaviour of cement paste using machine learning – XGBoost, J. Phys. Conf. 1451 (2020) 012026, https://doi.org/10.1088/1742-6596/1451/1/012026. IOP Publishing.
[45] A. Manoharan, K.M. Begam, V. Rau Aparow, D. Sooriamoorthy, Artificial neural networks, gradient boosting and Support vector machines for electric vehicle battery state estimation: a review, J. Energy Storage 55 (2022) 105384, https://doi.org/10.1016/j.est.2022.105384.
[46] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17), 2017, pp. 3149–3157.