A Computational Implementation to Forecast Electric Vehicles Usage in the Power System

Herbert Amezquita, Cindy P. Guzman, and Hugo Morais

INESC-ID, Department of Electrical and Computer Engineering, Instituto Superior Técnico—IST,
Universidade de Lisboa, 1049-001 Lisboa, Portugal
Email: {herbert.amezquita, cindy.lascano, hugo.morais}@tecnico.ulisboa.pt

***Abstract*—The mass adoption of electric vehicles (EVs) has been accelerated by the need for the energy transition. Nevertheless, the continuous growth of EVs can bring new challenges to electric power systems, which include voltage and congestion constraints in low-voltage (LV) and medium-voltage (MV) networks. To anticipate these constraints, it is necessary to develop algorithms to predict EV charging behaviour. Therefore, this paper proposes a computational implementation based on Random Forest for the accurate forecast of EV power consumption. The process starts with the data collection, followed by preprocessing, to clean and prepare the data to be used in the forecast method. Then, feature engineering and feature selection steps are applied to create and select the inputs (features) for the forecasting process, and finally, the forecast and validation steps are implemented. The case study analyzed uses real data from the charging station inside the University Instituto Superior Técnico and the results demonstrate the effectiveness of the proposed forecast method since for one year of training and one month of forecasting were obtained a Mean of the Absolute Error of 203.23W and a Normalized Root Mean Square Error of 3.44%.**

***Index Terms*—Charging Stations, Electric Vehicles, EV Power Consumption Forecast, Machine Learning, Random Forest.**

I. Introduction

Both the increasing penetration of renewable energy sources and the commitment to carbon neutrality have boosted electric vehicles (EVs) into one of the most reliable and economically viable solutions to reduce greenhouse gas emissions [1]. EV adoption in Europe has shown steadily growing, for instance, in 2021 EV sales achieved an annual growth of 65% over the previous years [1]. Due to the highly uncertain behaviour of EV users, EVs mass adoption represents a new and problematic power demand for the electric power system [2], by introducing high variation in the normal power demand, and degradation in the local transformer, among others. Therefore, EV power consumption forecasting is a key measure in power system planning, scheduling, and operation [3], which can help to better manage the scheduling of EV charging stations (CSs) in an optimal and secure way for the power system [4]. Several research works have been devoted to developing forecast strategies to support the optimal management of EV charging behaviour [3], [5]. A spatial-temporal model, based on Monte Carlo simulation, aiming to analyse the impact of large-scale deployment of EVs on the urban distribution network was proposed by the author in [5]. Considering the integration of information related to the power system, transportation network, EV technical specifications, and market data, it was possible to obtain the EV charging load forecast. In [6] the authors propose a methodology to predict the additional EV charging load in the mid-and-long term. The methodology proposed includes probabilistic modelling aiming to analyse EV charging profiles, and consequently predict the EV future pattern ownership. Results show information related to the hour of peak EV power consumption for 2025, indicating an increase of 11.08% of the total electric demand for the case study. A forecasting strategy for the EV charging load based on the Random Forest (RF) algorithm has been proposed in [7]. RF algorithm is used to realize short-term forecast focus on the charging station. A large amount of historical data is analyzed and learned to implement RF and to obtain an effective prediction of the EV charging load. Nevertheless, this proposal does not implement the RF algorithm considering other aspects such as pre-processing, feature engineering, or clustering in a complete framework. The authors in [8] implemented four different forecasting methods: nearest neighbour, modified pattern sequence forecasting (MPSF), Support Vector Regression, and RF. Hence, through a comparison of these methods was possible to analyze the EV charging load data and find that the most suitable method to forecast the EV charging demand was the MPSF.

This work adds a contribution to the existing literature by proposing: 1) A computational implementation based on a complete framework considering the analysis of raw data and implementing stages of pre-processing, feature engineering, feature selection, forecasting based on the RF method, and validation. 2) An evaluation of the performance of the RF algorithm to predict the charging behaviour of EVs, using real EV power consumption data from a case study. 3) A comparison of the results by using different error metrics.

II. Methodology

The methodological framework proposed in this paper is shown in Figure 1. Hence, it is composed of the following stages: 1. Pre-Processing, 2. Feature Engineering, 3. Feature Selection, 4. Forecasting Method, and 5. Validation.

This research work was funded by European Union's Horizon Europe R&I programme under grant agreement no. 101056765. Views and opinions expressed in this document are however those of the authors only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the grating authority can be held responsible for them. This work also was funded by the Portuguese Foundation for Science and Technology (FCT) under UIDB/50021/2020.

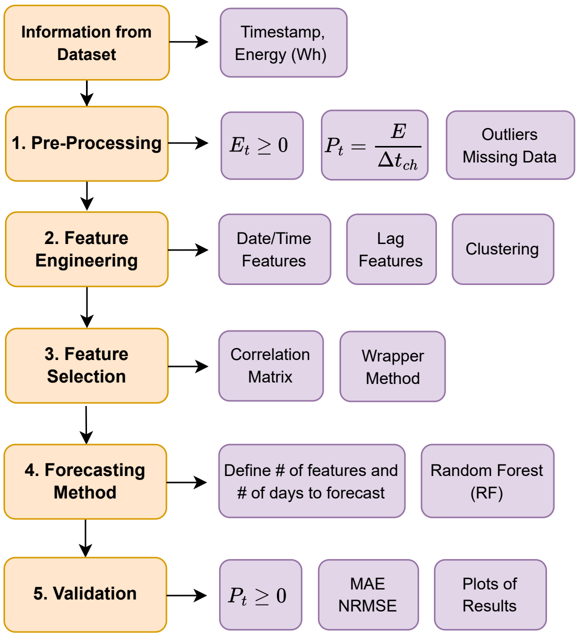

Fig. 1: Methodological framework proposed

A. *Pre-Processing*

In the pre-processing step, the raw dataset is analyzed, in which, only data that satisfy the condition $Energy \geq 0$ is considered. In case there are negative values of the energy, those values are removed. Then, the power consumption is calculated by dividing the energy by 0.25 (given that the time resolution of 15 minutes is equivalent to 0.25h). For handling missing data, two approaches are applied. 1) If there are less than two hours of missing values, linear interpolation is used to fill the gap. 2) For more than two hours of missing values, the gap is not filled because creating artificial values for long periods of time may have a negative impact on the forecast. Both approaches are included in the computational implementation (to use with future datasets), despite, no missing values existing in the dataset used.

B. *Feature Engineering*

In this stage, relevant and useful features that will be input into the forecasting algorithm are generated. Two categories of features are created, namely date/time features and lag features. In the first group, the following variables are generated from the timestamp of each record: Week of Year (from 1-52), Day of Year (from 1-365), Season (from 1-4), Month (from 1-12), Day of Month (from 1-30/31), Day of Week (from 1-7), Weekend (1 if weekend, 0 if not), Holiday (1 if holiday, 0 if not), Hour (from 1-24) and Minute (15, 30 or 45). In the second group, past values of power are used to create the following variables: Lag Consumption 1 (power value 15 minutes before), Lag Consumption 4 (power value 1 hour before), Mean Rolling 4 (mean power of last hour), Lag Consumption 96 (power value 1 day before) and Mean Rolling 96 (mean power of last day).

Finally, a clustering process is also performed at this stage, with the purpose of identifying daily patterns of EV power consumption in the CS. To do that, the power consumption is classified into clusters or groups, using K-means method [9]. The algorithm first chooses random points as centroids and then iterates adjusting them until full convergence [9]. To define the optimal number of clusters (K) for this case study, the elbow method, which is a line plot comparing the number of clusters and the total within clusters sum of squares was implemented [10]. The optimal number of clusters corresponds to the point where the curve starts to bend (the elbow of the curve). Once the clusters (groups) are created, a new variable called Clusters is generated assigning the cluster number to each power consumption data point.

C. *Feature Selection*

The feature selection intends to determine which variables have the higher correlation with the forecast variable, namely EV power consumption, and will be therefore used as input to the forecasting method. First, it is necessary to transform some of the date/time features such as Week of Year, Day of Year, Season, Month, Day of Month, Day of Week, Hour, and Minute; because they are cyclic variables by nature. Hence, each of these features is transformed into two components (x and y) according to:

$$f_x = \sin\left(2\pi f/max(f)\right) \tag{1}$$

$$f_y = \cos\left(2\pi f/max(f)\right) \tag{2}$$

Where f represents the cyclic feature to be transformed, f_x and f_y represent the first and second components of the cyclic feature, and $max(f)$ corresponds to the maximum value of the cyclic feature [11]. Once the cyclical features are transformed, a correlation matrix is used to observe the relationship between the different variables and the forecast variable (Ev power consumption). Afterwards, for feature selection, a wrapper method based on a specific machine learning algorithm is chosen. Wrapper methods allow to identify the best-performing set of features against the evaluation criterion [12]. For RF, the features are ranked from the higher to the lower score, thus it is decided to use the features that obtained the higher scores to perform the forecast. To determine the specific number of features to use in the forecast, a sensitivity analysis is executed.

D. *Forecasting Method*

RF was used to forecast the EV power consumption. RF is a commonly used machine learning algorithm, which combines the output of multiple decision trees to reach a single result [13]. It is an extension of the bagging method, as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. While decision trees are prone to problems, such as bias and overfitting, RF forms an ensemble with multiple decision trees and uses averaging to improve the predictive accuracy and control over-fitting [13], [14]. At this point, the number of features to use in the algorithm and the number of days to forecast are defined. Based on the time horizon of the forecast, the dataset is divided into two groups: the training set, which is the data used by the algorithm to discover and learn patterns between the features and the forecast variable; and the test set, which is the actual data used by the algorithm to generate the predictions. Both groups

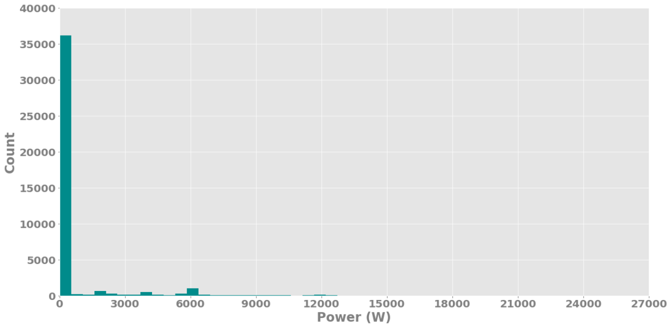

Fig. 2: Power consumption histogram

are subsequently passed to the RF method and the power consumption predictions are obtained. The steps involved in the algorithm are the following:

- Step 1: Randomly select a subset of data points and a subset of features to construct each decision tree.
- Step 2: Individual decision trees are constructed.
- Step 3: Each individual decision tree will generate a prediction.
- Step 4: The final prediction is calculated by averaging the predictions of the individual decision trees.

E. Validation

The final stage involves two steps. 1) The calculation of two error metrics aiming to evaluate the performance of the RF method. 2) The generation of the plots of results aims to compare the real values of power consumption with the predictions obtained.

III. Results and discussion

A. Data Description

The dataset used in this study was provided by Instituto Superior Técnico (IST), University of Lisbon. It consists of the timestamp and the energy consumption (Wh) records of the EV CS located inside the Alameda campus for the period 01/10/2021 to 30/11/2022. The temporal resolution of the data is 15 minutes, which means that the energy records were obtained every 15 minutes.

B. Pre-Processing

After cleaning the data, the final dataset is composed of 40,531 rows of data. Figure 2 presents the histogram of the power consumption in the CS, where it is possible to observe that the majority of the power consumption is lower than 1 kW, specifically 36,353 rows of data. However, there are also some important values of power consumption higher than 15 kW that are imperceptible in the figure but need to be considered in the forecast, as explained in Section II-A. The maximum power consumption registered in the CS for a 15 minutes period is 26.53 kW (imperceptible in Figure 2, because it corresponds only to one single point). It is important to mention that this dataset does not offer information related to how many cars were connected at the same time in a particular period, hence, the only information available is the power consumed in the station in those 15 minutes.

C. Feature Engineering

Figure 3, Figure 4, and Figure 5 show the EV power consumption behaviour with respect to some of the date/time variables created in the feature engineering stage (II-B). Figure 3 corresponds to the average power consumption per hour, which allows observing the daily pattern of the CS. There is a small power consumption for periods of time before 8h00 and after 20h00 and there is a peak power consumption of 2.2kW at 10h00 and 2kW at 11h00. This kind of behaviour is normal since the activities at the University start at 8h00 and people usually put the EV to charge as soon as they arrive at the University.

Figure 4 presents the average power consumption per day of the week in the CS. There is no major difference between weekdays, having an average power consumption of around 0.8 kW from Monday to Friday. The big difference appears during the weekends, in which the average power consumption decreases to 25% (0.2 MW) when compared with the power consumption on Monday. Nevertheless, this difference is to be expected, due to the fact that during the weekend there are no normal activities at the University and, consequently, the usage of the CS is significantly reduced.

Figure 5 presents the average power consumption per month in the CS. In this case, it was expected to observe a higher power consumption during the winter season (months 12, 1, and 2) because, in Winter, EV energy consumption increases due to the use of the heating system of the vehicles. However, by analyzing Figure 5 the opposite is observed. There is a reduction in power consumption during the Winter months and, this can be explained by the fact that from December to February there is not much teaching activity in the University, students only have final exams and vacations during that period. Moreover, the month with the lower power consumption is August (month 8), when the University is on vacation.

Regarding the clustering process performed, Figure 6 was created aiming to identify daily power consumption profiles in the CS. Multiple shapes are observed, representing the power consumption behavior for each day available in the dataset. The majority of the power consumption is concentrated between 8h00 and 17h00 (that corresponds to the working hours), but no repetitive behavior is distinguished between days. As mentioned in Section II-A, EV power consumption in CS presents high variability, hence, it is difficult to define the

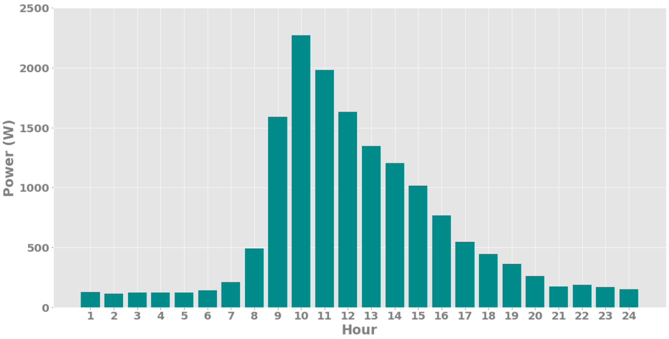

Fig. 3: Average power consumption per hour

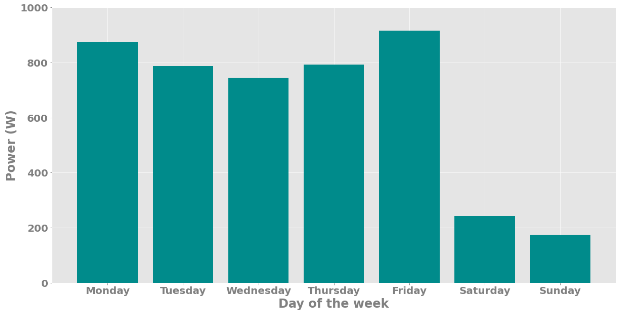

Fig. 4: Average power consumption per day of the week

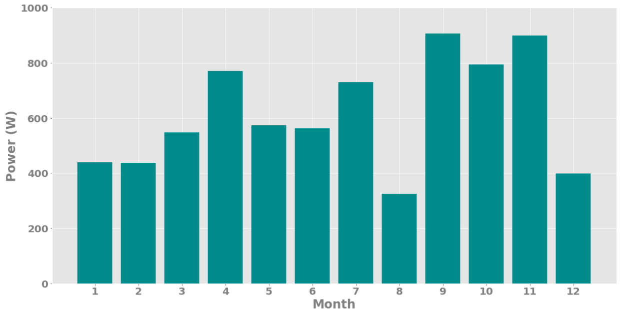

Fig. 5: Average power consumption per month

number of clusters (groups) visually using Figure 6. Therefore, Figure 7 was employed to define the optimal numbers of clusters (K) to use, based on the elbow method. Here again, it is not possible to clearly see the point corresponding to the elbow (bend) of the curve, but it should be a value between 12 to 25 clusters. After several trials doing the forecast using all cluster values in this range, it was determined that the optimal number of clusters is 20 (blue line in the figure).

D. Feature Selection

Figure 8 presents the correlation coefficient of each feature, obtained from the correlation matrix. The correlation matrix provides the relationship between variables on a scale between -1 and 1, where -1 shows a perfect, linear negative correlation, and 1 shows a perfect, linear positive correlation [15]. A negative correlation coefficient demonstrates a connection between two variables in the same way as a positive correlation coefficient, the only difference is that in the negative correlation, the two variables move in the opposite direction, while in the positive correlation, the two variables move in the same direction. The features that present the higher correlation with power consumption are Lag Consumption 1 (0.94), Mean Rolling 4 (0.85), Lag Consumption 4 (0.69), Clusters (0.29), Hour x (-0.29), Mean Rolling 96 (0.22) and Lag Consumption 96 (0.20), based on the correlation matrix.

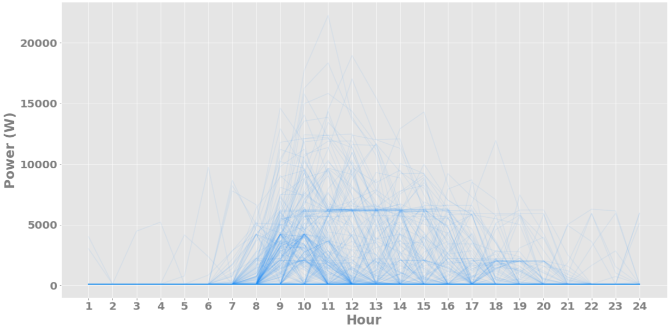

Fig. 6: Identifying daily patterns in EV power consumption profiles

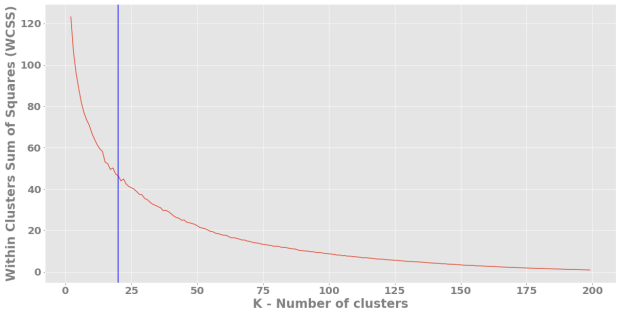

Fig. 7: Elbow method for the optimal number of clusters

Figure 9 presents the feature importance score obtained for each feature using the wrapper method based specifically on RF regression. The features that obtained the higher scores were Lag Consumption 1 (0.92), Clusters (0.036), Mean Rolling 4 (0.013), Lag Consumption 4 (0.012), and Lag Consumption 96 (0.004). To decide the number of features to use in the RF method (next stage), a sensitivity analysis was performed to determine the optimal number of features to maximize the accuracy and reduce the error between power consumption predictions and real values. Different forecasts were created for the same time horizon, changing only the number of features used in the forecast, starting from 1 feature up to 10 features, always considering the features from the higher score to the lower score, which means that features were added one by one from left to right when looking at Figure 9. After this sensitivity analysis, it was determined that the optimal number of features corresponds to the best 3 features; for that reason, only Lag Consumption 1, Clusters, and Mean Rolling 4 should be used to perform EV power consumption forecast using the RF method with this dataset.

E. Forecasting Method

Once the features and the number of features to use in the RF method have been determined, the other important

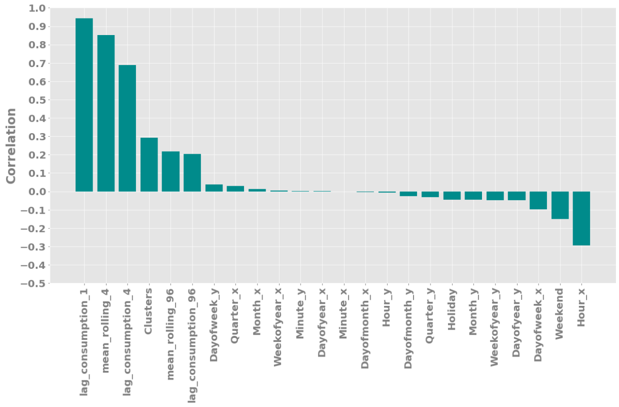

Fig. 8: Feature correlation coefficients

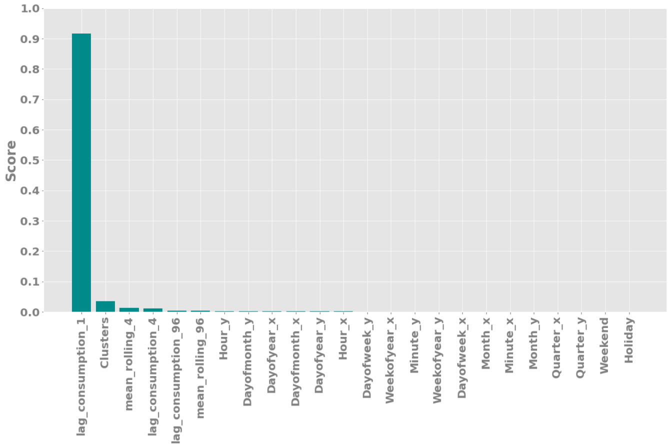

Fig. 9: Feature importance wrapper method

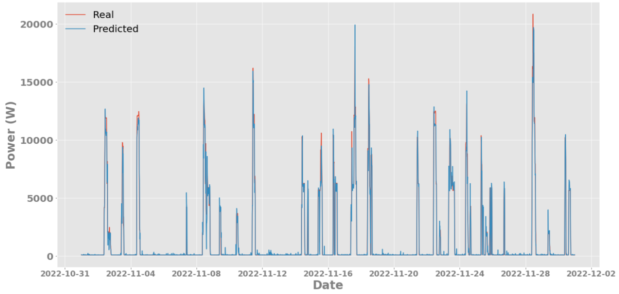

Fig. 10: Real power vs predicted power RF results

parameter to define is the number of days to forecast (time horizon of the forecast). In this case, the predictions, the plot of results, and the error calculations are obtained for one month of forecasting.

Afterward, the RF method is implemented using the following hyperparameters: $n\_estimators = 200$, $min\_samples\_leaf = 3$, $min\_samples\_split = 7$, $max\_depth = 30$, $bootstrap = True$ and $max\_leaf\_nodes = None$. Finally, the EV power consumption predictions for November of 2022 are produced.

The two metrics mentioned in Section II-E are calculated using the predictions obtained from the RF method and the real values of power consumption for November of 2022. The performance achieved for one year of training and one month of forecast is $MAE = 203.23\ W$ and $NRMSE = 3.44\ \%$ There is a high accuracy in the predictions since the NRMSE is only 3.44%. Moreover, when looking at the MAE, the results show that on average, the distance between the prediction value and the real value is 203.23 W. Based on this, the results demonstrate the effectiveness of the computational implementation. Once the performance of the RF method has been evaluated in terms of errors, Figure 10 is produced. Figure 10 compares the forecasting results (in blue) with the real values of power (in red) for the whole forecast period defined (one-month). The red parts observed in Figure 10 are equivalent to the errors that the forecasting model could not capture, but overall the results demonstrate the effectiveness of the proposed method.

IV. Conclusions

A methodological framework for accurate EVs power consumption forecast in charging stations has been proposed in this paper, applying RF method. The framework proposed is able to take advantage of real data from CSs to process through several stages including feature engineering, feature selection, forecasting method, and validation. The results verify that through the framework proposed is possible to forecast EV power consumption and its typical temporal patterns. For instance, it was observed that the EVs peak power consumption during a workday occurs between 10h-11h. Moreover, the results indicate a reduction in EV power consumption by 25% during the weekend and related to the EV CSs usage through the months, it was observed the lower average consumption during August. The forecasting stage by applying a RF algorithm allowed to validate the performance of the method, in which it was obtained a Normalized Root Mean Square Error of 3.44% for a one-month EV power consumption forecast.

References

[1] IEA, "Global ev outlook 2022: Securing supplies for an electric future," in *International Energy Agency: Paris, France*, 2022.

[2] C. P. Guzman, N. Bañol Arias, J. F. Franco, M. J. Rider, and R. Romero, "Enhanced coordination strategy for an aggregator of distributed energy resources participating in the day-ahead reserve market," *Energies*, vol. 13, no. 8, p. 1965, 2020.

[3] J. Zhu, Z. Yang, M. Mourshed, Y. Guo, Y. Zhou, Y. Chang, Y. Wei, and S. Feng, "Electric vehicle charging load forecasting: A comparative study of deep learning approaches," *Energies*, vol. 12, no. 14, 2019.

[4] T. Sousa, T. Soares, H. Morais, R. Castro, and Z. Vale, "Simulated annealing to handle energy and ancillary services joint management considering electric vehicles," *Electric Power Systems Research*, vol. 136, pp. 383–397, 2016.

[5] Y. Mu, J. Wu, N. Jenkins, H. Jia, and C. Wang, "A spatial–temporal model for grid impact analysis of plug-in electric vehicles," *Applied Energy*, vol. 114, pp. 456–465, 2014.

[6] Y. Zheng, Z. Shao, Y. Zhang, and L. Jian, "A systematic methodology for mid-and-long term electric vehicle charging load forecasting: The case study of shenzhen, china," *Sustainable Cities and Society*, vol. 56, p. 102084, 2020.

[7] Y. Lu, Y. Li, D. Xie, E. Wei, X. Bao, H. Chen, and X. Zhong, "The application of improved random forest algorithm on the prediction of electric vehicle charging load," *Energies*, vol. 11, no. 11, p. 3207, 2018.

[8] M. Majidpour, C. Qiu, P. Chu, H. R. Pota, and R. Gadh, "Forecasting the ev charging load based on customer profile or station measurement?" *Applied energy*, vol. 163, pp. 134–141, 2016.

[9] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE access*, vol. 8, pp. 80 716–80 727, 2020.

[10] M. Cui *et al.*, "Introduction to the k-means clustering algorithm based on the elbow method," *Accounting, Auditing and Finance*, 2020.

[11] S. Shahriar, A.-R. Al-Ali, A. H. Osman, S. Dhou, and M. Nijim, "Prediction of ev charging behavior using machine learning," *IEEE Access*, vol. 9, pp. 111 576–111 586, 2021.

[12] K. Bouzoubaa, Y. Taher, and B. Nsiir, "Predicting dos-ddos attacks: Review and evaluation study of feature selection methods based on wrapper process," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021.

[13] A. Antoniadis, S. Lambert-Lacroix, and J.-M. Poggi, "Random forests for global sensitivity analysis: A selective review," *Reliability Engineering & System Safety*, vol. 206, p. 107312, 2021.

[14] IBM, "What is random forest?" [Online]. Available: https://www.ibm.com/topics/random-forest

[15] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: appropriate use and interpretation," *Anesthesia & analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.