Ostermann and Haug *Energy Informatics* (2024) 7:13
https://doi.org/10.1186/s42162-024-00319-1

Energy Informatics

RESEARCH Open Access

Probabilistic forecast of electric vehicle charging demand: analysis of different aggregation levels and energy procurement

Adrian Ostermann[1,2*] and Theodor Haug[2]

*Correspondence: aostermann@ffe.de

[1] School of Engineering and Design, Technische Universität München (TUM), Munich, Germany
[2] FfE (Forschungsstelle für Energiewirtschaft e.V.), Munich, Germany

Abstract

Electric vehicles (EVs) are expected to be vital in transitioning to a low-carbon energy system. However, integrating EVs into the power grid poses significant challenges for grid operators and energy suppliers, especially regarding the uncertainty and variability of EV charging demand. Accurate forecasting of EV charging demand is essential for optimal power system integration, yet previous studies have often only considered point predictions that are inadequate for risk assessment. Therefore, this paper compares different probabilistic forecasting models for the short-term prediction of EV charging demand at various aggregation levels, using a large and novel dataset of over 350,000 charging processes at more than 500 locations across Germany. The performance of both machine learning and deep learning methods is evaluated against a naïve benchmark model, and the impact of data availability on the forecasting models is investigated. Further, the paper examines the effects of forecast accuracy on energy procurement, which has so far received minor attention in the literature. The results show that machine learning methods such as Ada Boosting and Random Forest yield robust results with a normalized root mean square error of 0.42 and 0.41 and a mean absolute scaled error of 0.36 and 0.34 at the highest aggregation level. Furthermore, the results show the influence of different site compositions on the forecast quality and how many charging points are likely to yield a robust forecast. Energy and fleet managers can use the described method to reliably predict the required energy quantities for fleets of sufficient size and procure them at low risk.

Keywords: Electric vehicles, Charging demand, Forecasting, Probabilistic, Machine learning, Deep learning, Aggregation level, Energy procurement

Introduction

Electromobility is in the fast lane. Significant obstacles to electromobility, such as range anxiety, battery life, and sustainability concerns, have been overcome, or work is underway to remove them (ev.energy 2023; Recurrent Auto 2024; Regett 2020; Wohlschlager et al. 2022). Global sales of electric vehicles amounted to around 10 million in 2022, with expected growth of over 30%, corresponding to about 14 million vehicles in 2023 (OECD 2023). Ultimately, electromobility, in combination with smart charging, is helping us to integrate renewable energies into the system and thus reduce transport emissions vital

to meeting the EU's climate neutrality objectives (Duscha et al. 2019). As a result, more electric consumers, such as electric vehicles (EVs) and heat pumps, will be added to the distribution grids. However, grid integration presents the energy sector with significant challenges (Gemassmer et al. 2021; Müller 2023). In contrast to electromobility's needed grid capacity and regulatory challenges, the flexible storage capacity in electric vehicles offers considerable potential for the system and the users (Müller 2023; Kern 2023). Today, electric vehicles can contribute their charging flexibility by integrating renewables through smart charging, reducing costs and CO_2 emissions. In the future, bidirectional electric vehicles can offer more flexibility by charging and discharging the battery. From an energy system perspective, this will happen in a cost-optimal way, with end customers, in turn, earning money or reducing their charging costs through the flexibility they provide (Kern 2023).

The following work is therefore dedicated to comparing different forecasting models for the short-term prediction of the charging energy demand of electric vehicles and the effects of forecast accuracy on energy procurement.

Since time series forecasting has been a significant study area, numerous prediction approaches have been created. It is common to refer to forecasting techniques as statistical or machine learning-based. Nonetheless, because most machine learning algorithms rely on maximum likelihood estimators, they are also statistical. Barker (2020) defines structured and unstructured and notes that both categories still require clarification. Prior knowledge of the forecast's target attributes is necessary for stochastic approaches. On the other hand, regression techniques are more data-driven and do not rely on prior information on the time series (Athiyarath et al. 2020). In the field of electric vehicles, many studies focus on predicting the energy demand of the battery, such as Shen et al. (2022), Mediouni et al. (2022), Chen et al. (2020), on the prediction of charging station occupancy, such as Ostermann et al. (2022), Aghsaee et al. (2023), Hecht et al. (2021), or the prediction of the charging load, using different methods and approaches.

For example, Xydas et al. (2013) proposed a support vector machines (SVM) model to forecast the EV charging demand using travel and driving patterns. They evaluate the accuracy of their method through a Monte Carlo forecasting technique and show that their SVM model has a mean absolute percentage error of 3.69% compared to 8.99% of the Monte Carlo model. Yi et al. (2022) use a long short-term memory (LSTM) model and a deep learning sequence-to-sequence (seq2seq) approach to forecast the monthly commercial EV charging demand. They use real-world datasets from the State of Utah and the City of Los Angeles to validate their models, showing that the seq2seq significantly outperforms other models performing multi-step prediction (Yi et al. 2022). Zhu et al. (2019) compare different deep-learning approaches to forecast the super-short-term charging load of plug-in EVs. Their results of twelve examples on several time steps demonstrate that deep learning methods, primarily LSTM, obtain high accuracy in super-short-term plug-in EV load forecasting (Zhu et al. 2019). As power suppliers do not have information about factors affecting a single car, such as car type, SOC, drive behavior, and destination, Kim and Kim (2021) focused on forecasting daily energy consumption using historical charging data, weather, and day effects. They use statistical methods such as auto-regressive-moving-average (ARMA) and autoregressive integrated moving average (ARIMA) and deep learning methods like the LSTM model using

past values and exogenous variables. Kim and Kim (2021) studied the importance of features on three different geographic scales. At the same time, the discrepancies between the statistical and machine learning approaches were not distinct in the case of micro-scale data with high variability (Kim and Kim 2021). Xie et al. (2011) use neural networks to forecast daily EV charging station load by training the model using similar historical data days. Majidpour et al. (2016) compare forecasting of the EV charging load based on customer profiles and charging station measurements and show that both datasets yield comparable forecasting errors. The Customer profile-based prediction is faster due to less preprocessing. However, this data is prone to privacy invasion (Majidpour et al. 2016). Their modified pattern sequence-based forecasting model has a symmetric mean absolute percentage error of 6.28% for the charging record and 7.85% for the station record. Besides models such as ARIMA and LSTM, Koohfar et al. (2023) use a transformers-based deep learning model to predict EV charging demand. However, they only forecast on a daily resolution. Van Kriekinge et al. (2021) apply a deep neural network to forecast the day-ahead charging demand of EVs in 15 15-min resolution. Additional features such as calendar and weather information reduce the root mean square error (RMSE) and mean absolute error (MAE) by 19.22% and 28.8%, respectively. Their final model has MAE lower than 1 kW for the day ahead horizon. While most studies focus on point forecasts, Buzna et al. (2021) explore a hierarchical probabilistic electric vehicle load forecasting approach at low-level and high-level resolutions. Using real charging data, they demonstrate that their approach outperforms non-hierarchical methods in hour-ahead and day-ahead forecasting EV energy consumption and increases the skill of probabilistic forecasting up to 9.5%. Rathore et al. (2023) use various machine learning models such as Random Forest, XGBoost, and neuronal network models to predict energy consumption by using the historical charging data of the EVs. Their RF and XGBoost models yield the best predictive results.

Energy and fleet managers are faced with the challenge of predicting the quantities of energy required to charge electric vehicles and subsequently procure them as cheaply as possible. However, previous studies have often only considered point forecasts, which need to be revised for a risk assessment. For energy and fleet managers and grid operators, for example, probabilistic forecasts can be advantageous as they show how confident the model is in its prediction. This information can be incorporated into the risk assessment. More than simply predicting the energy quantities probabilistically is required, as they also have to procure them as economically as possible on spot markets. The following paper, therefore, examines different procurement options based on the forecasts and considers the effects of forecast inaccuracies, which still need to be sufficiently addressed in the literature to date. The basis for this work is a new data set with over 350,000 charging processes at more than 500 locations across Germany. So far, the literature has mostly only considered models for small or large fleets and the prediction of point forecasts. This paper compares models based on different charging point numbers and geographical aggregation levels and evaluates the prediction quality based on point and quantile forecasts. To ensure the comparability of the models, we use a naive benchmark model and relate our metrics to its results. In addition, to evaluate the probabilistic predictive quality of the models, we give the pinball score (PS) and the interval score (IS). The forecast is done for the next 24-h horizon with a resolution of 15 min. In

addition, we use walk-forward validation to build robust models that come close to a real-world application. Furthermore, the influence of a shortened data set on the models is examined. In addition, a random composition of the sites is analyzed to provide information on when a group composition yields better prediction results. A detailed analysis of the characteristics of these group compositions is performed. Therefore, we examine the effects of forecast inaccuracies on energy procurement and different procurement strategies in detail.

The paper is structured in the following way: in "Materials and methods" section, we describe the data set and the features used. Further, we explain the methodology of how we developed our models and which metrics we used to evaluate their performance on the task of predicting the charging load. Subsequently, the methodology for analyzing the effects of forecast inaccuracies on energy procurement is presented. "Results" section details the model performance results for various aggregation levels, different training lengths, the effects of random site aggregation, and energy procurement. Finally, "Discussion and conclusion" section presents the discussion of the results and the conclusion.

Materials and methods

We use data for our analysis from the charging and energy management system ChargePilot, developed by The Mobility House. The data consists of over 350,000 charging sessions from over 500 locations or sites. The data begins on 01.06.2022 and covers almost 1 year of charging sessions until 06.05.2023. However, not all sites contain a year's data, as they have only been added over time. Figure 1 shows the methodology for the first part of the paper.

The raw data consists of the attributes listed in Table 1. Not all obtained attributes are listed, only the ones essential for the analysis.

First, we check for errors in the data, such as the plug-out time before the plug-in time, unplausible charging powers, or missing values. However, the data did not show any of those errors. Next, we transform the charging sessions into a time series format of 15-min resolution per charger while we round the plug-in and out time to the nearest quarterly hour. We assume that the electric vehicle is charged at its maximum charging power upon being plugged in, and the charging power is then reduced to zero once the targeted energy consumption is reached. Adding up the time series of every charger

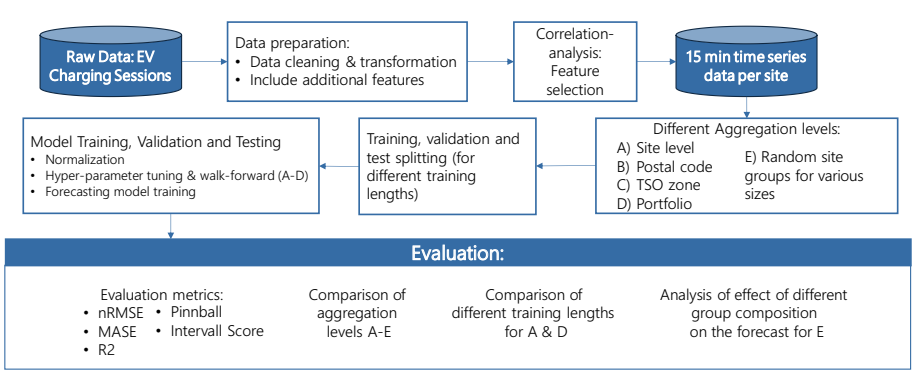

Fig. 1 Methodology for comparison of different models for various aggregation levels

Table 1 Data fields of the raw charging session data

Attribute	Description
Timestamp	Timestamp in UTC
Plugin time	The time when the EV is connected
Plug out time	The time when the EV is disconnected
Duration	Plugin duration in h
Site ID	Site identification
Number of chargers	The number of chargers for the site
Number of charging points	The number of charging points for the site
Site fuse limit	The fuse limit of the site in W
Postal Code	The first two digits of the postal code
TSO zone	Transmission grid operator zone
Charge power max	Maximal charge power of the session in W
Energy consumed	Charged energy during the session in Wh

Table 2 Included features

Attribute	Description
Power rolling week mean	Rolling mean considering a lag of 1 week, using a rolling window size of 1 week
Power rolling day mean	Rolling mean considering a lag of 1 day using a rolling window size of 1 day
Power week lag 3 h mean	Rolling means considering a 1-week lag, using a rolling window size of 3 h
Power day lag 3 h mean	Rolling mean considering a lag of 1 day, using a rolling window size of 3 h
Site fuse limit	The fuse limit of the site in W
Number of charging points	The number of charging points for the site
Holiday	Categorical encoded; 1 if German holiday, 0 if not
Weekday	Categorical encoded weekdays
Sine time of year	Encoded cyclical continuous feature time of year with sine
Cosine time of year	Encoded cyclical continuous feature time of year with cosine
Sine time of day	Encoded cyclical continuous feature time of day with sine
Cosine time of day	Encoded cyclical continuous feature time of day with cosine

from one site yields the 15-min resolution charging power time series per site, which serves as our target variable. We use the following features for each of these as input for the models described in Table 2.

The first two added features facilitate identifying trends and patterns in the data over a weekly or daily time frame while accounting for the specified lag. The third and fourth features help capture trends and patterns in the data over specified time intervals. We initially included several lag features. However, correlation analysis has shown that the ones in Table 2 had the highest correlation regarding the target variable. If the corresponding date is a public holiday in Germany, the feature is assigned 1; otherwise, it is 0. Furthermore, we extracted the days of the week from the timestamps (Mon–Sun) and used one-hot encoding to convert the categorical variables into numerical ones. Jump discontinuities are a problem for machine learning algorithms using cyclical data. Therefore, we took cyclic feature encoding into account for periodic patterns in the time of year and time of day features in the final step of data preparation. A simple method is dividing the features into sine and cosine parts. Since we use a rolling 1-week feature, the input data for the models starts on 08.06.2022. We used German weather data as

an input feature on a subsample of the data. However, this did add additional value nor showed forecast accuracy improvement. Other studies suggest using weather data as an input feature. Therefore, using regional weather data might further improve the forecast.

To compare the effect of various fleet sizes and aggregation levels on the prediction quality, we aggregate the time series per (A) site on (B) postal code, (C) TSO zone, and (D) portfolio, meaning all combined levels. For (A) and (B), we chose five different sites and five postal codes. The five sites have 3, 4, 8, 14, and 145 charging points. Further, for (E), we randomly sample from all sites to investigate the effect of random group compositions and different fleet sizes further. This is done for various group sizes in the range 10, 15, 20, 25, 30, 40, 50, 75, and 100. The random sampling is done 100 times per group size.

Next, we split the data into training, validation, and test sets in the following ratios: 75%, 15%, and 10%. To investigate the effect of less available data, we limit the length of the data set to the following starting dates: 01.09.2022, 01.12.2022, and 01.02.2023. By manually setting the start of the test set to 04.04.2023, we ensure that the models trained with a shortened data set are compared on the same test set. However, this analysis is limited to the aggregation levels (A)–(D) due to computational restrictions. Classical tree-based machine learning models could be better at extrapolating unseen data. To account for this, we normalize the charging power per charge point by dividing it by the number of charge points to accommodate the trend or increase in the energy charged by additional charge points and sites. In addition, this allows us to visualize the charging power, as otherwise, there would be concerns about commercial confidentiality.

As a benchmark model, we use a naïve model (Naïve WD Mean), which takes the average of the weekday in the according quarterly hour. We use the following machine and deep learning models for our analysis: Linear Regression (LinR), Bagging, Gradient Boosting (GradientB), Ada Boosting (Ada), Random Forest (RF), convolutional neural network (CNN), neural network (NN), and long short-term memory (LSTM). The underlying concepts of the models are described in detail in Breiman (2001), Freund and Schapire (1996), Friedman (2002), Hatalis et al. (2017), Wang and Raj (2017), Sharkawy (2020). The models were selected to encompass various machine learning and deep algorithms. While LinR represents a relatively simple linear model, RF and Bagging are non-linear tree-based ensemble learning techniques. Further, we include the non-linear tree-based models Ada and GradientB, which use boosting. The NN architecture represents one of the simpler deep learning models. CNNs were first developed to analyze pictures; they can also be used to predict time series. Due to their particular architecture, LSTMs know when to memorise and when to ignore past information and therefore, are widely used in time series forecasting. We used the Python sci-kit learn implementations for the machine learning models and implemented the deep learning models in PyTorch (Pedregosa et al. 2012; Paszke et al. 2019). Deep learning models have the innate capacity to recognize and retain patterns over a wide range of time scales, unlike typical machine learning models that could depend on manually designed lag characteristics to account for temporal patterns. Because of several attributes, deep learning models can independently manage temporal dependencies and create additional features. The basic NN model comprises three linear layers with dropout applied after the initial layer. ReLU functions are

activation functions between the layers, a pattern retained in subsequent models. The LSTM model features an LSTM layer with dropout, succeeded by two linear layers. In the CNN model, two convolutional layers with a kernel size of four are followed by a max-pooling layer, concluding with two linear layers.

Due to the temporal structure of the data, hyperparameter tuning—a critical step in maximizing the performance of machine learning models—becomes more difficult in the context of time series forecasting. In time series forecasting, the walk-forward validation technique is frequently employed to mimic real-world situations in which the model is trained on past data and subsequently evaluated on future data points. First, the model is trained using historical data. As our prediction horizon is 24 h, we predict the first day of the validation set and compare the predicted value with the actual value for the current time step using the performance metric mean squared error. Next, we move the time window to 24 h and update the training set with the actual values. For the subsequent time step, we repeat the training and prediction procedure and follow this step-by-step procedure, updating the model iteratively and assessing its effectiveness at every turn. This procedure is also used for the testing. We apply different combinations of hyperparameters using a grid search by validating each set of hyperparameters through the walk-forward validation process, calculating the average performance across all time steps. The combination of hyperparameters that produces the best overall performance is then selected. We test our final model on the test set not used for hyperparameter optimization to assess the model's generalization performance. The hyperparameters used for the grid search are listed in Table 5. Due to computational limitations, we apply this procedure only to aggregation levels (A)–(D). Furthermore, the deep learning models are not subjected to hyperparameter tuning due to computational costs, limiting their full potential. We use the 5 and 95% quantiles to calculate the quantiles. In boosting models that optimize individual estimators, quantile predictions were derived by extracting quantiles from the estimators. This approach is also suitable for Ada, where the quantiles need to consider the weights of the estimators. However, this quantile estimation method is not feasible for models optimizing the entire ensemble based on a specified loss function, such as Gradient Boosting. In these instances, the model must be trained with a different loss function, and the pinball loss was chosen for making quantile predictions. This is true for deep learning models as well.

We evaluate the models based on the following evaluation metrics: root mean squared error (RMSE), normalized RMSE (nRMSE), mean absolute error (MAE), mean fundamental scaled error (MASE), R^2, pinball score (PS), and interval score.

The Mean Absolute Error (MAE), is one of the most used error measurements and is referred to as (Hyndman and Athanasopoulos 2021):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=0}^{n} |\widehat{y}_i - y_i|, \tag{1}$$

where n is the number of observations and y_i defines the actual value, and $\hat{y}_i$ is the model's prediction. Another frequently used metric is the RMSE, which is defined as (Hyndman and Athanasopoulos 2021):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}\left(\widehat{y}_i - y_i\right)^2}. \tag{2}$$

The nRMSE is a frequent statistic for assessing a predictive model's accuracy, especially in regression analysis or forecasting. It offers a relative measurement of the error between expected and actual values and is a normalized form of RMSE. The MASE is often used to evaluate a forecasting model's accuracy and is the normalized form of the mean absolute error. We normalize based on our Naïve WD Mean model, meaning that values above 1 are worse than the benchmark model and below one are better than the benchmark model. This makes it easier to compare our model's performances. The R^2 is frequently employed to measure the regression model's goodness of fit since it shows how well the model's predictions correspond with the actual data, where one is a perfect fit. Zero indicates that the model does not explain any of the variability in the target variable. According to James et al. (2021), R2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=0}^{n}\left(\widehat{y}_i - y_i\right)^2}{\sum_{i=0}^{n}\left(\bar{y} - y_i\right)^2} \tag{3}$$

where $\bar{y}$ is the mean of the target. Let τ be the target quantile and $\widehat{q}_{i,\tau}$, the quantile forecast, then the $PS_{\tau,}$ which evaluates the upper and lower quantile separately, and according to Koenker and Machado (1999), can be defined as:

$$PS_\tau(y_i, \widehat{q}_{i,\tau}) = \begin{cases} \left(y_i - \widehat{q}_{i,\tau}\right)\tau, & \text{if } y_i \geq \widehat{q}_{i,\tau} \\ \left(\widehat{q}_{i,\tau} - y_i\right)(1-\tau), & \text{if } \widehat{q}_{i,\tau} \geq y_i \end{cases} \tag{4}$$

The PS is a metric that quantifies the difference between the actual and anticipated quantile value, weighted according to the quantile level. It evaluates the prediction interval's accuracy, with various quantiles generating distinct values. Better model performance is indicated by a lower PS, which penalizes deviations from actual values within the predicted quantile range less severely. Another metric to assess probability forecasts is the Interval Score (IS), which considers the width of the prediction interval, also known as sharpness (Hatalis et al. 2017). The IS is typically used with the PS to assess the prediction model's total predictive uncertainty because it cannot adequately characterize the dependability of the prediction interval (Hatalis et al. 2017). The narrower the interval and closer to the actual observations, the smaller the interval score.

Figure 2 shows the methodology for the second part of this work, in which we examine the effects of forecast inaccuracies and different trading strategies on energy procurement based on the German energy market. For the analysis, we use the German hourly day-ahead energy price, and for the intraday prices, we use the ID1 and ID3 prices from the ENTSO-E Transparency Platform (2024).

Based on the time series for the entire portfolio, we create a model for day-ahead (DA) procurement and two models for intraday procurement. The models differ in terms of their input features, the starting point of the forecast, the forecast horizon, and the resolution. In Germany, the day-ahead auction closes at noon for the next day, whereby hourly products can be traded. In continuous intraday day trading, 15-min products can be traded up to 5 min before the start of delivery. The DA model has an offset of 12 h to

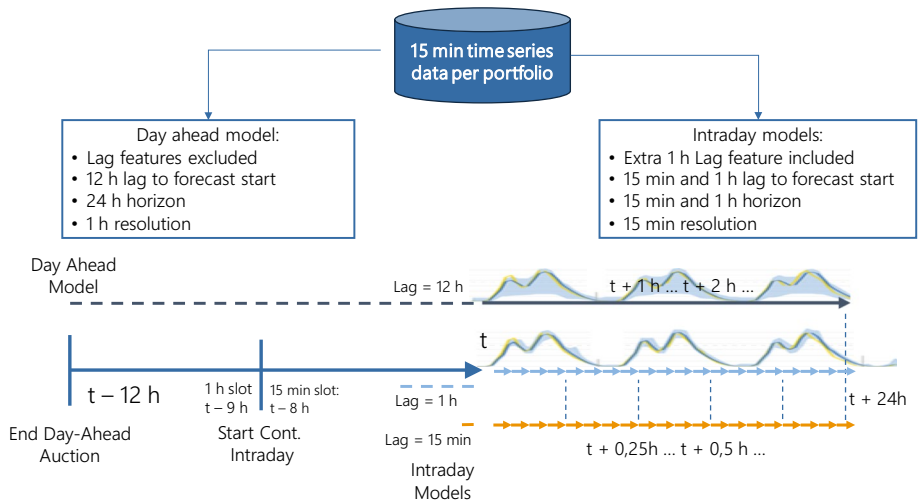

Fig. 2 Methodology for analyzing effects of forecast inaccuracies on energy procurement

the start of the forecast, a forecast horizon of 24 h, and an hourly resolution. The day-ahead model is designed so that it does not contain any lag features that provide the model with information from future observations, thereby preventing data leakage. The 12 h offset results from the gate closure of the day-ahead market, where we assume that the forecast and actual trading are instantaneous. The hourly resolution results from the hourly traded products. Thus, the day ahead model represents the best possible forecast to buy the hourly products for the next 24 h on the day ahead market. The intraday models have (a) a lag of 1 h, a forecast horizon of 1 h, and (b) a lag of 15 min and a forecast horizon of 15 min. Both intraday models have an additional 1-h lag feature as input and a resolution of 15 min. The intraday model b was chosen because it represents the best possible model with a resolution of 15 min. Since our time series is based on a 15-min resolution, it makes no difference whether we assume a 5-min or 15-min lag, as the last point in time or the last actual value is available to the model as information. The intraday model a was chosen to match the day ahead's hourly products with the hourly forecast horizon and to have a worse comparison model than b. Since the model has a lag of 1 h, it has less information available than the b model. The energy procurement process is as follows. We buy the energy amount forecasted from the DA model for various quantiles for the respective hours for the next day to the given day-ahead price. Next, we buy or sell (a) 1 h or (b) 15 min before delivery, depending on the intraday model a or b the difference from the predicted intraday value compared to the DA forecast to balance the energy according to the ID1 and ID3 intraday-price. The ID1 index is the weighted average price of all continuous trades completed within the last trading hour, and for the ID3, the previous 3 h are taken into account. Thus, we get the total energy procurement costs on the spot market for the energy per charge point in the portfolio during the test set. Afterward, we compare the actual values with the predicted ones to assess how much balancing energy would be needed to cover the difference. The balancing group managers in Germany are obliged to minimize their balancing group deviation. They must refrain from strategically exploiting the balancing energy; otherwise, there is a risk of high penalties. Therefore, we only examine the effects on the amount of energy that

must be balanced by the balancing energy, which should be as low as possible and not on the associated costs or revenues resulting from the balancing energy price and the amount of energy. However, in other European Countries, incorporating the balancing market into the strategy would be possible.

Results

The first part of this section reports the results of the various models based on the different aggregation levels. The second part of this section describes the outcomes of the energy procurement.

Model performance for various aggregation levels

This section compares the results of the different forecasting models described in the previous sections. The presented results correspond to the model's performance on the test set. The test set starts on 04.04.2023 at 00:00. It ends on 06.05.2023 at 23:45. The predictions are made at midnight with a 24-h horizon and 15-min resolution. Figure 3 shows the power per charging point in kW for (A) a random site, (B) for all sites in the zip code, (C) in the TSO zone, and (D) aggregated for all sites during the test set.

The power per charging point for the site displayed on top has high peaks during the day on the weekdays. During the night and the weekends, no charging events occur. The site shown in A, is also dominant at the ZIP code level B, as the time series are similar but not identical. For example, small charging events can be seen on Mondays and Sundays. The aggregation at the TSO level differs significantly from that at the ZIP code level, where charging power is already very regular. Comparing the TSO zone with the aggregation of all sites, it is noticeable that the charging power is smoothed out even further in the afternoons and on the weekends. Further, the factor of simultaneousness decreases with higher aggregation. Assuming an average charging capacity of 11 kW, the factor of simultaneousness is around 10%.

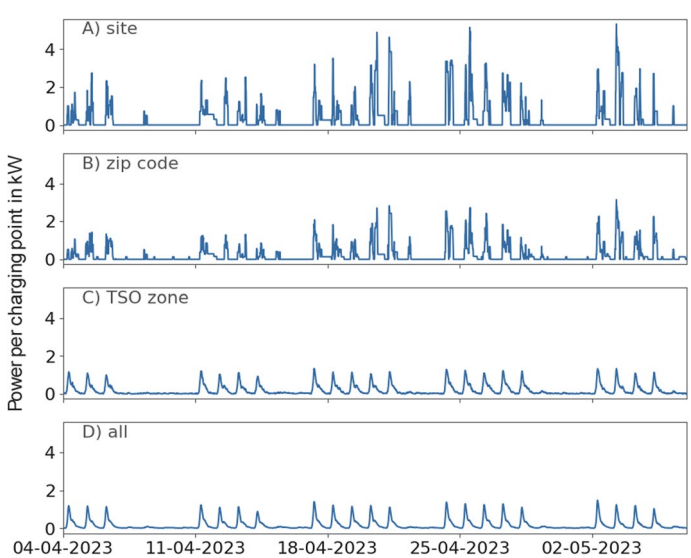

Fig. 3 Power per charging point for different aggregation levels during the test set

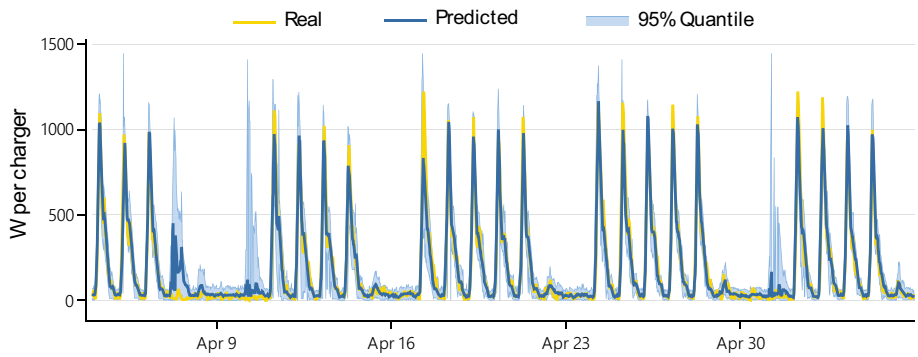

Fig. 4 Real power (yellow), predicted power (blue), and 95% quantile (light blue) per charger in W for the Ada model for the test set

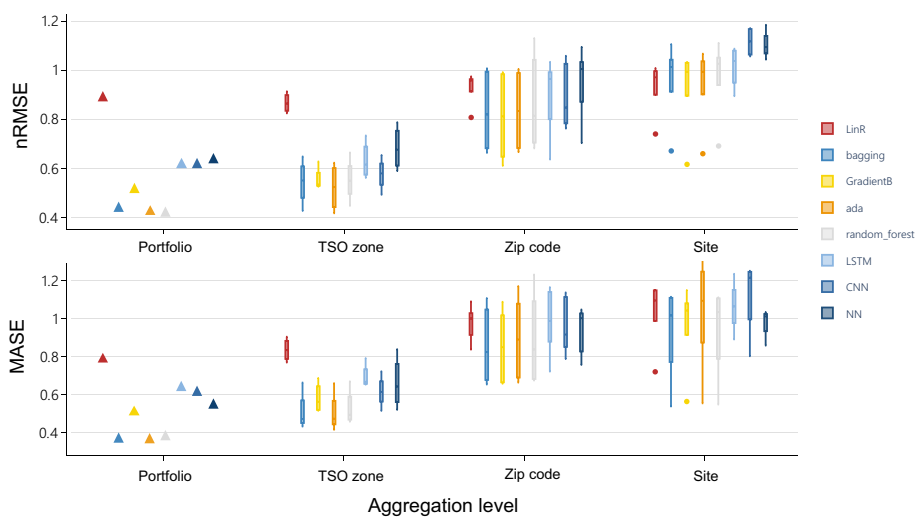

Fig. 5 Boxplot for nRMSE (top) and MASE (bottom) for different aggregation levels and models

Figure 4 shows the actual power per charger in W for the entire portfolio in yellow, the predicted power of the Ada model in blue, and the 95% quantile in light blue for the test set. Overall, the forecast follows the real power, and the 95% quantile also follows the real power with a low dispersion with a few exceptions. Large quantile deviations on 10.04.2023 and 01.05.2023 are noticeable. Nevertheless, the prediction is close to the actual value. This can be explained by the fact that both days are national holidays in Germany, and this information is available to the models as a feature. However, as we have no annual data available, the model is uncertain, as can be seen from the deflection of the quantile.

Figure 5 shows the nRMSE (top) and the MASE (bottom) as a boxplot for the different aggregation levels and the various models. The models are arranged from left to right, as shown in the legend from top to bottom.

Almost all models have values above 1 for the MASE and nRMSE for both the individual sites and the zip codes, which means that the benchmark model is better in some cases. Looking at the MASE and the nRMSE of the models for the TSO zones and the entire portfolio, it becomes clear that aggregation significantly increases the prediction

quality. Bagging, Ada, and RF have the most favorable MASE and nRMSE. Table 3 lists the metrics for the models for the portfolio. The best value is marked in bold, and the second best is underlined. The other model metrics for the different aggregation levels are listed in Tables 6, 7, 8.

The ensemble models Bagging, Ada, and RF excel in accurate point predictions, reflected by low nRMSE and MASE, and exhibit robust quantile prediction, as evidenced by the comparatively low PS. The Ada model has the lowest rRMSE with 0.355, the lowest PS for the high Quantile with 2.767, and the RF the lowest MASE with 0.411 and R2 with 0.954. Furthermore, these three provide narrower prediction intervals, as indicated by lower IS. The three deep learning models demonstrate moderate performance in point prediction, with the CNN having the lowest amongst them with 0.538, which is significantly better than the benchmark and better than the LinR. However, they are worse than the ensemble models, indicating a potential need for further refinement in capturing underlying patterns. Since we have chosen our hyperparameters to cover as wide a range as possible, but not every model has an oversized hyperparameter space, some models may be at a disadvantage. Specifically, looking at Table 5, this can be seen in the number of estimator parameters; for example, our RF model has a range of 500, 750, and 1000, GradientB 50, 100, 150, 200, 250, 300, 350, and Ada 10, 50, and 100. Large numbers of estimators often lead to overfitting. The models may have yet to reach their full potential and can achieve even improved results depending on the aggregation level. The fact that Ada and RF performed best in the overall portfolio does not necessarily mean that these two are generally the best. In particular, a variety of test pipelines that test different feature compositions, such as weather data, different lag features, district specific holidays, but also, for example, different encoding strategies for the cycling time features and weekdays, in combination with larger hyperparameters, would possibly improve the performance of all models. But these steps were limited by the lack of computational power. Further, the deep learning models face challenges in accurately predicting quantiles, reflected by higher PS. As mentioned in "Materials and methods" section, we did not fully fine-tune and optimize the deep learning architectures; therefore, doing so may enhance their predictive capabilities. Comparing the results of the individual sites (see Table 5), it is noticeable that only the values of site E with over 100 charging points are significantly better than those of the naive model. While sites A, B,

Table 3 Metrics for the models based on portfolio level

Model	nRMSE*	MASE*	R2**	PS low Q***	PS high Q***	IS***
LinR	0.883	0.777	0.788	–	–	–
Bagging	0.432	<u>0.358</u>	0.949	**1.815**	<u>2.815</u>	<u>133.882</u>
GradientB	0.508	0.498	0.930	3.526	4.156	203.921
Ada	<u>0.418</u>	**0.355**	<u>0.952</u>	<u>1.866</u>	**2.767**	147.959
Random Forest	**0.411**	0.374	**0.954**	2.843	4.474	**95.621**
LSTM	0.610	0.603	0.899	37.899	76.411	150.056
CNN	0.630	0.538	0.892	42.207	71.183	140.144
NN	0.609	0.627	0.899	115.093	95.773	247.339

Best result bold, second underlined

*Unitless, **in %, ***in W

and D all have a MASE greater than 1, some models at site C achieve a better value than the benchmark. It is interesting to note that, on the one hand, the site has fewer charging points than site D. On the other hand, the CNN model for the MASE and the LSTM for the nRMSE achieve the best values, although they perform worse than the machine learning models at the other aggregation levels.

In conclusion, Bagging, RF, and Ada perform best; they are particularly good at quantile estimation and point prediction with narrow prediction intervals. The deep learning models (LSTM, CNN, and NN) show moderate performance with room for improvement, especially regarding quantile prediction and narrow prediction interval widths. Comparing the results of Tables 6, 7, 8, finer aggregation levels (like zip code and site level) tend to pose more challenges for the models. The ensemble models Bagging, GradientB, and Ada maintain their robust performance across different aggregation levels, while RF shows performance degradation, indicating challenges in handling finer-grained data. The same is true for the deep learning models, as they exhibit more sensitivity to data granularity. The choice of the best-performing model may depend on the specific aggregation level and the trade-off between computational efficiency and predictive accuracy.

Model performance for different training lengths

To investigate the influence of available data on the prediction performance, we train the models on different data lengths. Figure 6 displays the nRMSE (top) and MASE (bottom) as a boxplot for the models over all aggregation levels for different start dates of the training set. The start date 08.06.2022 represents the results for the complete data set described in "Model performance for various aggregation levels" section. As mentioned in "Materials and methods" section the test set is the same for all data lengths and models to ensure comparability.

Comparing the different start dates, it is initially noticeable that the deep learning models perform particularly poorly with a start date of 01.02.2023. All models perform better with more data, although the dynamics and behavioral patterns can

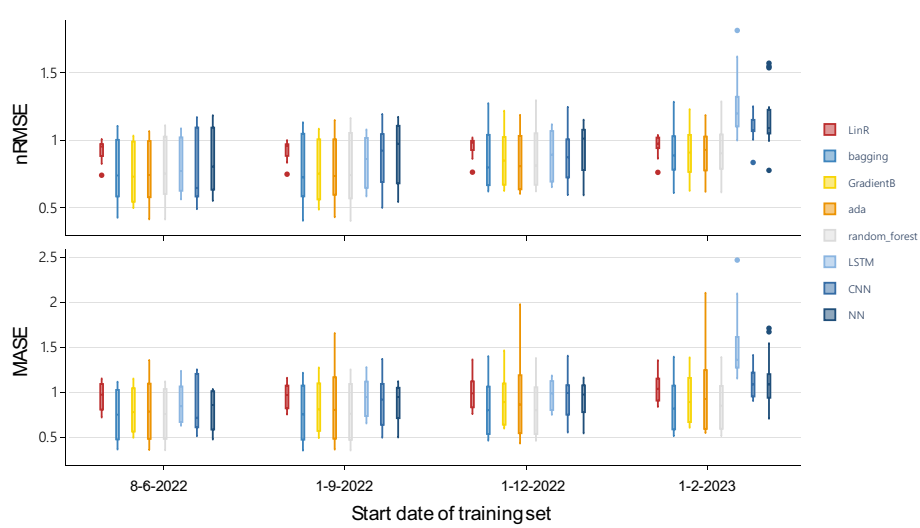

Fig. 6 Boxplot for nRMSE (top) and MASE (bottom) for different training lengths

change due to adding charging stations at individual sites. However, a certain saturation can be observed, as the minimum values of the models with the start date 08.06.2022 do not improve significantly compared to 01.09.2022. It would be interesting to investigate whether a further significant improvement occurs if more historical data is added, for example, to map seasonality. Unfortunately, however, we do not have more data. Shorter data sets can lead to overfitting, but the training is less computationally expensive and requires fewer computational resources. This is of particular importance when implementing real applications. Overall, it remains a case-by-case decision whether the simple benchmark model is superior to machine learning models in the case of limited historical data.

Analysis of random site aggregation

By forming random groups of different sizes from all the sites and aggregating them, we further investigate the influence of different fleet sizes and the number of charging points on the prediction quality. The random group sizes consist of 10, 15, 20, 25, 30, 40, 50, 75, and 100 sites, and the random draw is repeated 100 times. We thus formed a total of 36,500 different group compositions. We formed random groups of different sizes from all the sites, aggregated them, and then used them as input for the models to investigate the influence of fleet size and the number of charging points on the prediction quality. We formed the random group sizes of 10, 15, 20, 25, 30, 40, 50, 75, and 100 and repeated the random draw 100 times. We thus formed a total of 36,500 different group compositions. Due to the large number, we did not use hyperparameter tuning for the randomly composed time series. We only used the Ada model for the analysis based on the previous results, as it was among the best. Figure 7 shows the nRMSE (a) and MASE (b) for all different group compositions according to their number of charging points and frequency distribution.

Looking at the frequency distribution, it becomes clear that the benchmark model is only superior to the Ada model in a few exceptional cases when grouped according to

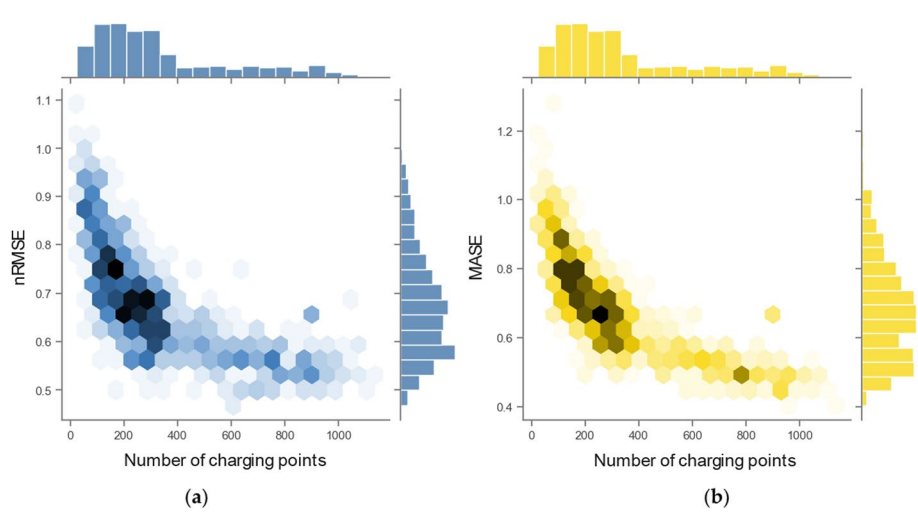

Fig. 7 Hexplot for nRMSE (**a**) and MASE (**b**) for different group compositions according to their number of charging points

the abovementioned quantities. Most groups have several charging points between 150 and 400 and a MASE or nRMSE of 0.8 to 0.55. Although the MASE and nRMSE drop significantly with increasing charging points to around 0.5, a few group compositions perform poorly despite many charging points. The group with a MASE and nRMSE of about 0.65 at 900 charging points is particularly striking. At the same time, however, random group compositions achieve an nRMSE or MASE of 0.55 or less with just around 200 charging points.

The question, therefore, arises as to which the remarkably predictable groups exhibit characteristics and whether these can be determined in advance. Charging point operators or aggregators could ensure their balancing groups are grouped to meet this characteristic. To investigate this question, we used a Wavelet analysis to examine the time series of a group composition with ten sites with an nRMSE of over one and one with an nRMSE of 0.55. Wavelet analysis is a mathematical method that breaks down signals or functions into their frequency components for analysis. Instead of conventional Fourier analysis, Wavelet analysis records both frequency and temporal localization. It analyzes data at various scales and reveals features at varied resolutions by using tiny, wave-shaped functions known as wavelets. This allows the identification of fleeting features in the data. Wavelet analysis is a potent tool for deciphering and obtaining information from complicated signals. It has applications in many domains, such as signal processing, image analysis, and compression methods. The wavelet analysis also has the advantage that the dynamics of charging behavior, which change to some extent over time, are visible. Figure 8 displays the resulting wavelet plot of the wavelet analysis while the y-axis depicts the frequency in days, and the x-axis represents the time for the group of tens with a nRMSE of 0.55 (a) and above one (b). The yellow areas highlight the occurrence at specific times and frequencies and illustrate how the frequencies contribute to the signal.

The left wavelet plot shows a strong periodicity of 1 day and 7 days. It can also be seen that the signal is weaker at the beginning and increases from September onwards. In addition, the period around Christmas is recognizable in which the periodicity visibly decreases. The right-hand wavelet plot initially shows little periodicity, especially around November 2022. From the end of January, an increased daily and weekly periodicity can be seen, but more clearly separated than in plot (a). This is because some of the sites were only integrated into the load management system at this time. The displayed wavelet plot thus enables a quick and easy visualization to analyze possible patterns, transients, and frequency components within the signal across different scales. The plot also shows that a shorter data set might lead to better results for the group composition (b).

It is advantageous for energy providers if the portfolio is as extensive as possible. However, energy and fleet managers require a location-specific forecast. In the case of a charging management system, a charging station-specific forecast is required. Whether

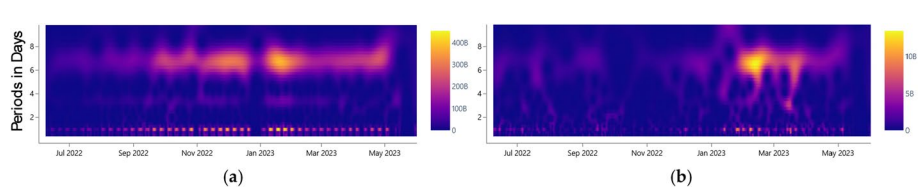

Fig. 8 Wavelet plot for a group of ten sites with a nRMSE of 0.55 (**a**) and above one (**b**)

a single location or a group composition can be predicted well can be estimated by looking at the wavelet plot, as described above. On the other hand, the correlation between the lag features and the charging load can be determined with the help of a correlation analysis. If these correlate strongly, it strongly indicates that the location or group composition can be predicted with improved accuracy.

Energy procurement

As described in "Materials and methods" section, we use two intraday models and a day-ahead model with a 12-h offset and 60-min resolution to examine energy procurement on the day-ahead and intraday markets. For this analysis, we use the median of the model predictions. The testing period is the same as mentioned in "Model performance for various aggregation levels" section. The day-ahead model has an MAE of 26.39 W/charger and an RMSE of 53.30 W/charger. The intraday model with an offset of 60 min (intra60) and resolution of 15 min has an MAE of 15.67 W/charger and an RMSE of 35.77 W/charger, while the intraday model with an offset of 15 min (intra15) and resolution of 15 min has an MAE of 13.94 W/charger and an RMSE of 29.33 W/charger.

Figure 9 shows the day ahead price in red, the intraday ID1 price in blue, and the balancing price in orange.

The upper graph shows only the day-ahead and intraday price, while the lower graph also shows the balancing price, with the y-axis scaled differently. The day-ahead price shows no significant outliers and ranges between −8.82 and 207.92 €/MWh. The intraday price, on the other hand, shows a significantly more extensive range of −1323.6 and 595.71 €/MWh. Looking at the chart below, it is clear that the balancing price shows even more significant outliers. Here, the minimum value is −6082.56, and the maximum is 9853.34 €/MWh. For

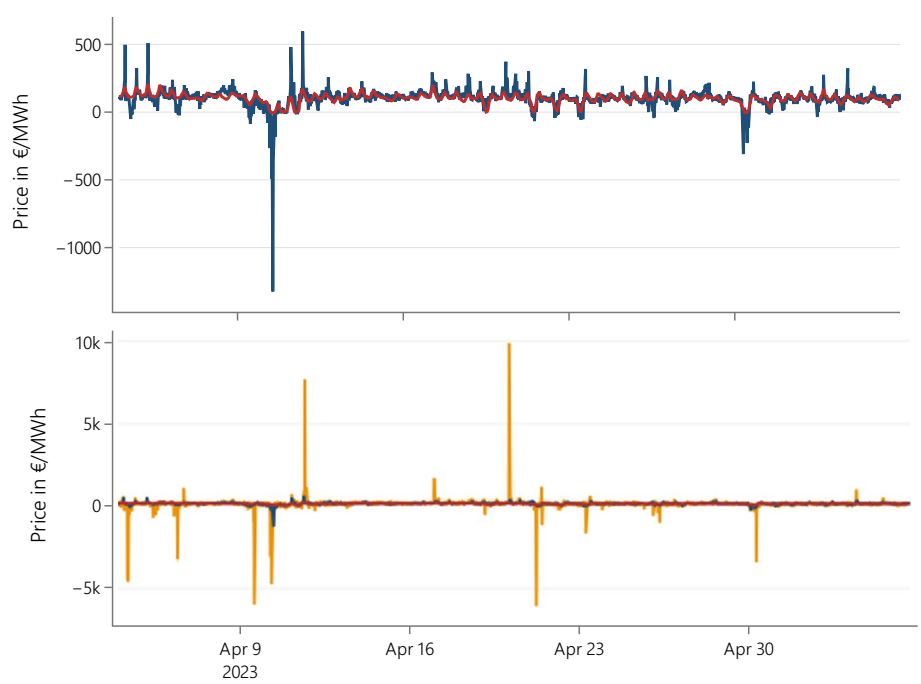

Fig. 9 Day-ahead, intraday ID1, and balancing price for the test set

example, if a trader had bought the required charging energy on the day-ahead market on 10.04.2023 at 11:00, he would have received €8.82 for each MWh. On the intraday market, it would be €252.96 at 11:00, and €1323.6 at 11:45 for each MWh consumed. At the same time, a MWh would have cost €65.54 on the day-ahead market and €14.71 on the intraday market at 17:00 on the same day. On 11.04.2023 around 06:00 a.m., however, the day-ahead price is significantly below the intraday price. The price range here is just under €370/MWh. At the same time, the balancing price was just under €690/MWh—every additional MWh required and not previously procured leads to considerable additional costs. The gap becomes even more extreme at 19:00, as the balancing price here is just under €7600/MWh, which is around 54 times higher than the day-ahead price at the same time. If the forecast is significantly too low at these times and too little energy is procured, this leads to considerable additional costs. Procurement purely on the intraday market would lead to significantly higher costs here. These enormous fluctuations clearly show the potential of smart energy procurement and the additional flexibilization of loads, for example, by postponing charging processes. For the entire portfolio, the procured amount results in 144.96 kWh of charging energy per charger in the test period. First, we assume that we have perfect foresight and purchase all the necessary charging energy once entirely on the day-ahead market and once completely on the intraday market at the ID1 and ID3 prices. This results in the following electricity costs for charging per charger: day-ahead € 15.87, intraday ID1 € 16.24, and Intraday ID3 € 16.17. Table 4 lists the results. This makes it clear that for the test period, it would be most cost-effective on average to procure all energy on the day-ahead market if one knew in advance exactly how much energy one would need, which in reality is not the case when procuring charging energy for a portfolio.

In the test period, procuring as little energy as possible would be financially advantageous, as the balancing energy price is primarily negative. As explained in "Materials and methods" section, it is not permitted to systematically and deliberately exploit balancing energy to gain financial advantages in Germany. If this were permitted, it would be possible to speculate on the negative price peaks with a certain degree of risk by procuring very little energy at this time. The negative price peaks would lead to such large profits that it would be cheaper overall than procuring on the day-ahead market, which is valid for the test set and the whole data period. Therefore, in the following, we look not at the total costs, including the balancing price, but at the influence of the two intraday models on balancing energy, as the amount must be minimal. The costs for procurement on the day-ahead market due to the day-ahead forecast per charger are € 15.637. After the purchase or sale of the deviation from the forecast of the intra60 model, the additional

Table 4 Results for energy procurement

Model	Day-ahead costs*	Intraday ID1 and ID3 costs*	Total costs ID1 and ID3*	Balancing energy**
Intra60	15.637	0.227 | 0.211	15.865 | 15.850	12.40
Intra15	15.637	0.360 | 0.340	16.000 | 15.978	9.74
Intra15-5% improved	15.637	0.363 | 0.343	16.001 | 15.980	9.25
Intra15-10% improved	15.637	0.366 | 0.346	16.004 | 15.983	8.76
Intra15-15% improved	15.637	0.369 | 0.349	16.007 | 15.986	8.28
Intra15-20% improved	15.637	0.371 | 0.351	16.009 | 15.989	7.79

*in € per charger, **in kWh per charger

costs per charger are € 0.227 for the ID1 and € 0.211 for the ID3, resulting in energy costs per charger of € 15.865 and € 15.850. As a result of the more accurate forecast of the intra15 model, more energy has to be procured, ultimately leading to higher costs. For the intra15 model, the costs per charger amount to € 0.36 for ID1 and € 0.34 for ID3 totaling € 16.00 and € 15.98. However, balancing energy quantities per charger of 12.40 kWh are required for the intra60 model and only 9.74 kWh for the intra15. To investigate the impact of a more accurate intraday forecast on costs and balancing energy, we reduced the error of the intra15 forecast by 5, 10, 15, and 20% compared to the actual value. The following costs in € per charger for intraday ID1 procurement result for the adjustment: 0.363, 0.366, 0.369, and 0.371. The total balancing energy in kWh per charger amounts to: 9.3, 8.8, 8.3, and 7.8. Thus, a reduction in the necessary balancing energy of 20% only increases intraday procurement by 3%. About the total costs, the additional costs are even less than 0.1%. In addition, this avoids the risk of compensating for the sometimes highly high balancing energy prices, such as on 20.04.2024. The average expected value of the forecast must be used for intraday procurement, as otherwise there is a strategic over- or under-procurement and thus an exploitation of the balancing energy price. To examine the effects of under- or over-procurement on the day-ahead market, we use the 5% and 95% quantile of the forecast of the day-ahead model and then procure the difference on the intraday market again. In this case, our analysis shows that it is more favorable for our predicted load energy in the test period to procure the lower quantile and then sell or buy the difference than to buy the upper quantile and then sell/buy the difference. However, the difference in the resulting total costs is less than 1%. In order to procure energy with as little risk as possible, energy and fleet managers should purchase the average forecast energy on the day-ahead market. Static procurement of charging energy results in lower costs on the one hand. On the other hand, they are not forced to buy or sell large quantities in the event of significant fluctuations in the intraday market. In particular, large quantities to be balanced out can exacerbate the price difference in situations with little liquidity. It is crucial to make the intraday volume forecast as accurate as possible because, as shown, the costs only increase marginally, and the balancing energy required decreases to the same extent as the improved forecast. Further, when taking flexibility into account the optimization and cost reduction potential increases substantially, by not only shifting the charging into times with low prices but also doing arbitrage trading.

Discussion and conclusion

This section discusses the potential limitations and directions for future work. As previous studies have shown, our results confirm that machine learning methods are suitable for predicting the charging load of electric vehicles and that the prediction improves with increasing fleet size. However, previous studies often only consider point predictions, whereas this paper also applied probabilistic predictions. Ensemble models, especially ada, bagging, and random forest, were shown to be robust across different aggregation levels, making them a reliable choice for different scenarios. These findings are comparable to those of Rathore et al. (2023). Although the machine learning models performed best and were superior to the benchmark model, our deep learning models do not utilize their full potential. This can be recognized by comparing the results of the models in Table 3 and Tables 6, 7, 8. With additional adjustments in the deep learning architectures

and additional hyperparameter tuning, they could improve their adaptability to different levels of data granularity and achieve enhanced results. Furthermore, exploring additional features and feature transformations can improve model performance, especially for deep learning models at finer levels of aggregation. Regarding the limitations of using only the used algorithms, it could be argued that using different models such as support vector machines, k-nearest neighbor or especially state of the art deep learning model architectures such as PLCNet or temporal fusion transformer might achieve better results (Lim et al. 2019; Farsi et al. 2021). They are including not only national but also state-specific holidays as a feature that could further improve the results of all models. Further, regional weather data as input feature might improve the forecast accuracy as we initially only tested with German-wide data. The influence of historical data of different lengths has shown that, as expected, more data provides improved results, but a certain degree of saturation is indicated. In the future, it would be interesting to investigate the influence of annual data, as this would allow the models to better account for seasonal effects. For new sites that have little or no historical data, one possible approach would be transfer learning. Here, a global model is trained on the existing sites. New sites that still need to have sufficient historical data can then be predicted. In addition, data augmentation techniques can artificially generate data to have more training data available and ultimately may reduce overfitting. These approaches would be further possibilities for future research to investigate.

The random aggregation of the individual sites has shown that robust results are achieved from approximately 200 chargers compared to the benchmark model. For the random aggregation of the individual sites, it would also be interesting to investigate how the models react to different behavioral dynamics, for example, by adding chargers to the existing sites or including new sites. Future research could aggregate sites by clustering hard-to-predict sites instead of randomly aggregating them or by training a global model on all data and then applying it to individual sites. Concerning the procurement of energy volumes, it was shown that an improved forecast leads to less balancing energy, but not to the same extent as higher procurement costs. It would be interesting for future work to investigate how procurement can be carried out in countries where it is also permitted to optimize expenses based on the balancing energy price. Additionally, this paper only examines static energy procurement without using the flexibility of electric vehicles, which results from the possibility of shifting charging processes. Future work should investigate how flexibility can be predicted. For example, one approach could be to predict the charging load and whether a vehicle plugs in as a time series per charging point to determine the shifting potential per charging point. In addition, the amount of energy and plugging duration could be determined as a regression for each plugging process, and the two models could be combined to increase reliability. The flexibility prediction and subsequent optimal energy procurement become particularly complex when bidirectional charging is considered, significantly changing the boundary conditions in energy procurement. A combination of prediction and optimization models, e.g., with reinforcement learning, is an approach that future research should examine.

The results presented relate to the charging processes of German companies, primarily from the commercial sector. If data from public or private charging stations is considered, different results will be obtained due to the significantly different charging behavior. By

standardizing the charging load to the number, we have considered the ramp-up of electric vehicles. Furthermore, the results are transferable to other countries with similar companies. It should be noted that the charging load may change in the future due to the following factors: on the one hand, bidirectional charging will play an increasingly important role, allowing not only charging but also discharging. This will not only change the load but also give users an even greater incentive to keep their vehicles plugged in for long periods of time. Further technological advances in battery technology will increase the charging speed and the battery capacity, which will directly impact the charging load if vehicles have to charge less frequently but for longer, provided the power remains the same.

As companies might have limited time and resources, the following factors should be considered when implementing charging load forecasting models and any forecast model: sometimes naive models, such as the mean per day and time, yield accurate results without any implementation and maintenance effort. At the same time, they are easy to understand, and the computational costs are comparatively small. Therefore, in a real world application, the added value of a more accurate prediction should always be compared to the additional effort of developing and maintaining more complex models.

Data availability is one of the most important aspects that should be considered in long-term planning. Therefore, companies should ensure that appropriate structures are put in place early to enable data processing. However, politicians and legislators must also ensure the right framework conditions are in place. For example, the EU is pushing to create the European mobility data space (European Commission 2023). The aim should be for organizations to recognize the added value data availability, which will ultimately help them to make better products and business decisions.

Appendix

See Tables 5, 6, 7, and 8.

Table 5 Hyperparameters for machine learning models grid search

Model	Hyperparameter	Space
Bagging	Base model	Decision tree
	Number of estimators	250, 375, 500
	Maximal samples	0.75. 1.0
	Bootstrapping	True, false
RF	Maximal depth	None, 3, 5, 10
	Number of estimators	500, 750, 1000
	Minimal samples split	2, 4, 8
GradientB	Learning rate	0.3
	Number of estimators	50, 100, 150, 200, 250, 300, 350
	Maximal depth	3
	Minimal sample split	2
	Subsample	1.0
Ada	Base model	Decision tree
	Number of estimators	10, 50, 100
	Learning rate	0.1, 0.5, 1.0
LinR	Fit intercept	True, false

Table 6 Metrics of the models for the TZO zones

TZO zone	Model	nRMSE*	MASE*	R2**	PS low Q***	PS high Q***	IS***
Tennet	LinR	0.883	0.806	134.983	0.793		
	Bagging	0.534	0.431	81.516	0.924	1.955	4.259
	GradientB	0.536	0.523	81.820	0.924	3.887	4.527
	Ada	0.467	0.414	71.388	0.942	1.945	4.142
	Random forest	0.545	0.456	83.233	0.921	3.650	6.052
	LSTM	0.493	0.513	75.266	0.936	34.397	69.342
	CNN	0.589	0.519	89.997	0.908	38.683	76.597
	NN	0.587	0.660	89.752	0.908	141.111	89.672
50Hertz	LinR	0.825	0.860	91.940	0.787		
	Bagging	0.648	0.662	72.269	0.869	2.782	3.765
	GradientB	0.628	0.686	70.013	0.877	3.481	4.488
	Ada	0.623	0.659	69.436	0.879	2.498	3.808
	Random forest	0.664	0.669	74.091	0.862	5.227	6.575
	LSTM	0.654	0.721	72.922	0.866	97.063	101.574
	CNN	0.788	0.838	87.828	0.806	88.751	92.326
	NN	0.734	0.791	81.815	0.832	115.450	108.417
Amprion	LinR	0.914	0.904	122.239	0.780		
	Bagging	0.427	0.473	57.084	0.952	2.115	2.819
	GradientB	0.525	0.599	70.274	0.927	3.555	4.009
	Ada	0.417	0.473	55.783	0.954	2.248	3.034
	Random forest	0.447	0.507	59.807	0.947	3.796	4.939
	LSTM	0.584	0.616	78.181	0.910	46.069	55.681
	CNN	0.718	0.683	96.105	0.864	61.516	60.636
	NN	0.561	0.652	75.076	0.917	117.008	91.710
TransnetBW	LinR	0.845	0.768	173.191	0.768		
	Bagging	0.568	0.467	116.333	0.895	3.571	5.436
	GradientB	0.533	0.514	109.334	0.908	4.468	6.936
	Ada	0.579	0.470	118.761	0.891	3.123	5.592
	Random forest	0.555	0.477	113.681	0.900	6.029	9.262
	LSTM	0.574	0.611	117.723	0.893	52.813	111.074
	CNN	0.634	0.600	129.904	0.869	63.730	108.037
	NN	0.643	0.673	131.848	0.866	188.707	125.658

*Unitless, **in %, ***in W

Table 7 Metrics of the models for the zip codes

Zip code	Model	nRMSE*	MASE*	R2**	PS low Q***	PS high Q***	IS***
I	LinR	0.975	1.09	0.351			
	Bagging	0.990	1.108	0.331	7.090	27.448	985.517
	GradientB	0.981	1.088	0.343	6.039	24.667	1033.998
	Ada	0.983	1.171	0.341	6.214	27.861	1176.891
	Random forest	1.131	1.232	0.127	125.155	78.103	548.477
	LSTM	1.015	1.138	0.296	252.223	235.476	605.484
	CNN	1.013	1.002	0.300	211.708	235.503	574.080
	NN	1.034	1.132	0.270	458.458	235.482	826.995

Table 7 (continued)

Zip code	Model	nRMSE*	MASE*	R2**	PS low Q***	PS high Q***	IS***
II	LinR	0.960	0.94	0.687			
	Bagging	0.663	0.684	0.851	2.053	3.936	154.005
	GradientB	0.611	0.657	0.873	2.620	6.178	191.541
	Ada	0.689	0.698	0.839	1.968	4.082	162.425
	Random forest	0.682	0.684	0.842	4.002	7.951	113.861
	LSTM	0.763	0.872	0.802	61.401	99.888	189.507
	CNN	0.703	0.756	0.832	67.554	99.889	194.910
	NN	0.855	0.931	0.751	108.516	99.888	240.200
III	LinR	0.958	1.008	0.375			
	Bagging	1.008	1.027	0.309	4.447	11.906	567.686
	GradientB	0.992	0.993	0.330	4.675	12.048	546.991
	Ada	1.005	1.047	0.312	4.196	12.550	633.464
	Random forest	1.013	1.045	0.302	44.326	32.429	299.845
	LSTM	1.059	1.106	0.237	279.547	163.712	509.868
	CNN	1.095	1.048	0.184	294.927	163.713	528.844
	NN	0.965	0.988	0.366	552.837	163.704	814.265
IV	LinR	0.949	1	0.734			
	Bagging	0.820	0.823	0.801	5.605	10.621	527.691
	GradientB	0.812	0.849	0.805	6.747	11.513	631.780
	Ada	0.834	0.889	0.795	5.583	11.247	602.754
	Random forest	0.814	0.837	0.804	15.005	20.154	340.988
	LSTM	0.848	0.916	0.788	310.657	264.919	658.047
	CNN	1.005	1.02	0.702	287.591	254.567	623.086
	NN	0.978	1.166	0.717	870.033	263.764	1289.562
V	LinR	0.807	0.837	0.815			
	Bagging	0.688	0.652	0.866	5.764	7.758	303.393
	GradientB	0.660	0.666	0.877	5.494	8.427	394.922
	Ada	0.666	0.663	0.874	5.020	8.045	320.075
	Random forest	0.713	0.674	0.856	8.923	13.661	229.896
	LSTM	0.791	0.787	0.823	85.854	209.834	347.807
	CNN	0.928	0.85	0.756	73.755	209.837	343.705
	NN	0.636	0.721	0.885	205.055	209.838	472.293

*Unitless, **in %, ***in W

Table 8 Metrics of the models for the individual sites

Site (number of charger)	Model	nRMSE*	MASE*	R2**	PS low Q***	PS high Q***	IS***
A (3)	LinR	1.009	1.151	584.168	0.15	0	0
	Bagging	1.106	1.115	640.565	−0.023	3.024	37.963
	GradientB	1.029	1.057	595.809	0.115	3.024	30.945
	Ada	1.067	1.356	618.029	0.048	3.024	50.004
	Random forest	1.111	1.105	643.396	−0.032	3.024	41.262
	LSTM	1.088	1.236	629.841	0.011	1943.876	117.946
	CNN	1.172	1.252	678.441	−0.147	666.728	117.946
	NN	1.185	1.01	686.159	−0.173	525.407	117.947

Table 8 (continued)

Site (number of charger)	Model	nRMSE*	MASE*	R2**	PS low Q***	PS high Q***	IS***
B (4)	LinR	0.992	1.077	870.695	0.189	0	0
	Bagging	1.013	1.017	889.115	0.154	10.281	82.067
	GradientB	0.994	1.032	872.227	0.186	10.281	73.75
	Ada	0.994	1.094	872.266	0.186	10.281	89.058
	Random forest	1.031	1.034	905.217	0.123	10.283	81.927
	LSTM	1.071	1.067	940.387	0.054	1695.592	400.964
	CNN	1.164	1.19	1022.103	-0.118	926.987	400.964
	NN	1.095	1.013	961.368	0.011	1211.949	400.964
C (8)	LinR	0.954	1.149	294.785	0.116	0	0
	Bagging	0.993	0.849	307.083	0.041	1.723	16.471
	GradientB	0.989	1.042	305.84	0.049	1.617	14.012
	Ada	0.982	0.981	303.652	0.062	1.797	15.972
	Random forest	1.022	0.869	316.031	− 0.016	2.739	17.262
	LSTM	0.894	0.89	276.51	0.222	807.033	63.078
	CNN	1.057	0.802	326.657	− 0.085	234.903	63.082
	NN	1.093	0.858	337.963	− 0.162	409.187	88.649
D (14)	LinR	0.971	1.096	644.271	0.32	0	0
	Bagging	1.022	1.108	677.737	0.247	9.038	56.199
	GradientB	1.034	1.15	685.814	0.229	9.054	47.525
	Ada	1.026	1.211	680.719	0.241	9.21	57.178
	Random forest	1.027	1.112	680.964	0.24	13.526	64.093
	LSTM	1.005	1.062	666.66	0.272	1060.265	353.12
	CNN	1.071	1.241	710.662	0.172	588.897	353.132
	NN	1.044	1.035	692.421	0.214	345.821	353.13
E (145)	LinR	0.74	0.719	301.995	0.764	0	0
	Bagging	0.671	0.537	273.832	0.806	10.581	12.018
	GradientB	0.616	0.562	251.491	0.837	6.126	14.434
	Ada	0.66	0.553	269.211	0.813	10.988	12.427
	Random forest	0.691	0.547	282.057	0.794	11.293	16.224
	LSTM	1.009	1.151	584.168	0.15	0	0
	CNN	1.106	1.115	640.565	− 0.023	3.024	37.963
	NN	1.029	1.057	595.809	0.115	3.024	30.945

*Unitless, **in %, ***in W

Acknowledgements
We would like to thank The Mobility House, especially Dr. Christopher Hecht and Dr. Michael Schreiber, for providing the data, their time, and valuable feedback during our regular meetings.

Author contributions
Conceptualization: AO and TH; methodology: AO, TH; software: TH and AO; validation: AO, TH; formal analysis: AO; investigation: AO; resources: AO; data curation: TH and AO; writing—original draft preparation: AO; writing—review and editing: AO and TH; visualization: AO; project administration: AO; funding acquisition: AO. All authors have read and agreed to the published version of the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. The described work was conducted within the project "unIT-e^2" by Forschungsstelle für Energiewirtschaft e.V. (FfE). The project is funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK) (funding code: 01MV21UN11).

Availability of data and materials
Data is unavailable due to privacy restrictions. However, for future research, we point to Zahler and Ostermann (2022), which provides a list of real-world charging datasets that are partly available.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no conflict of interest. The funders had no role in the study's design, in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Received: 23 January 2024 Accepted: 23 February 2024
Published online: 29 February 2024

References

Aghsaee R, Hecht C, Schwinger F, Figgener J, Jarke M, Sauer DU (2023) Data-driven, short-term prediction of charging station occupation. Electricity 4:134–153. https://doi.org/10.3390/electricity4020009

Athiyarath S, Paul M, Krishnaswamy S (2020) A comparative study and analysis of time series forecasting techniques. SN Comput Sci 1:175. https://doi.org/10.1007/s42979-020-00180-5

Barker J (2020) Machine learning in M4: what makes a good unstructured model? Int J Forecast 36:150–155. https://doi.org/10.1016/j.ijforecast.2019.06.001

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Buzna L, de Falco P, Ferruzzi G, Khormali S, Proto D, Refa N, Straka M, van der Poel G (2021) An ensemble methodology for hierarchical probabilistic electric vehicle load forecasting at regular charging stations. Appl Energy 283:116337. https://doi.org/10.1016/j.apenergy.2020.116337

Chen Y, Wu G, Sun R, Dubey A, Laszka A, Pugliese P (2020) A review and outlook of energy consumption estimation models for electric vehicles. arXiv preprint. arXiv:2003.12873

Duscha V, Wachsmuth J, Eckstein J, Pfluger B (2019) GHG-neutral EU2050—a scenario of an EU with net-zero greenhouse gas emissions and its implications. https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2019-11-26_cc_40-2019_ghg_neutral_eu2050_0.pdf

ENTSO-E (2024) Transparency platform. https://transparency.entsoe.eu/. Accessed 21 Feb 2024

European Commission (2023) Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions: creation of a common European mobility data space. COM (2023) 751 final. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52023DC0751. Accessed 18 Feb 2024

ev.energy (2023) Majority of EV owners confident with long-distance journeys as range anxiety now impacts less than a quarter of drivers. https://www.ev.energy/blog/majority-of-ev-owners-confident-with-long-distance-journeys-as-range-anxiety-now-impacts-less-than-a-quarter-of-drivers. Accessed 19 Jan 2024

Farsi B, Amayri M, Bouguila N, Eicker U (2021) On short-term load forecasting using machine learning techniques and a novel parallel deep LSTM-CNN approach. IEEE Access 9:31191–31212. https://doi.org/10.1109/ACCESS.2021.3060290

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proceedings of the thirteenth international conference on machine learning. Morgan Kaufmann Publishers Inc, San Francisco, pp 148–156. ISBN 1558604197.

Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gemassmer J, Daam C, Reibsch R (2021) Challenges in grid integration of electric vehicles in urban and rural areas. WEVJ 12:206. https://doi.org/10.3390/wevj12040206

Hatalis K, Lamadrid AJ, Scheinberg K, Kishore S (2017) Smooth pinball neural network for probabilistic forecasting of wind power. http://arxiv.org/pdf/1710.01720v1. Accessed 18 Feb 2024

Hecht C, Figgener J, Sauer DU (2021) Predicting electric vehicle charging station availability using ensemble machine learning. Energies 14:7834. https://doi.org/10.3390/en14237834

Hyndman R, Athanasopoulos G (2021) Forecasting: principles and practice, 3rd edn. OTexts, Melbourne

James G, Witten D, Hastie T, Tibshirani R (eds) (2021) An introduction to statistical learning. Springer, New York. ISBN 978-1-0716-1417-4

Kern T (2023) Assessment of the added value of bidirectionally chargeable electric vehicles for the user and the energy system. Technische Universität München, München

Kim Y, Kim S (2021) Forecasting charging demand of electric vehicles using time-series models. Energies 14:1487. https://doi.org/10.3390/en14051487

Koenker R, Machado JAF (1999) Goodness of fit and related inference processes for quantile regression. J Am Stat Assoc 94:1296–1310. https://doi.org/10.1080/01621459.1999.10473882

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

Lim B, Arik SO, Loeff N, Pfister T (2019) Temporal fusion transformers for interpretable multi-horizon time series forecasting

Majidpour M, Qiu C, Chu P, Pota HR, Gadh R (2016) Forecasting the EV charging load based on customer profile or station measurement? Appl Energy 163:134–141. https://doi.org/10.1016/j.apenergy.2015.10.184

Mediouni H, Ezzouhri A, Charouh Z, El Harouri K, El Hani S, Ghogho M (2022) Energy consumption prediction and analysis for electric vehicles: a hybrid approach. Energies 15:6490. https://doi.org/10.3390/en15176490

Müller MD (2023) Netzintegration dezentraler Flexibilitätsoptionen mit Fokus auf ausgewählte Anwendungsfälle für bidirektionale Elektrofahrzeuge. Technische Universität München, München
OECD (2023) Global EV outlook 2023. OECD, Paris. ISBN 978-92-64856-92-9
Ostermann A, Fabel Y, Ouan K, Koo H (2022) Forecasting charging point occupancy using supervised learning algorithms. Energies 15:3409. https://doi.org/10.3390/en15093409
Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) PyTorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G et al (2012) Scikit-learn: machine learning in python. J Mach Learn Res. https://doi.org/10.48550/arXiv.1201.0490
Rathore H, Meena HK, Jain P (2023) Prediction of EV energy consumption using random forest and XGBoost. In: 2023 international conference on power electronics and energy (ICPEE), Bhubaneswar, India, 03–05 Jan. 2023. IEEE, pp 1–6. ISBN 978-1-6654-7058-2
Recurrent Auto (2024) EV battery life exceeds expectations in study of 15,000 electric cars. https://www.globenewswire.com/en/news-release/2023/03/27/2634907/0/en/EV-Battery-Life-Exceeds-Expectations-in-Study-of-15-000-Electric-Cars.html. Accessed 19 Jan 2024
Regett A (2020) Development of instruments for a circular energy economy. Technische Universität München, München
Sharkawy A-N (2020) Principle of neural network and its main types: review. J Adv Appl Comput Math 7:8–19. https://doi.org/10.15377/2409-5761.2020.07.2
Shen H, Zhou X, Wang Z, Ahn H, Lamantia M, Chen P, Wang J (2022) Electric vehicle energy consumption estimation with consideration of longitudinal slip ratio and machine-learning-based powertrain efficiency. IFAC-PapersOnLine 55:158–163. https://doi.org/10.1016/j.ifacol.2022.11.177
van Kriekinge G, de Cauwer C, Sapountzoglou N, Coosemans T, Messagie M (2021) Day-ahead forecast of electric vehicle charging demand with deep neural networks. WEVJ 12:178. https://doi.org/10.3390/wevj12040178
Wang H, Raj B (2017) On the origin of deep learning. http://arxiv.org/pdf/1702.07800v4
Wohlschlager D, Haas S, Neitz-Regett A (2022) Comparative environmental impact assessment of ICT for smart charging of electric vehicles in Germany. Procedia CIRP 105:583–588. https://doi.org/10.1016/j.procir.2022.02.097
Xie F, Huang M, Zhang W, Li J (2011) Research on electric vehicle charging station load forecasting. In: 2011 international conference on advanced power system automation and protection (APAP), Beijing, China, 16–20 Oct. 2011. IEEE, pp 2055–2060. ISBN 978-1-4244-9621-1
Xydas ES, Marmaras CE, Cipcigan LM, Hassan AS, Jenkins N (2013) Forecasting electric vehicle charging demand using support vector machines. In: 2013 48th international universities' power engineering conference (UPEC), Dublin, 02–05 Sep. 2013. IEEE, pp 1–6. ISBN 978-1-4799-3254-2
Yi Z, Liu XC, Wei R, Chen X, Dai J (2022) Electric vehicle charging demand forecasting using deep learning model. J Intell Transp Syst 26:690–703. https://doi.org/10.1080/15472450.2021.1966627
Zahler J, Ostermann A (2022) Charging infrastructure and grid expansion needs: an overview of real-world charging load datasets. https://www.ffe.de/en/publications/charging-infrastructure-and-grid-expansion-needs-an-overview-of-real-world-charging-load-datasets/. Accessed 18 Feb 2024
Zhu J, Yang Z, Mourshed M, Guo Y, Zhou Y, Chang Y, Wei Y, Feng S (2019) Electric vehicle charging load forecasting: a comparative study of deep learning approaches. Energies 12(14):2692

Publisher's Note