



Data-driven available capacity estimation of lithium-ion batteries based on fragmented charge capacity



Zhen Zhang , Xin Gu, Yuhao Zhu, Teng Wang, Yichang Gong & Yunlong Shang

Efficient and accurate available capacity estimation of lithium-ion batteries is crucial for ensuring the safe and effective operation of electric vehicles. However, incomplete charging cycles in practical applications challenge conventional methods. Here we manipulate fragmented charge capacity data to estimate available capacity without complete charging information. Considering correlation, charging time, and initial state of charge, 36 feature combinations are available for estimation. The basic machine learning model is established on 11,500 cyclic samples, and a transfer learning model is fine-tuned and validated on multiple datasets. The validation results indicate that the best root-mean-square error for the basic model is 0.012. Furthermore, the RMSE demonstrates consistent stability across different datasets in the transfer learning model, with fluctuations within 0.5% when considering feature combinations across cycles with spacings of 5, 10, and 20. This work highlights the promise of available capacity estimation using actual, readily accessible fragmented charge capacity data.

Electric vehicles (EVs) are the prime contributors to promoting transportation electrification. Their energy storage equipment is lithium-ion batteries (LIBs), which have been extensively investigated and transformed^{1,2}. It is widely recognized that low-cost LIBs offer high energy density and a long lifespan³. In 2023, global shipments of LIBs reached 1202.6 GWh, with 865.2 GWh allocated to new energy vehicles, accounting for more than 70% of the total. However, LIBs in EVs experience irreversible capacity degradation during long-term operation due to factors such as solid electrolyte interface (SEI) thickening⁴, fast charging⁵, and overcharging⁶. Available capacity directly affects the energy storage level and driving range of EVs, making it a crucial indicator for evaluating battery safety^{7,8}. Achieving the accurate estimation of available capacity is a challenging task.

Typically, the available capacity is represented by the discharge capacity from a complete charge and discharge cycle, as illustrated in Supplementary Fig. 1(a). However, factors such as charging difficulties, driving purposes, and personal habits contribute to irregular and fragmented charging behaviors in real-world scenarios^{10–12}. Meanwhile, the discharge process depends on operational inputs and prevailing road conditions. In this context, estimating available capacity using complete charge/discharge data is not feasible¹³. Furthermore, the long duration of a single cycle and the complexity of feature data extraction make conventional methods impractical. Battery voltage and charge/discharge capacity are easily accessible parameters that are closely related to the electrochemical and thermodynamic properties of LIBs¹⁴. Consequently, estimating available

capacity using data from the daily operation of EVs represents a more promising avenue for investigation.

Currently, mainstream methods involve extracting degradation features and mapping them inversely to estimate available capacity¹⁵. Features commonly used to represent LIB degradation include electrical¹⁶, mechanical¹⁷, temperature¹⁸, electrochemical¹⁹, and acoustic features²⁰, among others. Data-driven methods based on machine learning (ML) and data processing techniques can establish mapping relationships using datasets and leverage excellent nonlinear fitting capabilities. With advancements in computational power and software, data-driven methods have emerged as the leading investigative tools in battery management, without requiring an in-depth understanding of electrochemical principles²¹. ML was employed to construct a prediction model that utilizes voltage data to accurately estimate the cycle life of commercial LiFePO₄/graphite batteries²². The Gaussian process regression was operated to automatically determine which impedance spectroscopy features predict degradation and identify the remaining useful life (RUL)²³. A deep learning framework was presented to evaluate battery state of health (SOH)²⁴. Laborious feature engineering can hinder the accuracy of data-driven estimation, particularly due to the complexity of feature acquisition and the necessary prerequisites. For example, extracting incremental capacity (IC) curves requires high-sampling accuracy datasets and curve-smoothing strategies, which can compromise data reliability²⁵. Voltage relaxation can only be obtained after the battery is fully charged and in a resting state (zero

current²⁶. Therefore, the simplicity of the feature extraction process, the flexibility of feature combination approaches, and the diversity of feature forms collectively determine the performance of data-driven methods in available capacity estimation.

Amid the trend toward transportation electrification, LIBs in EVs necessitate frequent charging to accommodate the stochastic discharge process. Recently, the prevalent charging protocols include constant current, constant current-constant voltage (CC-CV), and multi-stage constant current. In this context, the CC-CV method is regarded as a benchmark for other charging protocols due to its advantages in charging time, efficiency, and cycle life²⁷. However, charging a LIB is often influenced by user behavior, leading to variability in the starting and ending points of the charging process. Thus, considerable attention has been directed toward extracting features from fragmented charging data. These features can be broadly categorized into three groups: (I) voltage rise curves and their transformations during CC charging²⁸, (II) current decline curves and their transformations during CV charging²⁹, and (III) transformations of measurable parameters using statistical methods, including the mean, sum, standard deviation, and others³⁰. Fewer works have taken measurable parameters as feature vectors combined in a flexible manner for available capacity estimation. Many scholars evaluate battery degradation by mining diverse features^{31,32}. Unfortunately, the weak correlation between features and available capacity, the lack of interdependence among features, and the complexity of feature engineering make practical implementation challenging.

It has been proven that the charge capacity during CC charging can be directly obtained without additional filtering, and this complete charge capacity serves as a proxy for the available capacity of LIBs. Based on numerous reviews and investigations^{33,34}, we analyze more than 250 charging cycles of a commercial EV (Supplementary Figs. 2 and 3) and obtain the findings that (I) the start and end points of charging cycles are irregular, (II) the charging cycles exhibit fragmentation, and (III) complete charging/discharging processes are rare. In general, the CC charging time (charge capacity) over a fixed voltage interval can be used as an input feature for estimating the available capacity using the fragment data³⁵. However, the optimal voltage interval varies dynamically with battery type, operating conditions, and other factors. This variability complicates the standardization of the interval and undermines estimation accuracy. Therefore, we propose an approach for available capacity estimation based on fragmented charge capacity, emphasizing flexible feature combinations. This method embodies the principle of ‘less is more,’ as illustrated in Supplementary Fig. 1(b).

In this work, we segment the complete data, considering correlation and charging coverage, to serve as the basis for feature engineering. Charge capacities across various voltage intervals are selected as candidate features and are combined in a rational manner. This research employs basic ML models, including a linear model (LASSO³⁶) and non-linear models (XGBoost³⁷ and LightGBM³⁸). We also provide a large dataset comprising 11,500 samples with four LiFePO₄/graphite cells for feature extraction and model training. Multiple datasets are deployed to validate the performance of the transfer learning (TL) model. We observe that the best root-mean-square error (RMSE) for the dataset used to establish the basic model is 0.012. For the other datasets, the RMSE remains consistently stable. To meet practical requirements, we also examine the impact on estimation performance when input features are derived from different cycles. The results indicate that the proposed method adapts excellently to the spacing in the

feature combinations. This discussion underscores the potential of the proposed method for accurately estimating available capacity.

Results

Data generation

Data-driven methods for available capacity estimation of LIBs necessitate the underpinning of large datasets. To this end, we collect four degradation datasets for LiFePO₄/graphite batteries from the manufacturer's VALENCE, HUAWEI, GOTION, and A123. Table 1 provides basic information about these datasets. In this paper, the four datasets are referred to as dataset #1, dataset #2, dataset #3, and dataset #4. Dataset #1 is employed for the establishment, training, and testing of the basic model, while datasets #2 and #3 are intended for validation and evaluation under TL³⁹. Dataset #4 is extracted from the MATR dataset, which includes fast charging and CC-CV discharging. Detailed charge/discharge rate information is provided in Supplementary Table 1 and Supplementary Table 2. Dataset #4 is used to validate model performance when applying feature combinations across cycles without TL. Additionally, Supplementary Note 1 explains the expression format of the charge current rate as presented in Supplementary Tables 1 and 2. The degradation data for the VALENCE and HUAWEI LIBs are generated in our laboratory through cycling experiments. Dataset #3 from GOTION is a public dataset²⁴. Detailed experimental information for all datasets is provided in Supplementary Table 3. It is evident that the charging/discharging protocols for the first three datasets follow a CC-CV/CC pattern, with the primary difference being the varying charging/discharging currents. Similar to common practice in the field, this paper takes the C-rate to describe the charging/discharging current of the battery relative to its nominal capacity. For instance, dataset #1 has a charging current rate of 1 C (2.5 A) and a discharging current rate of 4 C (10 A).

Taking dataset #1 as an example, Fig. 1(a) illustrates three complete cycles from the experiment, each consisting of three main phases: (I) CC charging, (II) CV charging, and (III) CC discharging. The CC discharge capacity is considered the available capacity within a cycle. Figure 1(b) illustrates the voltage-charge capacity (V-Q) curve in the CC charging. The V-Q curve is roughly divided into four segments, with segment (I) and segment (IV) located at the beginning and end of charging, where the voltage increases rapidly, but the charge capacity is minimal. Segments (II) and (III) make up the voltage plateau phase, which is divided by SOC = 50%. The charge capacity within this phase accounts for 80% of the available capacity.

To illustrate the changes in V-Q curves during the degradation of LIBs, we plot the V-Q curves of the four cells in dataset #1. As shown in Fig. 1(c-f), the four sets of V-Q curves exhibit consistent trends during degradation and cover the SOH range of 80% to 100%. However, despite being produced by the same manufacturer and tested in the same experimental environment, these four cells show significant differences in SOH distribution. The reason is that the degradation of LIBs is influenced by multiple factors, with manufacturing variations being a key contributor⁴⁰. Focusing on the VALENCE #1 cell, as shown in Fig. 1(c), the magnitude of the voltage plateau phase decreases as the cell degrades, and the total charge capacity during CC charging declines. In Fig. 2, we present the available capacity degradation curves for the four datasets.

Feature extraction

Systematic and detailed profiling of compound datasets is an effective approach to uncovering internal patterns and extracting reliable features. As

Table 1 | Basic information about four datasets

Dataset	Manufacturer	Charge current rate (C) /discharge rate (C)	Nominal capacity (Ah)	Number of cells	Number of samples
Dataset #1	VALENCE	1/4	2.5	4	11500
Dataset #2	HUAWEI	0.5/0.5	280	4	4574
Dataset #3	GOTION	1/1	27	3	4262
Dataset #4	A123	Detailed information in Supplementary Tables 1 and 2	1.1	50	32800

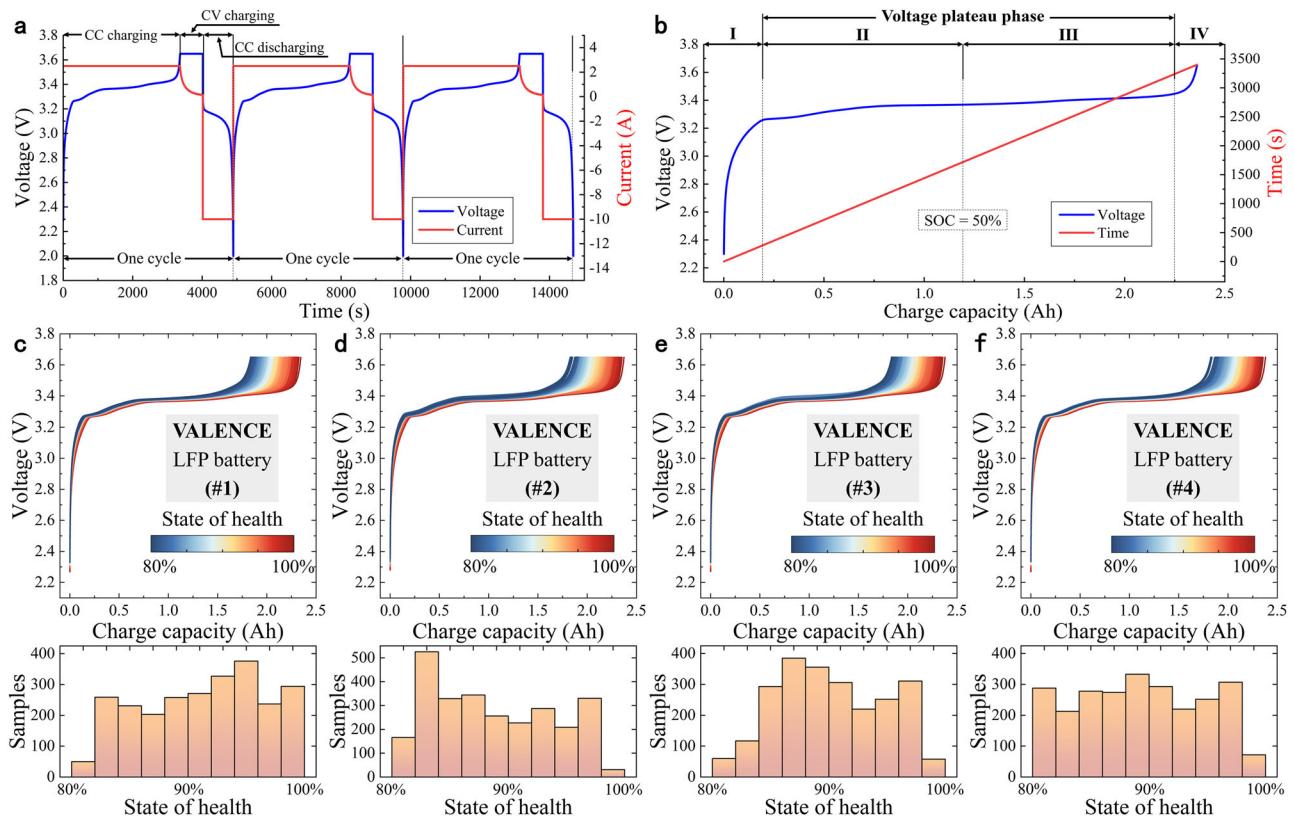


Fig. 1 | Details of dataset #1. Voltage and current profile over three complete cycles (a). V-Q curve in CC charging (b). (c–f) correspond to the degraded V-Q curves and SOH (state of health) histograms for cells #1, #2, #3, and #4, respectively. The

gradients in (c–f) correspond to changes in battery SOH. SOC denotes the state of charge. LFP denotes the LiFePO₄. CC denotes the constant current. CV denotes the constant voltage.

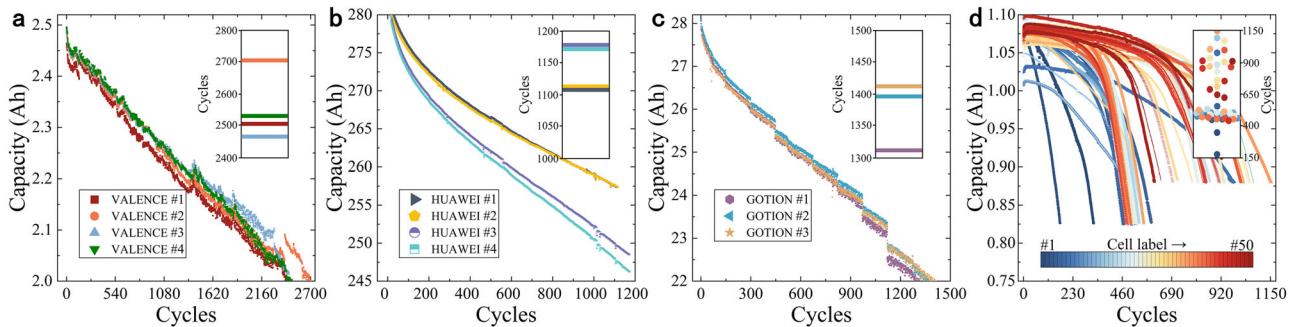


Fig. 2 | Available capacity degradation curves. Dataset #1 (a), dataset #2 (b), dataset #3 (c), and dataset #4 (d). The embedded figures in (a–d) are the cycle distribution of cells.

previously mentioned, variations in the V-Q curve with deeper cycling are inevitably linked to LIB degradation⁴¹. Convincingly, voltage and charge capacity are among the few real-time parameters that are easily accessible in most scenarios and under various operating conditions. Accordingly, research focused on the V-Q curve has both practical significance and a solid theoretical foundation. While obtaining a complete V-Q curve can be challenging and unlikely, fragmented, piecemeal data is frequently available. Segmenting the complete V-Q curve can simulate fragmented data as encountered in practice. Using the VALENCE #1 cell as an example, we divide the complete curve into segments based on multiple voltage thresholds. This paper considers the charge capacity within each segment as a candidate feature.

The selection of the aforementioned voltage thresholds necessitates careful consideration. First, the charging time between each voltage threshold should be balanced; if it is too long, it fails to capture

fragmentation, while if it is too short, it may lack distinguishing features. Additionally, since the SOC is an intuitive indicator for drivers that reflects the charging process, it is also important to observe the initial SOC at each segment. The segmented portions should exhibit low and similar charge capacities to facilitate the extraction of feature segments from fragmented data. Meanwhile, these smaller segments should be capable of being aggregated into larger, continuous voltage segments to accommodate extended charging behaviors. For this reason, we segment the V-Q curve and construct 11 segments, as presented in Fig. 3(a). These segments, labeled Segment 1 through Segment 11, span the voltage range from 3 V to 3.6 V. Figure 3(b) and (c) convey the performance of each segment with respect to charging time and initial SOC. The average and median charging time for each segment is within 550 s, which essentially splits the redundant voltage plateau phase (average time is 2682 s). Detailed charging time information can be found in Supplementary Table 4. The initial SOC of each segment

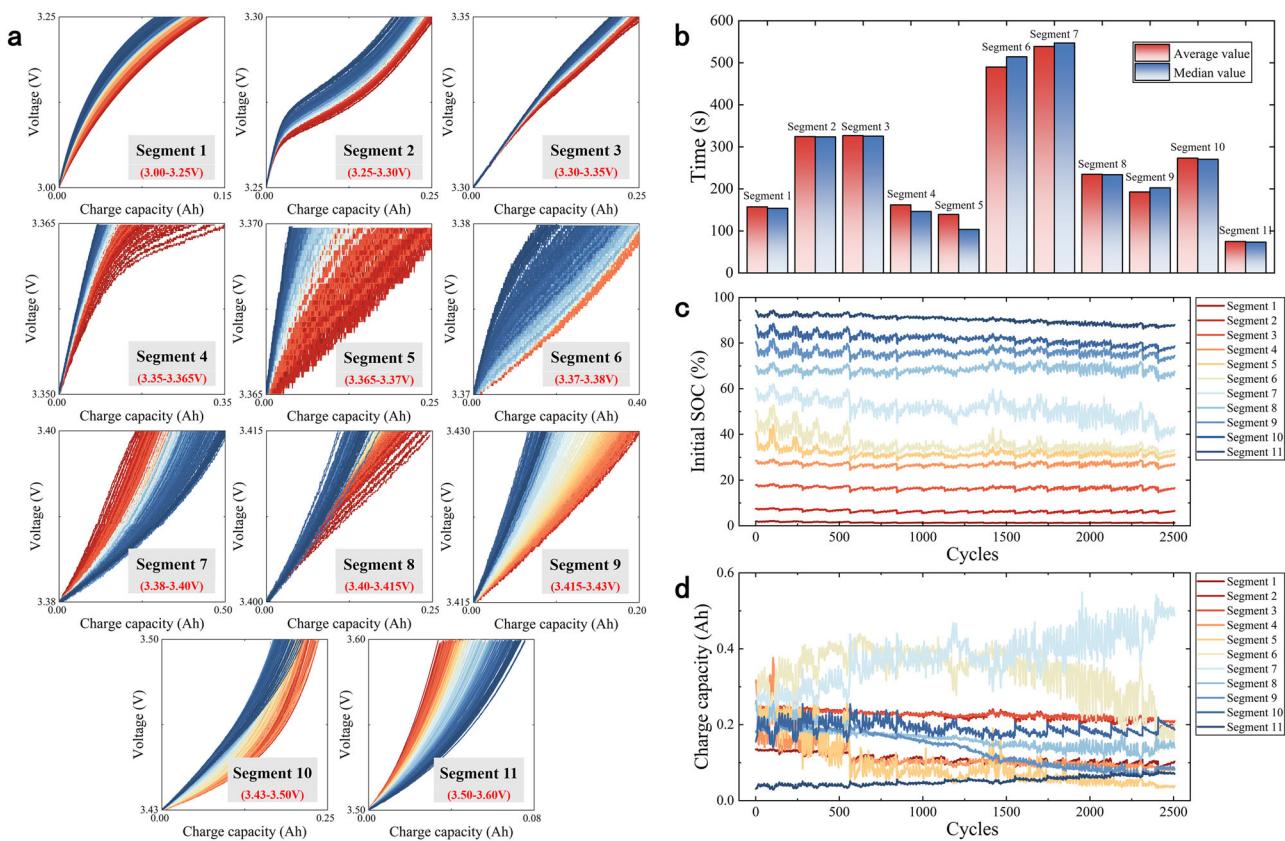


Fig. 3 | Extraction of candidate features from the complete V-Q curve. 11 segments divided by various voltage intervals (a). (b), (c) convey the performance of each candidate feature in terms of charging time and initial SOC (state of charge). (d)

depicts the variation trend of each candidate feature. The gradient in (a) corresponds to changes in battery SOH (state of health).

shows slight fluctuations as the LIB degrades. Figure 3(d) depicts the trend in charge capacity for each segment, which constitutes the candidate feature set. The division of the 11 candidate features is depicted in Supplementary Fig. 4.

Although the available capacity decreases with degradation, the variation in charge capacity differs across segments. This variation complicates the determination of whether a strong correlation exists between these candidate features and the available capacity. To address this, the Pearson correlation coefficient (PCC) serves as an effective tool, with the calculation principle detailed in Supplementary Note 2⁴². Figure 4, plotting in various color blocks, visualizes the correlation between the candidate features and the available capacity. Notably, a strong positive correlation is evident between the available capacity and the charge capacity within segments 1, 2, 3, 5, 8, and 9. Conversely, segments 7 and 11 exhibit a negative correlation, indicating that as the battery degrades, the charge capacity within these segments increases rather than decreases. Segments 4, 6, and 10 demonstrate weak performance in the correlation analyses and are thus excluded as candidate features. The PCCs for the eight segments with high correlations can be found in Supplementary Table 5.

Available capacity estimation

Based on the candidate features filtered through segmentation, the randomized feature combinations shall be inputted into the data-driven methods to implement the available capacity estimation. Firstly, including more candidates in the feature combinations generally improves estimation accuracy. However, an excessive number of candidates increases the time required and complicates the feature combination process. Secondly, for practical engineering applications, rapid and accurate available capacity estimation through simpler feature combinations is preferred. Therefore, we use combinations containing one or two candidate features as inputs,

resulting in a total of 36 combinations (see Supplementary Table 6). Compared to the conventional method of obtaining available capacity based on complete discharges, the proposed method reduces the estimation time by an average of 70.9%.

Linear (LASSO) and non-linear models (XGBoost and LightGBM) are selected as ML methods. Supplementary Table 7 details the hyperparameters for these three algorithms. Dataset #1 is employed for both training and testing the basic model, with the four cells in this dataset divided into training and test sets in a 3:1 ratio. For datasets #2 and #3, we validate the performance of the TL model by fine-tuning the basic model⁴³. The applied model fine-tuning strategy is outlined in Supplementary Note 3. In addition, the data in dataset #1 require standardization due to the order-of-magnitude differences between the candidate features⁴⁴. It is worth noting that all standardization procedures discussed in this paper are applied after splitting the complete dataset into training and test sets. Supplementary Note 4 illustrates the process of standardization. Furthermore, because of variations in nominal capacity among these datasets, the data in Datasets #2 and #3 will also need to be standardized.

In most ML tasks and competitions, the test set is pre-designated, fixed, and entirely isolated from the training process. Particularly in the field of battery intelligence management, the test set typically consists of the complete data from one or more cells, which is used solely for evaluating the model's estimation performance. This pre-determined way of splitting datasets is also more in line with what happens in engineering practice.

Focusing on 36 feature combinations, we first train and test the basic model on Dataset #1 using the LASSO algorithm. During model training, K-fold cross-validation with $K=4$ is employed to identify the optimal hyperparameters. The principle of K-fold cross-validation is depicted in Supplementary Fig. 5. Excluding the feature combinations that included

Fig. 4 | Correlation between the candidate features and the available capacity. Pearson correlation coefficient is applied, as in Supplementary Note 2. Various color blocks correspond to differentiated correlations. The gradient from blue to red denotes to the correlation coefficient from small to large.

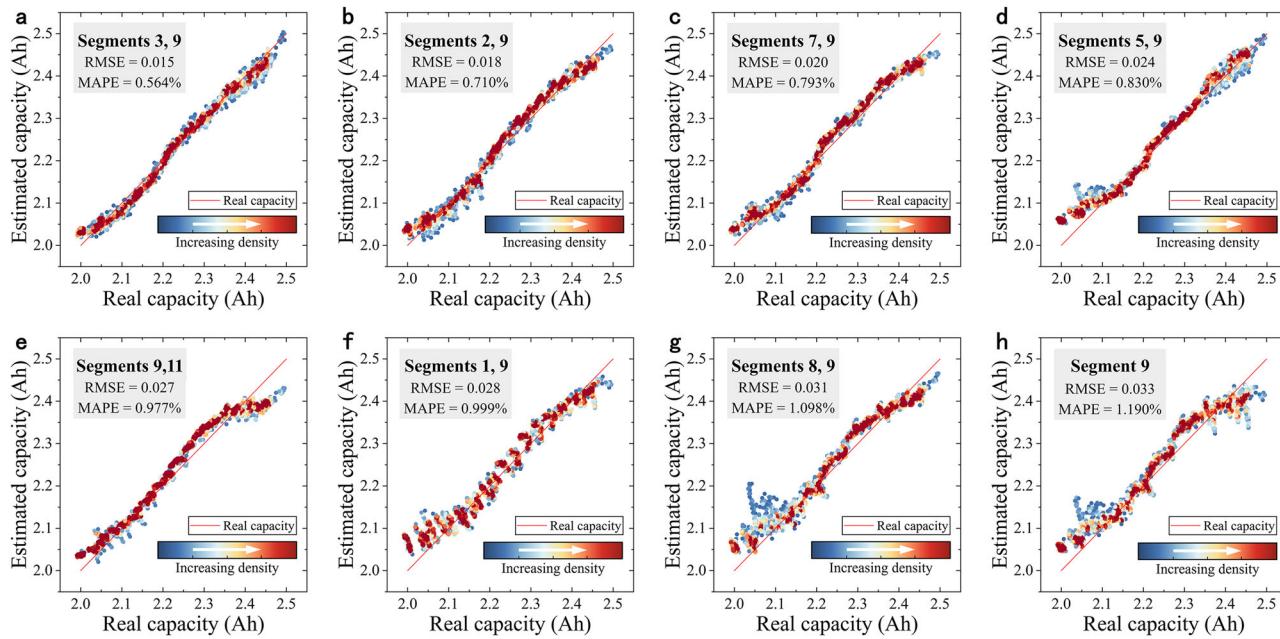
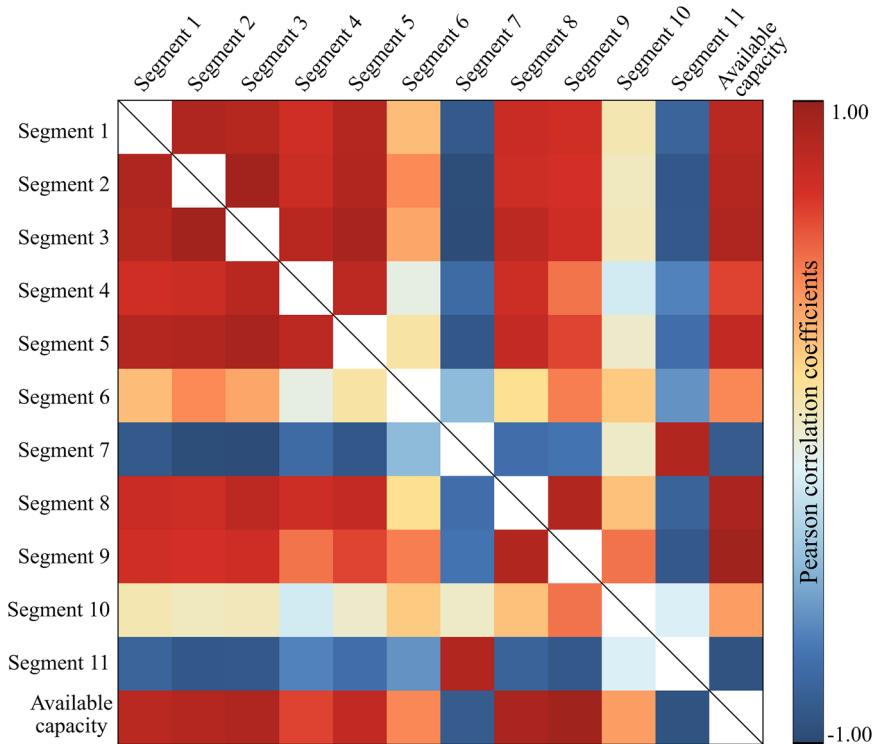


Fig. 5 | Visualization results of the LASSO algorithm for different feature combinations. The available capacity estimation performance is plotted for 8 feature combinations, and the results for the remaining 28 combinations are referenced in Supplementary Figs. 6–9. The 8 feature combinations include (segments 3, 9) (a),

(segments 2, 9) (b), (segments 7, 9) (c), (segments 5, 9) (d), (segments 9, 11) (e), (segments 1, 9) (f), (segments 8, 9) (g), and (segment 9) (h). The gradients correspond to varying point-aggregation densities. RMSE denotes root-mean-square error and MAPE denotes mean absolute percentage error.

only one candidate, the remaining 28 combinations comprised two candidate features each. To initially validate the feasibility of the proposed method and demonstrate its estimation accuracy, the two candidate features used for validation come from the same cycle or neighboring cycles. Figure 5 presents the visualization results from the LASSO algorithm for eight feature combinations, comparing the estimated capacity with the real capacity. It also includes the corresponding root-mean-square error (RMSE) and mean absolute percentage error (MAPE). The validation results under the

remaining 28 feature combinations are depicted in Supplementary Fig. 6 to Supplementary Fig. 9. It can be noticed that the RMSE and MAPE of a feature combination containing two candidates is not necessarily lower than that of a combination including only one candidate. Furthermore, it is interesting to consider that the feature combinations containing segment 9 exhibit even better estimation performance. Subsequently, we conduct model training and testing using XGBoost and LightGBM, with the results presented in Supplementary Tables 8–10. It can be concluded that both

XGBoost and LightGBM achieve an optimal RMSE of 0.012, demonstrating better estimation performance compared to the linear model (LASSO).

As previously mentioned, we thoroughly consider the estimation performance when one or two candidate features are included in the input feature combination. While feature combinations can theoretically include two or more candidates, this increases the complexity of obtaining such combinations. Hence, it is imperative to investigate whether ML models can demonstrate more reliable estimation performance when the input feature combination includes multiple candidates. To address these concerns, we construct six typical feature combinations based on correlation analysis and explore the model performance using the LASSO algorithm. Supplementary Table 11 presents the estimation performance for input feature combinations containing 3 to 8 candidates, with RMSE and MAPE used as evaluation metrics. The validation results indicate that input combinations containing more candidate features do not obviously enhance the estimation performance compared to the feature combination (segments 3, 9), as evidenced by the RMSEs remaining near 0.015. This suggests that more complex feature combinations do not improve estimation accuracy but rather increase computational costs and complicate feature engineering.

To further substantiate the practical applicability, reasonableness, and superiority of the proposed method in engineering practice, this paper conducts validation experiments using a public dataset under actual operating conditions⁴⁵. This dataset contains three LFP cells with a nominal capacity of 180 Ah, which were degraded using a realistic forklift load profile at three elevated temperatures (45 °C, 40 °C, and 35 °C), respectively. The three cells are designated as cell #1, cell #2, and cell #3. The profile used for aging is based on actual forklift operations, which resemble the operating mode of electric vehicles (i.e., dynamic discharge followed by fragmented charging). Each aging experiment cycle lasted two weeks, resulting in a total of 169 experiments conducted across the three cells. Capacity tests were performed between every two aging experiments to generate the available capacity labels. Supplementary Fig. 10(a) presents the available capacity degradation curves for the three cells. We conduct a detailed analysis of the fragmented charging processes during the aging experiments and draw the following conclusions: (I) each aging experiment includes more than 100 charging processes, (II) the primary charging protocol is 24 A CC charging, and (III) over 90% of the charging processes cover the voltage range of 3.3V–3.36 V. Consequently, we slightly adjust the voltage ranges of the 11 segments to better align with this dataset. If the model maintains good estimation performance with these adjustments, it will underscore the applicability of the method.

Since each aging experiment provides only one available capacity label, we apply linear interpolation to ensure that each charging process is associated with a corresponding label. Under the LightGBM algorithm, cells #1 and #2 are used as the training set, while cell #3 serves as the test set. Standardization is then performed separately for the training and test sets.

(Segment 3), (segment 4), and the combination of (segments 3, 4) are used as input features to evaluate the performance of the proposed method. Supplementary Fig. 10(b-d) describes the visualization results. The validation results indicate that the test RMSEs for (segment 3) and (segment 4) are 0.024 and 0.027, respectively. The optimal RMSE of 0.010 is achieved with the feature combination of (segments 3 and 4). The above results further validate the practical applicability, reasonableness, and superiority of the proposed method in engineering practice. A recent study achieved available capacity estimation on the forklift dataset using a wide voltage interval⁴⁵. They reported an estimation performance with an RMSE of 0.017, which is weaker than the 0.010 noted in this paper. In a certain sense, the fragmented charging arising from actual operating conditions does not adversely affect the performance of the proposed method.

Feature combinations across cycles

In contrast to feature combinations containing only one candidate, more complex combinations encounter challenges in effective feature extraction. Specifically, suitable candidate features may not be present within the same cycle or neighboring cycles. This difficulty arises because EV charging is an irregular and fragmented process, influenced by the subjective behavior of drivers during transportation electrification. Over 3.7 million charging sessions from public charging stations were analyzed, demonstrating that charging behavior is influenced by a multitude of complex factors⁴⁶. EV data from 79 users was discussed, revealing obvious individual variability in charging preferences; for example, some users chose to charge even when the battery state of charge (SOC) was still high⁴⁷. In addition, the configuration of charging stations or piles drastically affects the charging choice, necessitating long-term coordination and optimization between charging habits and infrastructure⁴⁸. Relying solely on candidate features from the same or neighboring cycles for available capacity estimation lacks flexibility. Consequently, incorporating feature combinations across cycles is likely to become a prospective trend in the proposed method. We anticipate that combining two candidates from separate cycles as input features will enable accurate estimation of the available capacity of LIBs. The process for creating feature combinations across cycles is detailed in Supplementary Note 5, with the visualization procedure illustrated in Supplementary Fig. 11.

Following the above description, we adjust the dataset for the initial validation of the basic model so that the spacing n between the two candidate features at 5, 10, and 20. The XGBoost algorithm is selected to validate several typical feature combinations, with the visualization results for the combination (segments 3, 9) illustrated in Fig. 6. Supplementary Fig. 12 and Supplementary Table 12 demonstrate the visualization results and quantification findings from the other feature combinations. It can be intuitively represented that the RMSE and MAPE of the feature combination tend to increase as the spacing n expands; however, the magnitude of this increase is minor. For this feature combination (segments 3, 9), compared with $n = 0/1$,

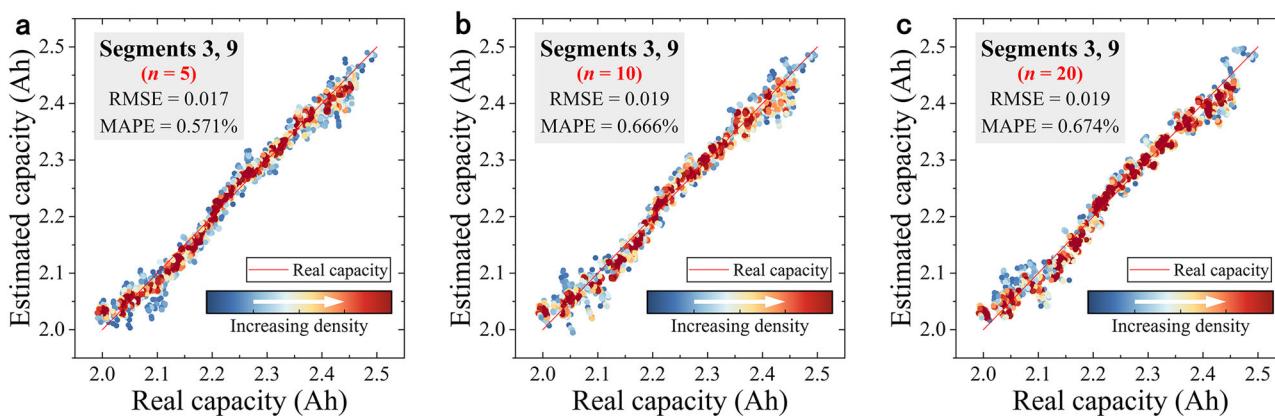


Fig. 6 | Performance of feature combinations across cycles under (Segments 3, 9). The spacing n between the two candidate features is taken as 5 (a), 10 (b), and 20 (c). The gradients correspond to varying point-aggregation densities. RMSE denotes root-mean-square error and MAPE denotes mean absolute percentage error.

Table 2 | Comparison with existing methods

Methods	Input features	Best test RMSE on dataset #1
Direct extraction method	Charge capacity in the optimal voltage interval	0.034
	Charging time in the optimal voltage interval	0.035
Indirect extraction method	Incremental capacity peak	0.020
Proposed method	Multiple feature combinations	0.012

the RMSE and MAPE increase by 0.005 and 0.160%, respectively, when $n = 5$. Compared with $n = 5$, the RMSE and MAPE increase by an additional 0.002 and 0.103% when $n = 20$. It can be concluded that the estimation performance of the proposed method is not markedly affected when the two candidate features are derived from separate cycles. The preceding discussion confirms the applicability and robustness of the proposed method in practical scenarios, suggesting its potential for online application in EVs.

It is evident that feature combinations across cycles provide an intriguing solution for feature engineering in practical applications. This research demonstrates the effectiveness of this method in efficiently, accurately, and continuously estimating available capacity and assessing battery performance. While our investigation and discussions have preliminarily validated the feasibility and applicability of the method on a limited number of cells, highlighting its broad potential for application, caution is warranted. The similarity in degradation mechanisms among these cells may lead to model overfitting and overestimation of performance. Therefore, further in-depth and extensive research is essential to achieve accurate estimation using fragmented data from different cycles. Dataset #4, comprising 32,800 cyclic samples from 50 cells, provides a diverse array of degradation mechanisms and degradation curves, as evidenced by the distribution of cyclic samples per cell, which ranges from 170 to 1154. To comprehensively validate the implementability of feature combinations across cycles, we conduct model training and evaluation without relying on TL techniques or additional datasets. For dataset #4, the proposed method is extended to apply to discharge curves rather than charge curves, while simultaneously adjusting the voltage range distribution of the 11 segments. If the model maintains strong estimation performance under these adjustments and fluctuations, it will demonstrate the robustness of the method.

Under the XGBoost algorithm, 35 cells from dataset #4 participate in model training, while the remaining 15 cells serve as the test set. A detailed partitioning of the training and test sets can be found in Supplementary Table 13. The feature combination (segments 7, 8) is chosen for use in feature combinations across cycles. Initially, we assess the estimation performance when the two candidate features are extracted within the same cycle or adjacent cycles, corresponding to the spacing n of 0/1. Subsequently, we adjust the test set to achieve spacings n of 5, 10, and 20 between the two candidate features. Finally, the validation results from 15 cells are summarized. Supplementary Table 14 elucidates the estimation performance of feature combinations across cycles on dataset #4, utilizing RMSE as the evaluation metric. The validation results reveal that the best RMSE is 0.011 on cell #3 when $n = 0/1$, while the average RMSE on the test set is 0.0156. Similar to the performance observed on dataset #1, there is a slight increase in RMSE as the spacing n gradually expands. Compared with $n = 0/1$, the average RMSE under $n = 5$ increases by only 0.0004, indicating an improvement in estimation performance over dataset #1. Furthermore, compared with $n = 5$, the average RMSE under $n = 20$ increases by 0.002, ensuring the stability of estimation performance as the spacing n expands. Therefore, on dataset #4, feature combinations across cycles consistently maintain excellent estimation performance. Additionally, the exploration of integrating fragmented discharge data without relying on TL highlights the robustness and superiority of the proposed method.

Comparison with existing methods

Reliable available capacity estimation for LIBs provides a crucial reference for SOH and RUL calculations. The feature extraction process greatly

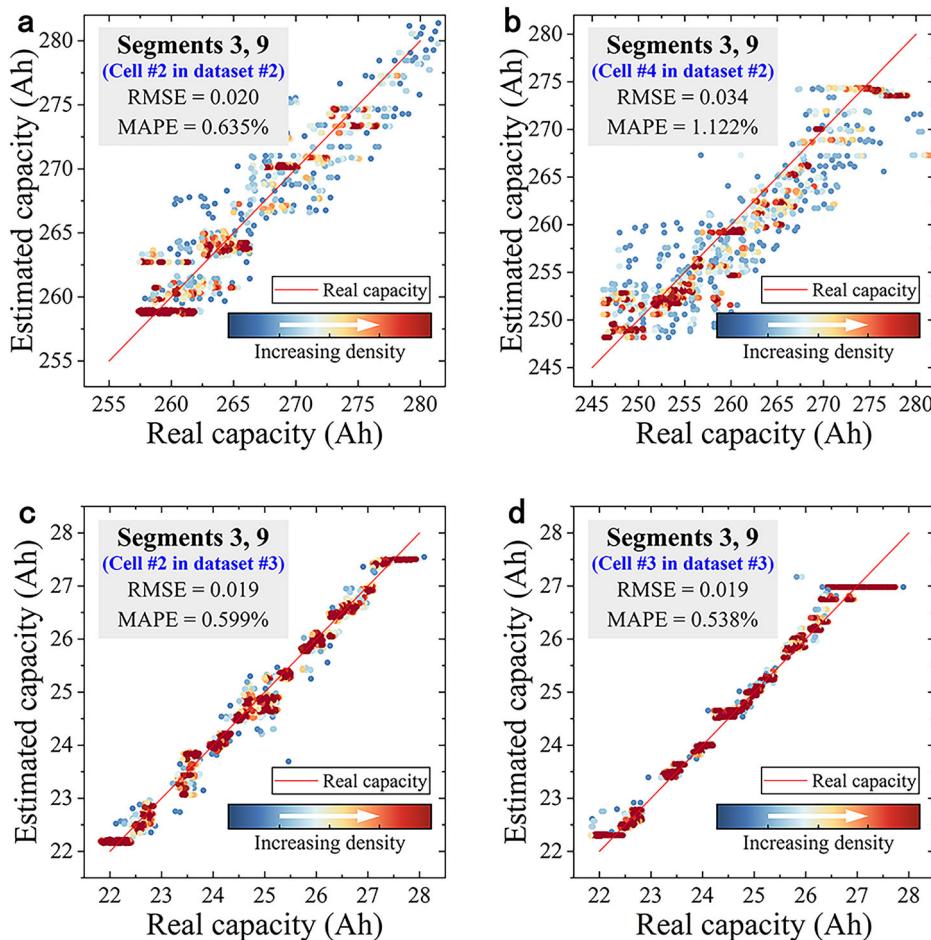
influences both the practical application and estimation performance of data-driven methods. Relevant literature suggests that feature engineering should consider the actual usage characteristics of LIBs, with a focus on extracting features from various charging stages. Feature extraction methods can be broadly categorized into four types: (I) CC charging-based, (II) CC-CV charging-based, (III) CV charging-based, and (IV) voltage relaxation-based. However, due to the infrequent occurrence of CV charging in engineering applications, the implementation of methods (II) and (III) poses enormous challenges. Voltage relaxation is defined as the open-circuit voltage of the battery measured after a full charge and a 30-minute rest period⁴⁹. However, full charging and extended resting periods are not aligned with the actual operational practices of EVs. Therefore, CC charging-based feature extraction methods offer superior practicality and interpretability. The XGBoost algorithm is chosen for comparison with existing methods.

CC charging-based methods can be further divided into direct and indirect extraction methods, as outlined in Supplementary Table 15. Since the datasets used in other literature differ from those in this paper, past indicators cannot serve as reliable comparisons. To address this discrepancy, we extract the corresponding features from our dataset and perform available capacity estimation, as demonstrated in Table 2. In the mainstream literature, charge capacity⁵⁰ or charging time⁵¹ within the optimal voltage interval is commonly used as input features. The optimal voltage interval is defined as the range in which the strongest linear relationship exists between internal features and battery degradation. In this paper, the optimal voltage interval is identified as (3.415 V, 3.43 V) within segment 9, as illustrated in Supplementary Table 5. The RMSE values for dataset #1 are 0.034 and 0.035 when using the charge capacity and charging time of segment 9, respectively. While some scholars have opted for larger voltage intervals for feature extraction, this approach clearly increases the difficulty of data acquisition and reduces generalizability^{52,53}. The incremental capacity (IC) curve, typically obtained through data interpolation or filtering methods, is also frequently utilized for estimation⁵⁴. When the IC peak⁵⁵ is selected as an input feature, the RMSE on dataset #1 is 0.020. Additionally, features such as the IC peak area⁵⁶ and IC peak slope⁵⁷ are commonly applied as input features. However, the complexity of curve processing and the rigidity of feature extraction in these methods make them challenging to use in online scenarios. In contrast, the proposed method offers more flexible feature engineering and achieves better estimation accuracy, especially when dealing with fragmented charging data.

Performance validation by transfer learning

Scholars have explored how a basic model trained on an original dataset can maintain excellent performance across different datasets. Variations in usage intensity, ambient temperature, and unbalanced charging, among other factors, contribute to differences between batteries. For the first three proposed datasets, we analyze the correlations among them, as discussed in Supplementary Note 2. Supplementary Table 16 and Supplementary Fig. 13 examine the correlations between the same candidate features and between available capacities. First, in comparison to dataset #1, datasets #2 and #3 show a strong positive correlation in available capacity, with PCCs exceeding 0.8, indicating similar degradation patterns. Subsequently, in dataset #2, cells #1 and #2 (with an ambient temperature of 35 °C) exhibit weak correlations in eight segments, while cells #3 and #4 (with an ambient temperature of 45 °C) show weak correlations in six segments. However, in

Fig. 7 | Test results of estimated capacity versus real capacity by transfer learning under (segments 3, 9). Tests are performed on 4 cells in datasets #2 and #3. Cell #2 in dataset #2 (a), cell #4 in dataset #2 (b), cell #2 in dataset #3 (c), and cell #3 in dataset #3 (d). Results under (segments 3, 8) are presented in Supplementary Fig. 14. The gradients correspond to varying point-aggregation densities. RMSE denotes root-mean-square error and MAPE denotes mean absolute percentage error.



dataset #3, only two segments display weak correlations. These differences among the datasets impact the performance of the basic model.

To accommodate the variations in cycle conditions in datasets #2 and #3, we apply TL technology, which enhances estimation performance by fine-tuning the model with small amounts of new data. LightGBM is used to support the performance verification of TL. The model fine-tuning strategy is detailed in Supplementary Note 3. TL is implemented by adding one or more fully connected layers to the top layer of the basic model. Zero-shot learning (ZSL), without model fine-tuning, is employed as a comparison and reference⁵⁸. Figure 7 presents the visualization results for the feature combination (segments 3, 9). TL performs well on cells #2 and #3 in dataset #3, achieving a test RMSE of 0.019 for both. However, for cell #4 in dataset #2, the RMSE reaches 0.034, likely due to differences between the datasets. Subsequently, we test the feature combination (segments 3, 8), as shown in Supplementary Fig. 14. The estimation performance for cell #4 in dataset #2 improves, with a test RMSE of only 0.016. Nonetheless, this improvement is accompanied by a decline in performance for the other cells. The RMSEs are compared in Table 3. It can be concluded that the application of TL helps mitigate the differences between datasets and enhances estimation accuracy. Any performance fluctuations observed in individual cells across certain datasets can be optimized by adjusting the feature combinations. In summary, the proposed method offers flexible and effective available capacity estimation and demonstrates satisfactory performance without the need for model reconstruction under TL.

Discussion

Existing technologies for estimating battery available capacity are constrained by labor-intensive feature engineering, making them inefficient for application in EVs with irregular charging and discharging patterns. In this

Table 3 | RMSEs of battery available capacity estimation using zero-shot learning (ZSL) and transfer learning (TL) on dataset #2 and dataset #3

Dataset	Input feature combinations	Methods	
		TL	ZSL
Cell #2 in dataset #2	Segments 3, 9	0.020	0.304
	Segments 3, 8	0.022	0.213
Cell #4 in dataset #2	Segments 3, 9	0.034	0.266
	Segments 3, 8	0.016	0.170
Cell #2 in dataset #3	Segments 3, 9	0.019	0.242
	Segments 3, 8	0.024	0.215
Cell #3 in dataset #3	Segments 3, 9	0.019	0.250
	Segments 3, 8	0.027	0.222

work, we propose a solution to estimate available capacity by effectively manipulating fragmented charging data without relying on complete charging/discharging information. Briefly, we have successfully incorporated the concept of 'less is more' into our technological applications. By extracting charge capacity across various voltage intervals as candidate features, we form 36 potential feature combinations to serve as inputs to the method. Compared to estimation methods that rely on a fixed (optimal) voltage interval, the proposed method offers a more flexible, correlated, and practical feature engineering. As a forward-looking research, four datasets from different experimental conditions and battery manufacturers, encompassing more than 50,000 samples, are utilized in the data-driven

method. Three ML algorithms guide the basic model training and test. We first demonstrate that the optimal RMSE for the dataset used to establish the basic model is 0.012. Moreover, the estimation performance remains stable when the feature combination process is extended across cycles, with the average RMSE increasing by only 0.002, demonstrating its robustness in real-world application scenarios. In addition, the method applicability is ensured by model fine-tuning for TL retraining on other datasets. Compared to ZSL, the RMSE under TL is lower by an average of 0.213.

In conclusion, our work highlights the potential of utilizing fragmented data for battery available capacity estimation in the context of irregular EV charging and discharging. Multiple feature combinations and an extraction process without post-processing make feature engineering more flexible and convenient. This contribution advances the online application of data-driven available capacity estimation methods in EVs and introduces a technological approach for areas such as the rapid detection of LIBs. In future research, we plan to further validate and refine the proposed method by testing it on large-scale battery packs or energy storage systems. Additionally, through collaboration with battery manufacturers and automotive companies, we aim to actively promote the application of the proposed method in real-world scenarios, particularly within EVs.

Methods

Cell cycling and dataset generation

Datasets #1 and #2 are obtained by cycling cells under various experimental conditions. Specifically, the cell in dataset #1 is tested in a thermal chamber at 25 °C. Cells #1 and #2 in dataset #2 are exposed to an ambient temperature of 35 °C, while cells #3 and #4 are tested at 45 °C. Detailed environmental information, such as charging protocol, discharging C-rate, and type, can be found in Supplementary Table 3. The battery charging and discharging test system is employed to perform the cycling steps, while the thermal chamber provides a temperature-stabilized test environment. The equipment connection principle schematic is presented in Supplementary Fig. 15. In the experiment, the resting process is essential for enhancing the performance and stability of the cell. However, this paper does not address the resting process, so this part is eliminated before the dataset generation. The final dataset contains the complete charge and discharge curves from the cycles.

Machine learning methods

Our work involves three basic algorithms in ML, including the LASSO, XGBoost, and LightGBM algorithms. Each of these algorithms has a well-established history and distinct characteristics, making them widely applicable across various fields. Detailed introductions to the LASSO, XGBoost, and LightGBM algorithms are provided in Supplementary Notes 6–8, respectively. The procedures for basic model training and the implementation of transfer learning are briefly outlined below:

(1) In dataset #1, cells #1, #2, and #3 are designated as the training set, while cell #4 serves as the test set. All three algorithms are utilized to train and test the basic model to verify its feasibility.

(2) In this paper, zero-shot learning (ZSL) is employed as a comparison method for transfer learning (TL). The basic model, without any modifications, is directly validated on datasets #2 and #3 to implement ZSL.

(3) TL is initiated by adding one or more fully connected layers to the top layer of the basic model. This research adopts a model fine-tuning strategy (as detailed in Supplementary Note 3) to adjust the basic model. The transfer learning model is validated on partial cells in datasets #2 and #3.

In addition, this work uses RMSE and MAPE to evaluate the estimation performance of the proposed method. Both are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_{i,\text{Real}} - y_{i,\text{Estimated}})^2} \quad (1)$$

$$\text{MAPE} = \frac{100\%}{m} \sum_{i=1}^m \left| \frac{y_{i,\text{Estimated}} - y_{i,\text{Real}}}{y_{i,\text{Real}}} \right| \quad (2)$$

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability

Code for the modeling work is available from the corresponding authors upon request.

Received: 24 January 2024; Accepted: 12 February 2025;

Published online: 25 February 2025

References

1. Tarascon, J. M. & Armand, M. Issues and challenges facing rechargeable lithium batteries. *Nature* **414**, 359–367 (2001).
2. Harper, G. et al. Recycling lithium-ion batteries from electric vehicles. *Nature* **575**, 75–86 (2019).
3. Camargos, P. H., dos Santos, P. H., dos Santos, I. R., Ribeiro, G. S. & Caetano, R. E. Perspectives on Li-ion battery categories for electric vehicle applications: a review of state of the art. *Int. J. Energy Res.* **46**, 19258–19268 (2022).
4. Baek, M., Kim, J., Jin, J. & Choi, J. W. Photochemically driven solid electrolyte interphase for extremely fast-charging lithium-ion batteries. *Nat. Commun.* **12**, 6807 (2021).
5. Liu, Y., Zhu, Y. & Cui, Y. Challenges and opportunities towards fast-charging battery materials. *Nat. Energy* **4**, 540–550 (2019).
6. Togasaki, N., Yokoshima, T., Oguma, Y. & Osaka, T. Prediction of overcharge-induced serious capacity fading in nickel cobalt aluminum oxide lithium-ion batteries using electrochemical impedance spectroscopy. *J. Power Sources* **461**, 228168 (2020).
7. Palacín, M. R. & de Guibert, A. Why do batteries fail? *Science* **351**, 1253292 (2016).
8. Lu, L., Han, X., Li, J., Hua, J. & Ouyang, M. A review on the key issues for lithium-ion battery management in electric vehicles. *J. Power Sources* **226**, 272–288 (2013).
9. Hu, X., Xu, L., Lin, X. & Pecht, M. Battery lifetime prognostics. *Joule* **4**, 310–346 (2020).
10. Yi, T., Zhang, C., Lin, T. & Liu, J. Research on the spatial-temporal distribution of electric vehicle charging load demand: a case study in China. *J. Clean. Prod.* **242**, 118457 (2020).
11. Philipsen, R., Brell, T., Brost, W., Eickels, T. & Ziefle, M. Running on empty—users' charging behavior of electric vehicles versus traditional refueling. *Transp. Res. Part F Traffic Psychol. Behav.* **59**, 475–492 (2018).
12. Hidrue, M. K., Parsons, G. R., Kempton, W. & Gardner, M. P. Willingness to pay for electric vehicles and their attributes. *Resour. Energy Econ.* **33**, 686–705 (2011).
13. Roman, D., Saxena, S., Robu, V., Pecht, M. & Flynn, D. Machine learning pipeline for battery state-of-health estimation. *Nat. Mach. Intell.* **3**, 447–456 (2021).
14. Wang, X. et al. Non-damaged lithium-ion batteries integrated functional electrode for operando temperature sensing. *Energy Storage Mater.* **65**, 103160 (2024).
15. Liu, K., Shang, Y., Ouyang, Q. & Widanage, W. D. A data-driven approach with uncertainty quantification for predicting future capacities and remaining useful life of lithium-ion battery. *IEEE Trans. Ind. Electron.* **68**, 3170–3180 (2020).
16. Fly, A. & Chen, R. Rate dependency of incremental capacity analysis (dQ/dV) as a diagnostic tool for lithium-ion batteries. *J. Energy Storage* **29**, 101329 (2020).
17. Mohtat, P., Lee, S., Siegel, J. B. & Stefanopoulou, A. G. Comparison of expansion and voltage differential indicators for battery capacity fade. *J. Power Sources* **518**, 230714 (2022).
18. Wu, Y. & Jossen, A. Entropy-induced temperature variation as a new indicator for state of health estimation of lithium-ion cells. *Electrochim. Acta* **276**, 370–376 (2018).

19. Khodadadi Sadabadi, K., Jin, X. & Rizzoni, G. Prediction of remaining useful life for a composite electrode lithium ion battery cell using an electrochemical model to estimate the state of health. *J. Power Sources* **481**, 228861 (2021).
20. Knehr, K. W. et al. Understanding full-cell evolution and nonchemical electrode crosstalk of Li-ion batteries. *Joule* **2**, 1146–1159 (2018).
21. Zou, Y., Lin, Z., Li, D. & Liu, Z. Advancements in artificial neural networks for health management of energy storage lithium-ion batteries: a comprehensive review. *J. Energy Storage* **73**, 109069 (2023).
22. Severson, K. A. et al. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **4**, 383–391 (2019).
23. Zhang, Y. et al. Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. *Nat. Commun.* **11**, 1706 (2020).
24. Lu, J., Xiong, R., Tian, J., Wang, C. & Sun, F. Deep learning to estimate lithium-ion battery state of health without additional degradation experiments. *Nat. Commun.* **14**, 2760 (2023).
25. Jiang, B., Dai, H. & Wei, X. Incremental capacity analysis based adaptive capacity estimation for lithium-ion battery considering charging condition. *Appl. Energy* **269**, 115074 (2020).
26. Zhu, J. et al. Data-driven capacity estimation of commercial lithium-ion batteries from voltage relaxation. *Nat. Commun.* **13**, 2261 (2022).
27. Duru, K. K. et al. Critical insights into fast charging techniques for lithium-ion batteries in electric vehicles. *IEEE Trans. Device Mater. Reliab.* **21**, 137–152 (2021).
28. Khaleghi, S. et al. Developing an online data-driven approach for prognostics and health management of lithium-ion batteries. *Appl. Energy* **308**, 118348 (2022).
29. Yang, J., Xia, B., Huang, W., Fu, Y. & Mi, C. Online state-of-health estimation for lithium-ion batteries using constant-voltage charging current analysis. *Appl. Energy* **212**, 1589–1600 (2018).
30. Deng, Z. et al. Prognostics of battery capacity based on charging data and data-driven methods for on-road vehicles. *Appl. Energy* **339**, 120954 (2023).
31. Wang, J. et al. A novel aging characteristics-based feature engineering for battery state of health estimation. *Energy* **273**, 127169 (2023).
32. Lin, C., Xu, J., Shi, M. & Mei, X. Constant current charging time based fast state-of-health estimation for lithium-ion batteries. *Energy* **247**, 123556 (2022).
33. Li, K. et al. Battery life estimation based on cloud data for electric vehicles. *J. Power Sources* **468**, 228192 (2020).
34. Attia, P. M. et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature* **578**, 397–402 (2020).
35. Zheng, Y., Qin, C., Lai, X., Han, X. & Xie, Y. A novel capacity estimation method for lithium-ion batteries using fusion estimation of charging curve sections and discrete Arrhenius aging model. *Appl. Energy* **251**, 113327 (2019).
36. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
37. Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 785–794 (2016).
38. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30** 3149–3157 (2017).
39. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2009).
40. Cripps, E. & Pecht, M. A Bayesian nonlinear random effects model for identification of defective batteries from lot samples. *J. Power Sources* **342**, 342–350 (2017).
41. Zhu, J. et al. A method to prolong lithium-ion battery life during the full life cycle. *Cell Rep. Phys. Sci.* **4**, 101464 (2023).
42. Schober, P., Boer, C. & Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. *Anesth. Analg.* **126**, 1763–1768 (2018).
43. Shin, H. C. et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).
44. Gal, M. S. & Rubinfeld, D. L. Data standardization. *NYUL Rev.* **94**, 737 (2019).
45. Vilzen, S. B. & Stroe, D. I. On the use of randomly selected partial charges to predict battery state-of-health. *Batteries* **10**, 193 (2024).
46. Wolbertus, R., Kroesen, M., Van Den Hoed, R. & Chorus, C. Fully charged: An empirical study into the factors that influence connection times at EV-charging stations. *Energy Policy* **123**, 1–7 (2018).
47. Franke, T. & Krems, J. F. Understanding charging behaviour of electric vehicle users. *Transp. Res. Part F Traffic Psychol. Behav.* **21**, 75–89 (2013).
48. Chen, J., Li, F., Yang, R. & Ma, D. Impacts of increasing private charging piles on electric vehicles' charging profiles: a case study in Hefei City, China. *Energies* **13**, 4387 (2020).
49. Baghdadi, I., Briat, O., Gyan, P. & Vinassa, J. M. State of health assessment for lithium batteries based on voltage-time relaxation measure. *Electrochim. Acta* **194**, 461–472 (2016).
50. Zheng, Y. et al. A novel capacity estimation method based on charging curve sections for lithium-ion batteries in electric vehicles. *Energy* **185**, 361–371 (2019).
51. Shu, X. et al. A flexible state-of-health prediction scheme for lithium-ion battery packs with long short-term memory network and transfer learning. *IEEE Trans. Transp. Electrif.* **7**, 2238–2248 (2021).
52. Li, W. et al. Online capacity estimation of lithium-ion batteries with deep long short-term memory networks. *J. Power Sources* **482**, 228863 (2021).
53. Li, Y. et al. Random forest regression for online capacity estimation of lithium-ion batteries. *Appl. Energy* **232**, 197–210 (2018).
54. Zhang, S. et al. Synchronous estimation of state of health and remaining useful lifetime for lithium-ion battery using the incremental capacity and artificial neural networks. *J. Energy Storage* **26**, 100951 (2019).
55. Weng, C., Feng, X., Sun, J. & Peng, H. State-of-health monitoring of lithium-ion battery modules and packs via incremental capacity peak tracking. *Appl. Energy* **180**, 360–368 (2016).
56. Agudelo, B. O., Zamboni, W. & Monmasson, E. Application domain extension of incremental capacity-based battery SoH indicators. *Energy* **234**, 121224 (2021).
57. Li, X., Yuan, C. & Wang, Z. Multi-time-scale framework for prognostic health condition of lithium battery using modified Gaussian process regression and nonlinear regression. *J. Power Sources* **467**, 228358 (2020).
58. Wang, W., Zheng, V. W., Yu, H. & Miao, C. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.* **10**, 1–37 (2019).

Acknowledgements

The project is supported in part by the National Natural Science Foundation of China under Grant U24A20159, Grant 62122041, Grant 62333013, and Grant 62173211, in part by the Natural Science Foundation of Shandong Province, China, under Grant ZR2021JQ25.

Author contributions

Z.Z. conceived the original idea of available capacity estimation, designed the research, and wrote the initial draft of the manuscript. The experimental studies were performed by Z.Z., X.G., and Y.Z. The computational studies were performed by Z.Z., T.W., and Y.G. Y.S. led and supervised the project, participated in paper writing and revision, and provided guidance to all co-authors. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at
<https://doi.org/10.1038/s44172-025-00372-y>.

Correspondence and requests for materials should be addressed to Yunlong Shang.

Reprints and permissions information is available at
<http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025