Recovering large-scale battery aging dataset with machine learning

Graphical abstract

Highlights

- A new approach to generate high-quality large-volume battery aging datasets

- Our data-generating method saves up to 90% of the battery aging experimental time

- The generated datasets exhibit an ultra-low error bounded of 1% only

- The first attempt to regenerate data from industrial calibrations like EV maintenance
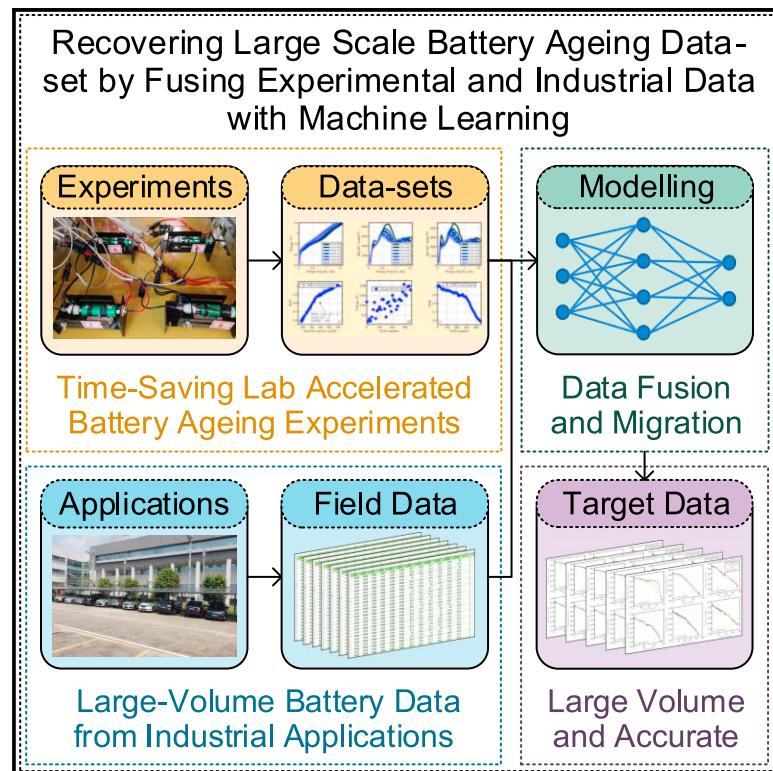
Authors

Xiaopeng Tang, Kailong Liu, Kang Li,
Widanalage Dhammika Widanage,
Emma Kendrick, Furong Gao

Correspondence

kailong.liu@warwick.ac.uk

In brief

Limited access to sufficient battery aging data significantly hinders the advancement of battery health assessment, which is critical for building a clean and sustainable society. Compared with carrying out large-scale aging experiments in the laboratory, this work provides an alternative solution to produce large-volume, high-quality aging datasets. While saving up to 90% experimental time, aging data can be recovered with ultra-low error within 1%, significantly improving data shortage issues in the field of battery and energy storage aging assessment.

CellPress

Article

Recovering large-scale battery aging dataset with machine learning

Xiaopeng Tang,[1] Kailong Liu,[2,7,*] Kang Li,[4] Widanalage Dhammika Widanage,[2,3] Emma Kendrick,[5,3] and Furong Gao[1,6]

[1]Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR 999077, China
[2]WMG, The University of Warwick, Coventry CV4 7AL, UK
[3]The Faraday Institution, Quad One, Becquerel Avenue, Harwell Campus, Didcot OX11 0RA, UK
[4]School of Electronic and Electric Engineering, University of Leeds, Leeds LS2 9JT, UK
[5]School of Metallurgy and Materials, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
[6]Guangzhou HKUST Fok Ying Tung Research Insititue, Guangdong 511458, China
[7]Lead contact
*Correspondence: kailong.liu@warwick.ac.uk
https://doi.org/10.1016/j.patter.2021.100302

THE BIGGER PICTURE Battery health assessment is crucial for future societal sustainability but it requires extensive and considerable data-sets with a diverse range of usage scenarios. Access to sufficient battery aging data is challenging due to both the complicated battery operations and also the significantly laborious and time-consuming battery aging experiments.

We here highlight the potential of combining industrial data with accelerated aging tests for rapid aging data acquisition. We present a machine learning approach to generate high-fidelity large-volume datasets with only 1% error, while the required experimental time can be saved by up to 90%. Our new solution can significantly improve the data-shortage issues for battery aging assessment. The regenerated dataset with representative battery aging dynamic information can benefit many data-based applications such as battery health estimation, lifespan prediction, second-life utilization, and fault diagnostics.

1 2 3 4 5   **Concept:** Basic principles of a new data science output observed and reported

---

SUMMARY

Batteries are crucial for building a clean and sustainable society, and their performance is highly affected by aging status. Reliable battery health assessment, however, is currently restrained by limited access to sufficient aging data, resulting from not only complicated battery operations but also long aging experimental time (several months or even years). Refining industrial datasets (e.g., the field data from electric vehicle batteries) unlocks opportunities to acquire large-volume aging data with low experimental efforts. We introduce the potential of combining industrial data with accelerated aging tests to recover high-quality battery aging datasets, through a migration-based machine learning. A comprehensive dataset containing 8,947 aging cycles with 15 operational modes is collected for evaluation. While saving up to 90% experimental time, aging data can be recovered with ultra-low error within 1%. It provides an alternative solution to significantly improve data shortage issues for assessment of battery and energy storage aging.

INTRODUCTION

Lithium-ion (Li-ion) batteries have been widely viewed as a key energy storage technology to support the transition to a clean and sustainable society.[1,2] However, the battery aging process will inevitably reduce the battery performance and reliability, further influence users' confidence, and hinder the advancement of the related battery applications, e.g., in transportation de-carboniza-

tion and renewable energy storage.[3,4] The data-based battery aging assessment is emerging as a complementary approach to address the inherent complexity of battery system modeling, achieve accurate estimation and prediction of battery capacity, and accelerate technology transfer from academic research to industry.[5–7] Moreover, it could unlock new opportunities for the health-aware control/management of batteries. For examples, fast charging scheme could be designed for different

aging conditions to balance the charging speed and degradation rate.[8–10] Likewise, users can get early warnings of the battery failure,[11,12] determine whether a battery should be replaced,[13,14] and if the retired batteries could be adopted for the less-demanding applications to allow cascade utilization.[15–17] Most of these data-based applications, by nature, require a large amount of available battery aging data.[18] Although the importance of battery degradation has been well addressed, how to quickly generate large-scale battery aging datasets is still an open research question. This task is critically important given the ever-increasing utilities of data-based aging assessment but significantly challenging due to not only complicated battery operations but also high experimental cost and long aging time of the batteries.

The state-of-the-art and most straightforward solution to generate aging data is designing large-scale lab experiments, where numerous batteries are tested with hundreds or thousands of equivalent cycles.[19,20] However, the aging-related testings in the lab are intensely laborious and time consuming. For instance, a battery aging experiment to obtain the entire degraded capacity trajectory generally requires several months or even years.[21,22] In addition, the aging process can be significantly influenced by the user's operating modes, for example, the charging rates. The coverage of these modes is commonly limited by the laboratory resources. Even for some highly acclaimed open-access datasets, e.g., provided by the Center for Advanced Life Cycle Engineering (CALCE)[21] and the National Aeronautics and Space Administration (NASA),[23] the lab experiments could only test limited battery types with a few load profiles. The resulting data shortage issue significantly limits the accuracy, reliability, and scalability of existing data-based approaches.

To tackle the aforementioned open challenges, we propose to generate the battery aging dataset by exploring field data from industrial systems, such as electric vehicles (EVs) or smart grid energy storage facilities. In these applications, parameters, such as the battery's voltage or current, are measured constantly by the embedded battery management systems in real time, and uploaded to big data platforms periodically for the purpose of data recording and further analysis.[24,25] Compared with the lab-generated aging datasets, the field-acquired ones have two distinguishable characteristics.

First, these kinds of datasets are often significantly larger than those collected from laboratory experiments. For instance, the National Monitoring and Management Center for EVs in China[26,27] can collect the real-time operating data of EVs at a frequency of one sample per 10 s, resulting in an accumulated dataset of 20 TB per day. By 2019, the dataset reached 7 PB, which contained the real-time battery information from more than 6,700 types of vehicles. This scale is approximately 1,000,000 times larger than the MIT-Stanford open-access battery aging dataset[19] (7.8 GB, tested on A123 type LiFePO$_4$ [LFP] batteries only). Excavating the potential utilization of (even a small proportion of) such industrial datasets could, without a doubt, provide an alternative approach to handle the data shortage issue for battery and energy storage aging assessment.

Second, the collected real-time data from field applications are usually not suitable for supervised learning applications due to the incompleteness of data, which is a primary challenge to address in this paper. To be specific, the batteries used in industrial systems, such as EVs, marine propulsion, and grid-tied energy storage, are rarely fully charged or discharged.[28,29] The actual capacity of these industry batteries can only be calibrated when they leave the production line as new batteries or during periodic maintenance periods.[30] In general, Li-ion batteries in industrial applications need real-time monitoring rather than frequent offline maintenance.[31] The maintenance frequency is often lower than once a year,[32,33] resulting in a lack of capacity measurement. In this case, a data-driven model for battery aging assessment is difficult to establish, even though some aging-related features, such as the incremental capacity (IC) information, can be readily extracted from current and voltage signals collected in real time.[34,35] This difficulty is by and large caused by the inherent nature of the general supervised training algorithms—a training target (battery capacity) is required to map input features. To the best of our knowledge, how to handle the field battery capacity data shortage in supervised learning is still an open question and no solutions have yet been reported.

In this paper, we introduce the potential of combining industrial datasets with the accelerated aging test to recover high-quality battery aging datasets, through a migration-based machine learning approach. Thereby the completeness of a field dataset can be greatly enhanced by appending the recovered battery capacity. To test our method in different scenarios, a comprehensive dataset consisting of 8,947 aging cycles for three different battery types using 15 load profiles is utilized. Our solution can subsequently recover the whole battery capacity trajectory with a maximum error of 0.86%, while saving up to 90% of the experimental time compared with the mainstream solution of carrying out cyclic aging experiments in the lab directly. The recovered dataset could be further used for training other data-based battery aging models, with a negligible loss of accuracy (<0.5%) in comparison with using the aging dataset collected from laboratory experiments. In addition, our recovered dataset could also be utilized in data-based approaches for the broader health-aware applications, such as failure predictions, lifespan evaluations, fast charging optimization, and battery "second-life" utilization, supporting more complex working conditions with better generalization performance. This paper not only provides an alternative solution to the data shortage issue in aging assessment, but also illustrates the potential of fusing easier-to-implement accelerated lab experiments with industrial data to generate large-volume high-quality datasets.

RESULTS

Recovering unmeasured field capacity data

We first consider the scenario of refining the battery aging datasets by regenerating the entire capacity degradation trajectories. In this scenario, the battery's voltage and current can be collected in real time. However, the capacity of each operating cycle is not available unless calibrated. Here, we start with the case that the calibrations are conducted, respectively, at 0%, 50%, and 100% of the battery lifespan. The data of these points could be obtained, respectively, from the factory calibration of new batteries, battery maintenance during use, and battery performance test before second-life applications. Then, our target here is to recover hundreds of uncollected capacity values from only three groups of calibration data, together with the easily collected voltage and current information.
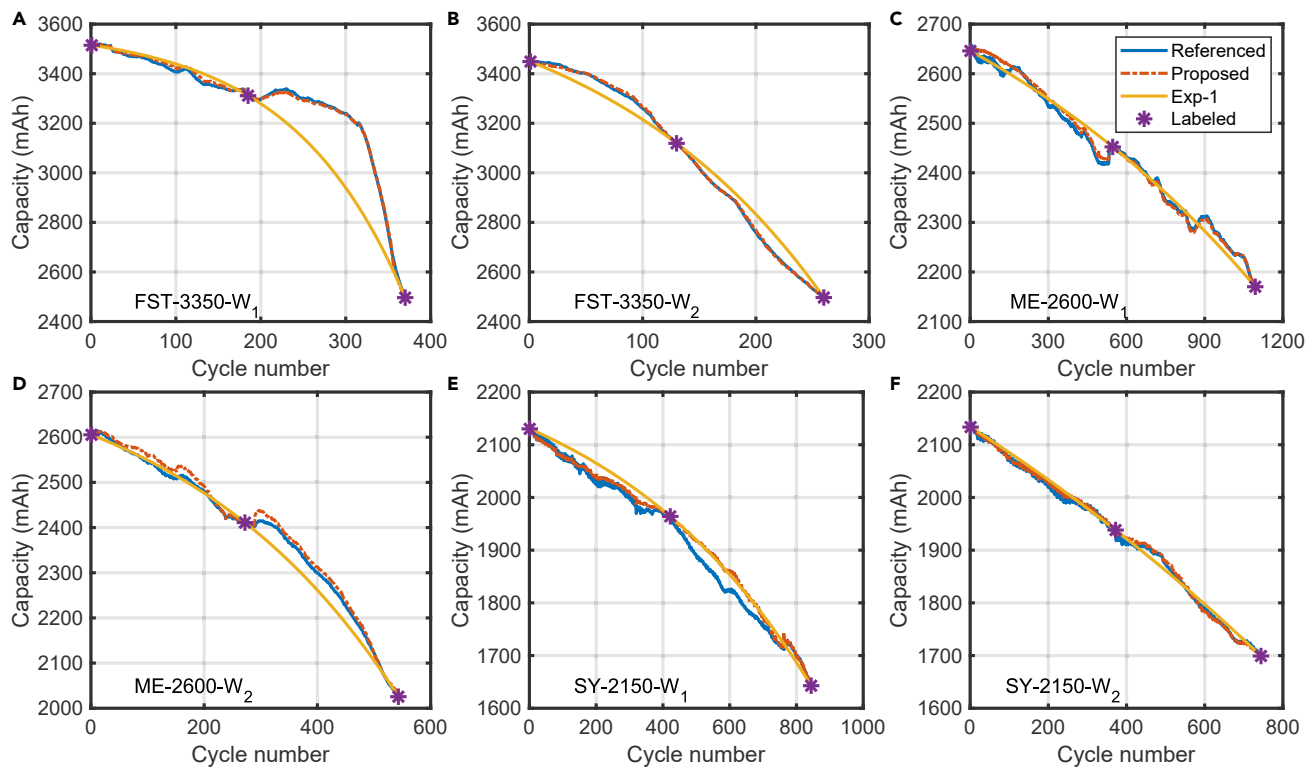
Figure 1. Capacity recovering results

(A) Results of the FST-3350-W$_1$ dataset; (B) results of the FST3350-W$_2$ dataset; (C) results of the ME-2600-W$_1$ dataset; (D) results of the ME-2600-W$_2$ dataset; (E) results of the SY-2150-W$_1$ dataset; (F) results of the SY-2150-W$_2$ dataset. In all six sub-figures, the blue curves represent the referenced battery capacity, three groups of labeled data are selected from this curve and marked with the purple stars. They correspond to the first, last, and 50% point of the aging trajectory. The red curves represent the recovered capacity using these three labels, and the yellow curve is the fitting results of the empirical exponential curve.

To verify our designed migration-based capacity recovering method (as detailed in the experimental procedures), we tested three commercial battery types (FST-3350, ME-2600, and SY-2150) and, for each battery type, two different aging methods (marked by W$_1$ and W$_2$) are conducted. The full details can be found in Section S1, Table S1, and Figure S1 of the supplemental information. It is worth noting that accurate laboratory-collected data have to be adopted in this work because reliable "referenced values" are required to evaluate the accuracy of our data recovery algorithm. Comparisons between the referenced and recovered capacity trajectories of the six scenarios are shown in Figures 1A–1F, respectively. Considering that the training results of a machine learning method may vary with the initialization of the neural network,[36] in addition to the results shown in Figure 1, we run our algorithm for another ten iterations to evaluate the consistency of the method. The root-mean-square errors (RMSEs) of all tests are summarized in Figure 2, while the full results can be found in Section S2, Table S2, and Figures S2–S11 of the supplemental information. To further illustrate the superiority of our method, a curve-fitting result of the capacity to the cycle number is also provided. With only three groups of labeled data, a popular and widely utilizsed empirical model with an exponential form $f(k) = a \cdot \exp(-b \cdot k) + c$[22] is selected as the fitting function, and the corresponding results are denoted as "Exp-1". Here a, b, and c are the fitted parameters, and k is the cycle number.

As illustrated in Figure 1, our proposed approach is not only able to effectively recover the general trend of the aging curves but also to capture the local fluctuations within the capacity trajectories. Even in the cases where the aging trends present significant differences for various battery types and cyclic operations, the proposed method limits the relevant RMSE to within 0.86%. Such a state-of-the-art accuracy[37,38] implies that our approach can recover the unmeasured capacity values accurately for general industrial datasets that store the partial voltage and current measurements, as long as only three regular capacity measuring maintenance operations are carried out during the entire battery lifespan. This in return will enrich the field battery aging datasets and guarantee their usability for battery aging assessment. It is worth noting that, when combining the existed industrial datasets with the accelerated aging tests, the overall experimental time of our approach is dominated by that of the accelerated aging tests. As illustrated in Table S1 of the supplemental information, the time required by the accelerated aging test is significantly lower than that of the normal-speed aging case (the time required by the test marked with "ME-2600-A" is only ~3% of that marked with "ME-2600-W$_1$"). The detailed proportion could vary with the applied accelerated and normal-speed aging conditions, but for the general normal-speed battery aging profiles with a relatively low current rate (e.g., <0.5C), our proposed method could be capable of saving up to 90% of the experimental time.
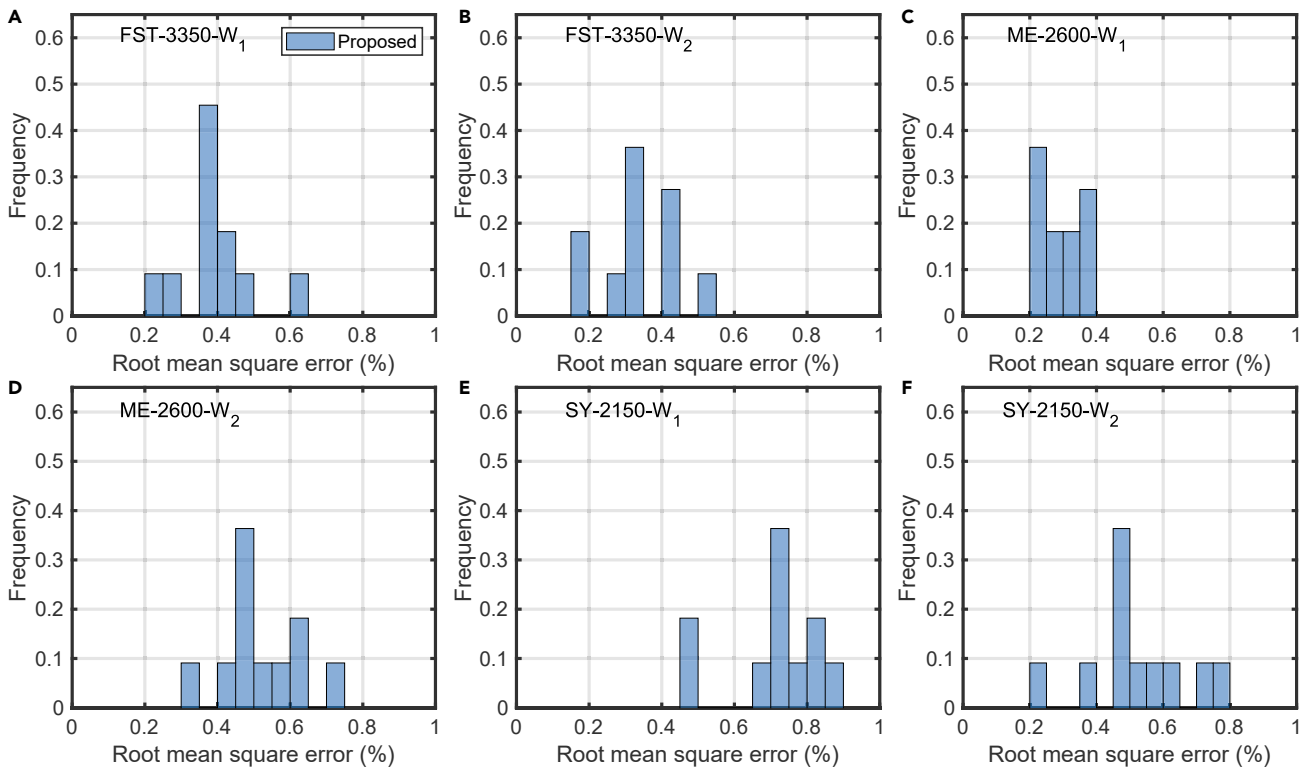
Figure 2. The distribution of the RMSE over 11 trials
(A) Results of the FST-3350-W$_1$ dataset; (B) results of the FST3350-W$_1$ dataset; (C) results of the ME-2600-W$_1$ dataset; (D) results of the ME-2600-W$_2$ dataset; (E) results of the SY-2150-W$_1$ dataset; (F) results of the SY-2150-W$_2$ dataset. In all six sub-figures, the x axis is the root-mean-square error of the methods, and the y axis is the frequency that the error locates in the specific range.

On the contrary, the performance of the curve-fitting approach shown as Exp-1 is rather inconsistent with the dataset. In some cases, such as SY-2150-W$_2$, Exp-1 provides regeneration results that closely match the referenced aging trajectories. However, for other cases, such as FST-3350-W$_1$, its deviations to the referenced values are much larger. This inconsistency is mainly caused by the inherent limitations of empirical fittings: the fitting is often only optimized for the given training data, and its generalization to unseen data is often poor. On the other hand, the fitting results are also determined by the structure of the selected fitting function and the choice of the input variables.

Influence of the "reference position"
As batteries are often operated and maintained differently in various applications, the maintenance check points may not be evenly distributed along the aging trajectory (0%, 50%, 100%, etc.). In light of this, it is also important to evaluate the impact of different referenced positions on the accuracy of recovered capacity trajectories with our proposed method. Here, the FST-3350-W$_1$ dataset is used for illustration as its capacity degradation trajectory exhibits the highest non-linearity among all the six cases as shown in Figure 1. The testing results are shown in Figure 3, and the RMSE of the recovered capacity trajectories are given in Table 1. Further results of the other batteries can be found in Section S3, Table S3, and Figure S12 of the supplemental information, while Table S4 and Figure S13

list the results when the distribution of the maintenance check points are heavily unbalanced.

The accuracy of Exp-1, as expected, is low, and its RMSE reaches 4.83%. In addition, the curve-fitting results can vary significantly with the change of the reference position. In contrast, the RMSEs of our proposed approach are smaller than 0.5% in all cases when the maintenance checkpoints are distributed along the aging trajectory in a close-to-uniform way (RMSEs can be bounded within 1.5% for heavily unbalanced cases, even though it is not recommended to use our algorithm in these extreme cases), once again confirming that this approach is insensitive to the position of the referenced capacity points. As shown in Figure 3, the recovered trajectories using the proposed approach matches sufficiently close to the true values even if the testing conditions are significantly different. This feature is critically important, as in real-life applications the timing of battery maintenance can vary from user to user. The requirement of the reference capacity values or maintenance points to recover the whole capacity trajectory is flexible in our proposed method.

DISCUSSIONS

Summary
Data-based methods are powerful to model and evaluate the complex battery degradation processes, but their performance
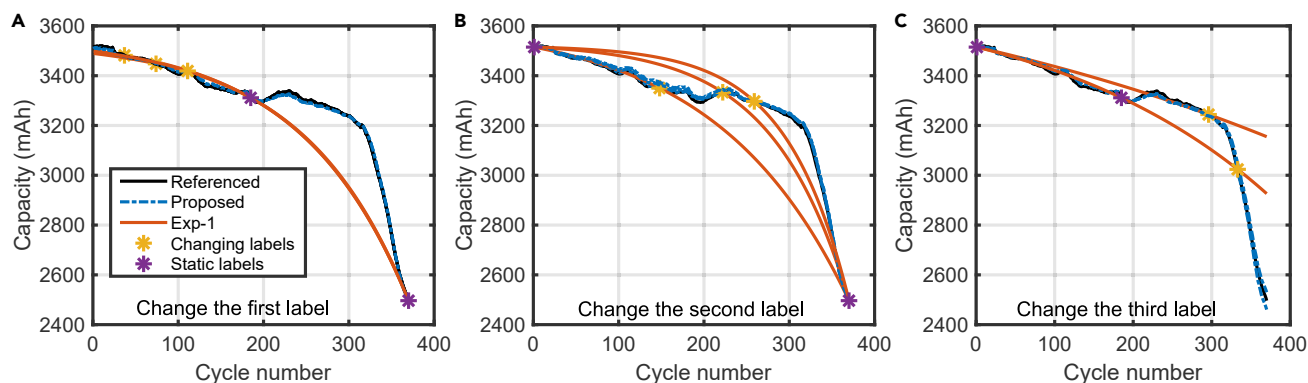
Figure 3. Influence of the "reference position" tested with FST-3350-W₁ dataset

(A) Results of moving the first label from the first cycle to the cycles corresponding to 10%, 20%, and 30% of the aging trajectory; (B) results of moving the second label from the cycle corresponding to 50% of the aging trajectory to 40%, 60%, and 70%; (C) results of moving the third label from the last cycle to the cycles corresponding to 80% and 90% of the aging trajectory.

is quite often limited by the scale of available battery aging datasets. Aiming at this typical but challenging issue, we have made an attempt to explore the potential of field datasets by proposing a rapid approach to effectively recover the unmeasured capacity information from industrial applications, for the first time. Using only three groups of training data, an accurate recovery result with 0.86% RMSE has been achieved for three commercial battery types that have undergone thousands of cycles, while the required experimental time can be significantly reduced by up to 90%. The highly accurate capacity values recovered from our approach could be used for training data-based battery aging models, with a negligible loss of accuracy (<0.5%) in comparison with using the aging dataset collected from laboratory experiments. (The corresponding comparative results are given in Section S4, Table S5, and Figure S14 of the supplemental information.) These results illustrate that, for general data-based battery health assessments, our recovered datasets can complement the ones acquired from the large-scale aging experiments carried out in the laboratory. By augmenting large-volume field datasets (commonly collected in a data center) with our proposed solution, the data shortage issue in the area of battery and energy storage health assessment can be significantly improved. It will directly support the advancement of the data-based battery health diagnostics and prognostics and, in turn, significantly enhance the safety, lifespan, and efficiency of industrial systems, such as EVs or energy storage power stations. From a broader point of view, this work highlights the promise of combining industrial calibration data and accelerated lab experiments to generate high-quality datasets with ultra-low experimental effort, and this strategy can also used in other industrial applications for generating useful large-volume datasets.

Research limitations and future works

Due to the lack of fully labeled battery aging datasets caused by complicated battery operations and long aging experimental time, the aim of this study is to recover high-quality battery aging datasets rather than predicting battery lifetime. This is important as it could significantly improve data shortage issues for assessment of battery and energy storage aging. Although developing data-driven strategies to carry out battery lifetime prognostic actually lies beyond the scope of this particular work, we also believe it is another key and challenging topic in the battery community. In this context, some related topics, such as the discovery of new health indicators and the establishment of advanced battery data-driven aging models, are important and considered as other valuable research directions to avoid extending the scope of the paper unnecessarily.

In addition, following the fact that battery charging in industrial applications is usually implemented in a relatively stable environment with a standard current rate, we assumed that the charging conditions of target battery for data recovery remain similar along its life. Although uncommon, if the battery of interest does violate this assumption, the proposed migration-based data recovering strategy could still be implemented, by developing a more comprehensive base model (the concept of "base model" is introduced in the experimental procedures) that considers the impact of different charging conditions. Given the safety requirement of real-life applications, we also assumed that battery systems should be trimmed before releasing to the market, maintained at least one time during their lifespan, and tested again before second-life usage, making it possible to collect data from at least three maintenance checkpoints. However, recovering the data with only two or even one checkpoint(s) is still an interesting future research topic.

Table 1. RMSE of the recovered capacity (in %) with different placement of the labels

Location	[10, 50, 100]	[20, 50, 100]	[30, 50, 100]	[0, 40, 100]	[0, 60, 100]	[0, 70, 100]	[0, 50, 80]	[0, 50, 90]
Proposed	0.21	0.22	0.21	0.20	0.32	0.42	0.34	0.32
Exponential	4.12	4.03	4.09	4.83	2.83	2.55	4.50	3.20

"Location" refers to the placement of the three labels along the battery's aging trajectory (in %).
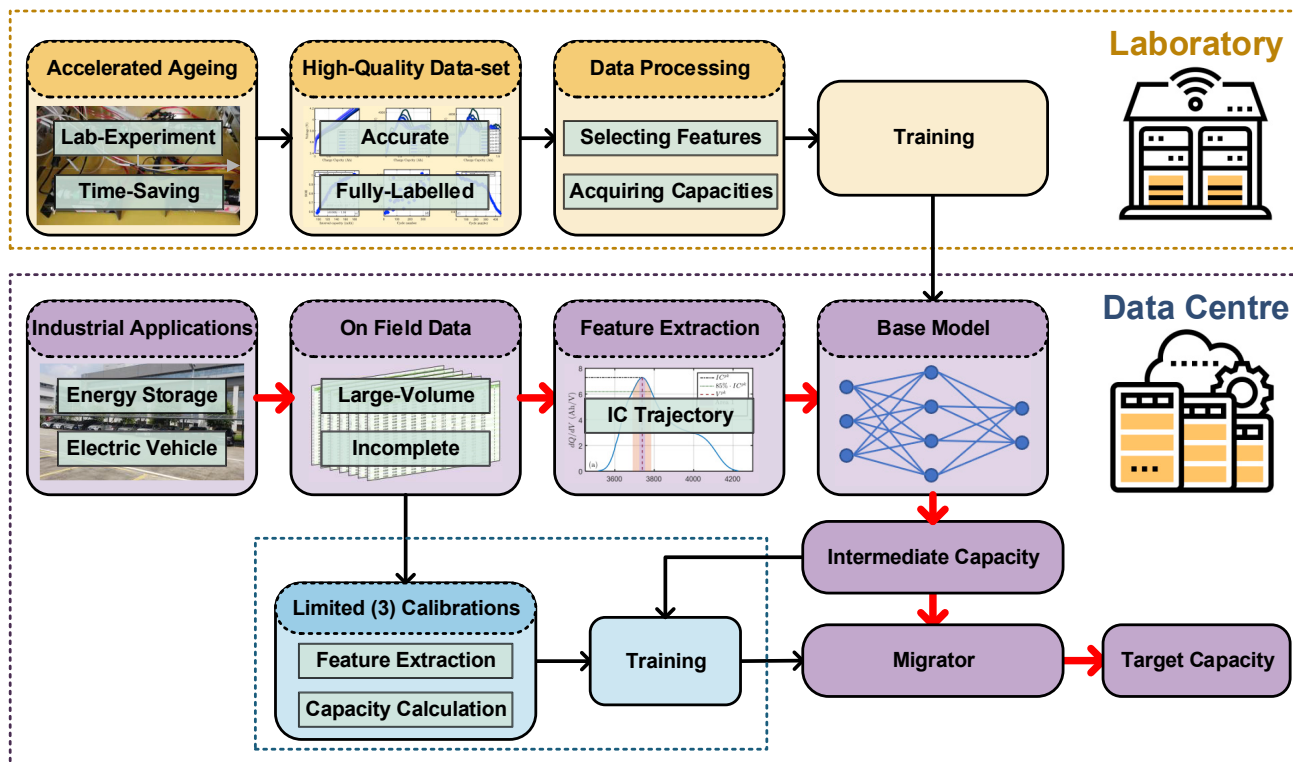
Figure 4. The diagram of the proposed data-regeneration method
The method consists of three main steps. First, aging-related features are extracted from the field data collected from industrial applications. Second, a base model is utilized to predict the intermediate battery capacity values with respect to these features. Finally, the intermediate capacity values obtained in the second step are migrated to generate the target capacity data. Here, the base model is trained using the data acquired from the lab experiments, and accelerated aging strategies could be used to reduce the experimental effort. The last migration step is used to compensate for the model mismatch between the base model and the target process of interest. The number of data samples required for training the migration rule is reduced to three even if the battery's lifespan exceeds hundreds or thousands of cycles.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact
Please contact the lead contact, K. Liu (kailong.liu@warwick.ac.uk, kliu02@qub.ac.uk), for information related to the data and code described in the following section.

Materials availability
This study did not generate new unique materials.

Data and code availability
All data and code related to this work would be found at https://github.com/xtangai/Ageing-Data-Set or https://doi.org/10.5281/zenodo.4867041 and upon request to the lead contact. Alternatively, the file could be downloaded from the supplemental information—Data S1.

Method overview
Given the challenges of directly building a model with a high degree of freedom using only three data points, an alternative "base + migration" strategy is designed here with the diagram shown in Figure 4. Specifically, a base model that maps the selected features to the battery capacity is first built offline from the accelerated aging experiments in the laboratory (our designed accelerated aging experiments are detailed in Section S1 of the supplemental information). Then, a migration model is established, such that the intermediate outputs from base model could be transferred to the desired target. Such a migration model can be trained using only three capacity measurements obtained from related maintenance checkpoints. Finally, the unmeasured capacities in between the checkup instances can be recovered by using our base + migration

method. The following discussions start with the feature extraction and establishment of the base model.

Feature extraction and base model establishment
In the general model training phase, it is preferred to choose input features with low noise sensitivity and high correlations to battery capacity as the inputs of the base model. Motivated by the fact that the IC-based features have several key merits, such as containing electrode thermodynamic information,[19,35] and having the potential to exhibit close-to-linear relationship with battery aging,[34] we develop an IC-based neural network model as the base model. For a constant current charging process, the IC is defined as:

$$IC_k = \frac{dQ}{dV}\bigg|_k \approx \frac{Q_k - Q_{k-\Delta k}}{V_k - V_{k-\Delta k}}, \qquad \text{(Equation 1)}$$

where Q refers to the capacity charged into the battery in a specific charging process, V stands for the terminal voltage, Δk is the interval of finite difference, and k represents the time index within the charging cycle. The raw IC value calculated from Equation 1 usually contains noise. Therefore, a Gaussian moving average filter is then employed to smooth the IC curves. Detailed specifications for the filters can be found in Section S5 of the supplemental information.

From the smooth IC curve characteristic, four features are selected as the inputs of the neural network. To facilitate the description, these features are illustrated in Figure 5A. Here the first feature is the peak value of the IC curve, also known as the IC peak, IC^{pk}. The second feature is the voltage corresponding to the IC peak, also known as the peak voltage, V^{pk}. The third feature is the capacity contained in "Area 1" with a boundary of $V^{pk} \pm 15$ mV, as shown in
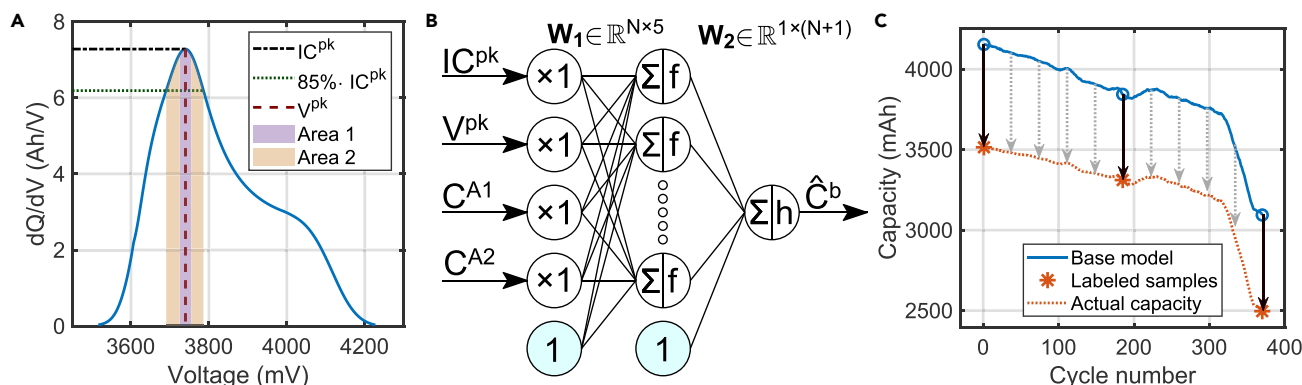
Figure 5. Methodology
(A) Illustration of the selected features, including the peak of the IC trajectory IC^{pk}, the voltage corresponding to the peak V^{pk}, the capacity contained in the Areas 1 and 2, C^{A1} and C^{A2}. (B) The three-layer feedforward neural network for establishing the base model. (C) Illustration of the capacity regeneration procedure. When using the base model to estimate capacity, its related results (blue curve) would be biased. A strategy is required, which can shift the biased output (solid blue curve) to the actual capacity (dotted red curve) by using only three data points (red stars).

Figure 5A. In other words, the capacity change corresponding to the period when the battery's terminal voltage rises from $(V^{pk} - 15)$ mV to $(V^{pk} + 15)$ mV is defined as the third feature, which is termed as the "interval capacity-A1," C^{A1}. Accordingly, the fourth feature is the capacity contained in "Area 2" of Figure 5A, whose boundary is defined as the intersection point between the IC trajectory and a horizontal line $y = 85\% \times IC^{pk}$. For simplicity, the last feature is termed as "interval capacity-A2," C^{A2}.

Next, these features are mapped to the referenced battery capacity through a three-layer neural network with N hidden neurons, as illustrated in Figure 5B. Under this framework, the overall output is expressed by $F_{NET}^{b}(\cdot)$ as:

$$y = \widehat{C}^{b} = F_{NET}^{b}(\boldsymbol{x}) = h\left(\sum_{i=1}^{N} W_2^{(1,i)} \cdot f\left(\sum_{k=1}^{4} W_1^{(i,k)} \cdot x_i + W_1^{(i,5)}\right) + W_2^{(1,N+1)}\right), \quad \text{(Equation 2)}$$

where $\boldsymbol{x} = [IC^{pk}; V^{pk}; C^{A1}; C^{A2}]$.

Without loss of generality, the activation function of the hidden layer is selected as the positive linear function ($poslin(\cdot)$), while a pure linear function ($purlin(\cdot)$) is chosen for the output layer as:

$$f(x) = poslin(x) = \max\{0, x\}, \quad \text{(Equation 3)}$$

$$h(x) = purlin(x) = x. \quad \text{(Equation 4)}$$

For the purpose of simplifying the algorithm implementation, the MATLAB `nftool` toolbox is used for the network training in this study. Please refer to Section S5.2 of the supplemental information for details. It is also worth mentioning that a traditional feedforward neural network is sufficient to describe the relationship between the selected IC-based features and the battery's capacities in this work. But readers may also use more advanced recurrent neural network[39] or convolutional neural network[40] for enhanced performances.

Remark 1. The detailed selection of the features will not affect the logic of our base + migration framework, and multiple aging-related parameters could be used. However, noting that a battery is rarely fully charged or discharged in real-life industrial applications, the selected features should be extracted without using the data segments of high or low state-of-charge (SoC). In addition, as the battery SoC is inconvenient to be estimated when considering the joint influence of aging and temperature,[41,42] the features that are strongly coupled with battery SoC should be avoided.

Interpolation-based migration

Since the base model is built on the accelerated aging dataset performed in the laboratory, when using it to predict battery capacities during field operations, where the degradation rate is lower, an error caused by the model mismatch can exist, as illustrated in Figure 5C. Therefore, it is necessary to define a migration rule for the base model to efficiently transfer its outputs to the actual battery capacity trajectories for the field operations.

As only limited capacity reference points from the field operation are available, an electrochemical-based migration rule with tens of parameters is difficult to establish. To handle this issue, an interpolation-based corrector is utilized as the easy-to-implement alternative. Here, we assume N referenced capacity points are available and denote these capacities as $C_{[l_{1:N}]}^{ref}$, where $l_1, l_2, \cdots, l_N$ are the corresponding cycle when the capacity is calibrated. Accordingly, the intermediate output points predicted by the base model at $l_{1:N}^{th}$ cycles can be denoted as $\widehat{C}_{[l_{1:N}]}^{b}$. The interpolation-based migration can then be described by:

$$L(x) = p_1(x) \cdot C_{l_1}^{ref} + p_2(x) \cdot C_{l_2}^{ref} + \cdots + p_N(x) \cdot C_{l_N}^{ref}, \quad \text{(Equation 5)}$$

where

$$p_k(x) = \frac{\left(x - \widehat{C}_{l_1}^{b}\right) \cdots \left(x - \widehat{C}_{l_{k-1}}^{b}\right)\left(x - \widehat{C}_{l_{k+1}}^{b}\right) \cdots \left(x - \widehat{C}_{l_N}^{b}\right)}{\left(\widehat{C}_{l_k}^{b} - \widehat{C}_{l_1}^{b}\right) \cdots \left(\widehat{C}_{l_k}^{b} - \widehat{C}_{l_{k-1}}^{b}\right)\left(\widehat{C}_{l_k}^{b} - \widehat{C}_{l_{k+1}}^{b}\right) \cdots \left(\widehat{C}_{l_k}^{b} - \widehat{C}_{l_N}^{b}\right)} \quad \text{(Equation 6)}$$

holds for $k \in [1, N]$.

With the established migration rule $L(\cdot)$ and the features $\boldsymbol{x}$ extracted from the battery's charging voltage curves of a certain cycle, we can use Equation 7 to estimate the referenced battery's capacity (denoted as $\widehat{C}^{m}$) for this specific cycle:

$$\widehat{C}^{m} = L\left(\widehat{C}^{b}\right) = L\left(F_{NET}^{b}(\boldsymbol{x})\right). \quad \text{(Equation 7)}$$

Remark 2. With the joint effects of the piece-wise linear migration rule (Equation 5), the positive linear activation in (Equation 3), and the pure linear function in (Equation 4), higher capacity regeneration accuracy could be achieved if the selected features in the above-mentioned feature extraction and base model establishment section have close-to-linear relationships with the capacity.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100302.

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

X.T. and F.G. conceived the study and carried out the experiments. X.T. and K.L. developed the machine learning solution, implemented the programming, analyzed the experimental results, and drafted the manuscript. K.L., W.D.W., and E.K. discussed the technical details, analyzed the data, and revised and polished the manuscript. All authors commented on the manuscript.

DECLARATION OF INTERESTS

The authors are preparing a CN patent and a US patent related to this work.

REFERENCES

1. Stephan, A.K. (2019). The age of li-ion batteries. Joule *3*, 2583–2584. https://doi.org/10.1016/j.joule.2019.11.004.

2. Boles, S.T., and Tahmasebi, M.H. (2020). Are foils the future of anodes? Joule *4*, 1342–1346. https://doi.org/10.1016/j.joule.2020.05.009.

3. Berecibar, M. (2019). Machine-learning techniques used to accurately predict battery life. Nature *568*, 325–326. https://doi.org/10.1038/d41586-019-01138-1.

4. Hu, X., Xu, L., Lin, X., and Pecht, M. (2020). Battery lifetime prognostics. Joule *4*, 310–346. https://doi.org/10.1016/j.joule.2019.11.018.

5. Ng, M.F., Zhao, J., Yan, Q., Conduit, G.J., and Seh, Z.W. (2020). Predicting the state of charge and health of batteries using data-driven machine learning. Nat. Mach. Intell. *2*, 161–170. https://doi.org/10.1038/s42256-020-0156-7.

6. Wang, Y., Tian, J., Sun, Z., Wang, L., Xu, R., Li, M., and Chen, Z. (2020). A comprehensive review of battery modeling and state estimation approaches for advanced battery management systems. Renew. Sustain. Energy Rev. *131*, 110015. https://doi.org/10.1016/j.rser.2020.110015.

7. Deng, J., Bae, C., Denlinger, A., and Miller, T. (2020). Electric vehicles batteries: requirements and challenges. Joule *4*, 511–515. https://doi.org/10.1016/j.joule.2020.01.013.

8. Attia, P.M., Grover, A., Jin, N., Severson, K.A., Markov, T.M., Liao, Y.H., Chen, M.H., Cheong, B., Perkins, N., Yang, Z., et al. (2020). Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature *578*, 397–402. https://doi.org/10.1038/s41586-020-1994-5.

9. Bhowmik, A., and Vegge, T. (2020). AI fast track to battery fast charge. Joule *4*, 717–719. https://doi.org/10.1016/j.joule.2020.03.016.

10. Liu, K., Zou, C., Li, K., and Wik, T. (2018). Charging pattern optimization for lithium-ion batteries with an electrothermal-aging model. IEEE Trans. Ind. Inform. *14*, 5463–5474. https://doi.org/10.1109/tii.2018.2866493.

11. Li, W., Zhu, J., Xia, Y., Gorji, M.B., and Wierzbicki, T. (2019). Data-driven safety envelope of lithium-ion batteries for electric vehicles. Joule *3*, 2703–2715. https://doi.org/10.1016/j.joule.2019.07.026.

12. Finegan, D.P., Zhu, J., Feng, X., Keyser, M., Ulmefors, M., Li, W., Bazant, M.Z., and Cooper, S.J. (2020). The application of data-driven methods and physics-based learning for improving battery safety. Joule. https://doi.org/10.1016/j.joule.2020.11.018.

13. Baars, J., Domenech, T., Bleischwitz, R., Melin, H.E., and Heidrich, O. (2020). Circular economy strategies for electric vehicle batteries reduce reliance on raw materials. Nat. Sustain. *4*, 71–79. https://doi.org/10.1038/s41893-020-00607-0.

14. Ciez, R.E., and Whitacre, J.F. (2019). Examining different recycling processes for lithium-ion batteries. Nat. Sustain. *2*, 148–156. https://doi.org/10.1038/s41893-019-0222-5.

15. Chen, M., Ma, X., Chen, B., Arsenault, R., Karlson, P., Simon, N., and Wang, Y. (2019). Recycling end-of-life electric vehicle lithium-ion batteries. Joule *3*, 2622–2646. https://doi.org/10.1016/j.joule.2019.09.014.

16. Harper, G., Sommerville, R., Kendrick, E., Driscoll, L., Slater, P., Stolkin, R., Walton, A., Christensen, P., Heidrich, O., Lambert, S., et al. (2019). Recycling lithium-ion batteries from electric vehicles. Nature *575*, 75–86. https://doi.org/10.1038/s41586-019-1682-5.

17. Xu, P., Dai, Q., Gao, H., Liu, H., Zhang, M., Li, M., Chen, Y., An, K., Meng, Y.S., Liu, P., et al. (2020). Efficient direct recycling of lithium-ion battery cathodes by targeted healing. Joule *4*, 2609–2626. https://doi.org/10.1016/j.joule.2020.10.008.

18. Aykol, M., Herring, P., and Anapolsky, A. (2020). Machine learning for continuous innovation in battery technologies. Nat. Rev. Mater. https://doi.org/10.1038/s41578-020-0216-y.

19. Severson, K.A., Attia, P.M., Jin, N., Perkins, N., Jiang, B., Yang, Z., Chen, M.H., Aykol, M., Herring, P.K., Fraggedakis, D., et al. (2019). Data-driven prediction of battery cycle life before capacity degradation. Nat. Energy *4*, 383–391. https://doi.org/10.1038/s41560-019-0356-8.

20. Zhang, Y., Tang, Q., Zhang, Y., Wang, J., Stimming, U., and Lee, A.A. (2020). Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. Nat. Commun. *11*. https://doi.org/10.1038/s41467-020-15235-7.

21. He, W., Williard, N., Osterman, M., and Pecht, M. (2011). Prognostics of lithium-ion batteries based on Dempster–Shafer theory and the Bayesian Monte Carlo method. J. Power Sources *196*, 10314–10321. https://doi.org/10.1016/j.jpowsour.2011.08.040.

22. Tang, X., Liu, K., Wang, X., Gao, F., Macro, J., and Widanage, W.D. (2020). Model migration neural network for predicting battery aging trajectories. IEEE Trans. Transp. Electr. *6*, 363–374. https://doi.org/10.1109/tte.2020.2979547.

23. Saha, B., and Goebel, K.. Battery data set. 2007. NASA Ames Research Center, Moffett Field, CA. https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/.

24. Li, B., Kisacikoglu, M.C., Liu, C., Singh, N., and Erol-Kantarci, M. (2017). Big data analytics for electric vehicle integration in green smart cities. IEEE Commun. Magazine *55*, 19–25. https://doi.org/10.1109/mcom.2017.1700133.

25. Xiong, R., Sun, W., Yu, Q., and Sun, F. (2020). Research progress, challenges and prospects of fault diagnosis on battery system of electric vehicles. Appl. Energy *279*, 115855. https://doi.org/10.1016/j.apenergy.2020.115855.

26. Hong, J., Wang, Z., and Yao, Y. (2019). Fault prognosis of battery system based on accurate voltage abnormity prognosis using long short-term memory neural networks. Appl. Energy *251*, 113381. https://doi.org/10.1016/j.apenergy.2019.113381.

27. Zhao, Y., Liu, P., Wang, Z., Zhang, L., and Hong, J. (2017). Fault and defect diagnosis of battery for electric vehicles based on big data analysis methods. Appl. Energy *207*, 354–362. https://doi.org/10.1016/j.apenergy.2017.05.139.

28. Lu, L., Han, X., Li, J., Hua, J., and Ouyang, M. (2013). A review on the key issues for lithium-ion battery management in electric vehicles. J. Power Sources *226*, 272–288. https://doi.org/10.1016/j.jpowsour.2012.10.060.

29. Balsamo, F., Capasso, C., Lauria, D., and Veneri, O. (2020). Optimal design and energy management of hybrid storage systems for marine

propulsion applications. Appl. Energy *278*, 115629. https://doi.org/10.1016/j.apenergy.2020.115629.

30. IEEE guide for design, operation, and maintenance of battery storage systems, both stationary and mobile, and applications integrated with electric power systems. 2019. https://doi.org/10.1109/IEEESTD.2019.8930450.

31. IEEE guide for the characterization and evaluation of lithium-based batteries in stationary applications. 2018. https://doi.org/10.1109/IEEESTD.2018.8262521.

32. Gomez-Parra, M., Munoz-Condes, P., Sancho, C., Gonzalez-Fernandez, F., San Andres, M.A.G., Carpio, J., et al. Implementation of a new predictive maintenance methodology for batteries. Application to railway operations. In: 2009 IEEE Vehicle Power and Propulsion Conference. 2009, p. 1236–1243. doi:10.1109/VPPC.2009.5289709.

33. Mathew, M., Kong, Q., McGrory, J., and Fowler, M. (2017). Simulation of lithium ion battery replacement in a battery pack for application in electric vehicles. J. Power Sources *349*, 94–104. https://doi.org/10.1016/j.jpowsour.2017.03.010.

34. Tang, X., Zou, C., Yao, K., Chen, G., Liu, B., He, Z., and Gao, F. (2018). A fast estimation algorithm for lithium-ion battery state of health. J. Power Sources *396*, 453–458. https://doi.org/10.1016/j.jpowsour.2018.06.036.

35. Naha, A., Han, S., Agarwal, S., Guha, A., Khandelwal, A., Tagade, P., Hariharan, K.S., Kolake, S.M., Yoon, J., and Oh, B. (2020). An incremental voltage difference based technique for online state of health estimation of li-ion batteries. Sci. Rep. *10*. https://doi.org/10.1038/s41598-020-66424-9.

36. Humbird, K.D., Peterson, J.L., and Mcclarren, R.G. (2019). Deep neural network initialization with decision trees. IEEE Trans. Neural Netwk. Learn. Syst. *30*, 1286–1295. https://doi.org/10.1109/tnnls.2018.2869694.

37. Tian, H., Qin, P., Li, K., and Zhao, Z. (2020). A review of the state of health for lithium-ion batteries: research status and suggestions. J. Clean. Prod. *261*, 120813. https://doi.org/10.1016/j.jclepro.2020.120813.

38. Lipu, M.H., Hannan, M., Hussain, A., Hoque, M., Ker, P.J., Saad, M., and Ayob, A. (2018). A review of state of health and remaining useful life estimation methods for lithium-ion battery in electric vehicles: challenges and recommendations. J. Clean. Prod. *205*, 115–133. https://doi.org/10.1016/j.jclepro.2018.09.065.

39. Liu, K., Shang, Y., Ouyang, Q., and Widanage, W.D. (2021). 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. https://doi.org/10.1109/TIE.2020.2973876.

40. Li, Y., Li, K., Liu, X., Wang, Y., and Zhang, L. (2021). Lithium-ion battery capacity estimation—a pruned convolutional neural network approach assisted with transfer learning. Appl. Energy *285*, 116410. https://doi.org/10.1016/j.apenergy.2020.116410.

41. Tang, X., Wang, Y., Zou, C., Yao, K., Xia, Y., and Gao, F. (2019). A novel framework for lithium-ion battery modeling considering uncertainties of temperature and aging. Energy Convers. Manag. *180*, 162–170. https://doi.org/10.1016/j.enconman.2018.10.082.

42. Diao, W., Saxena, S., and Pecht, M. (2019). Accelerated cycle life testing and capacity degradation modeling of LiCoO2-graphite cells. J. Power Sources *435*, 226830. https://doi.org/10.1016/j.jpowsour.2019.226830.