Energy and AI 18 (2024) 100451

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.sciencedirect.com/journal/energy-and-ai

Machine learning for battery quality classification and lifetime prediction using formation data

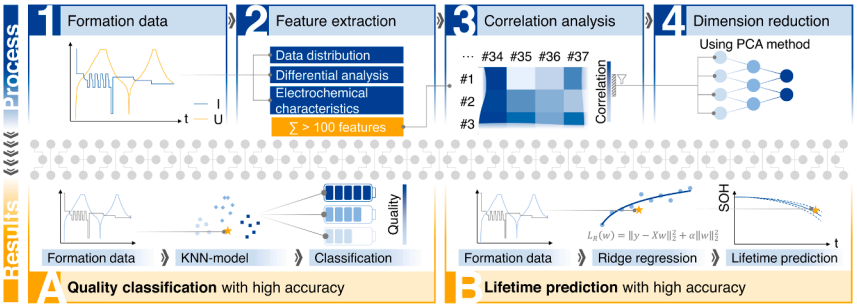
Jiayu Zou [a,b,e], Yingbo Gao [c], Moritz H. Frieges [d], Martin F. Börner [a,b,e], Achim Kampker [d], Weihan Li [a,b,e,*]

[a] Chair for Electrochemical Energy Conversion and Storage Systems, Institute for Power Electronics and Electrical Drives (ISEA), RWTH Aachen University, Campus-Boulevard 89, 52074 Aachen, Germany
[b] Center for Ageing, Reliability and Lifetime Prediction of Electrochemical and Power Electronic Systems (CARL), RWTH Aachen University, Campus-Boulevard 89, Aachen, Germany
[c] Department of Computer Science, RWTH Aachen University, 52056 Aachen, Germany
[d] Chair of Production Engineering of E-Mobility Components (PEM), RWTH Aachen University, Bohr 12, 52072 Aachen, Germany
[e] Juelich Aachen Research Alliance, JARA-Energy, Germany

HIGHLIGHTS

- Battery quality classification and lifetime prediction only using formation data.
- Data-driven framework with machine learning for quality control in battery production.
- Over 100 features extracted from formation data for enhanced ageing analysis.
- Achieved 89.74% accuracy in classification and 5.45% error in lifetime prediction.
- Potential to reduce costly end-of-line testing in battery production processes.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Battery
Formation
Quality classification
Life prediction
Machine learning

ABSTRACT

Accurate classification of battery quality and prediction of battery lifetime before leaving the factory would bring economic and safety benefits. Here, we propose a data-driven approach with machine learning to classify the battery quality and predict the battery lifetime before usage only using formation data. We extract three classes of features from the raw formation data, considering the statistical aspects, differential analysis, and electrochemical characteristics. The correlation between over 100 extracted features and the battery lifetime is analysed based on the ageing mechanisms. Machine learning models are developed to classify battery quality and predict battery lifetime by features with a high correlation with battery ageing. The validation results show that the quality classification model achieved accuracies of 89.74% and 89.47% for the batteries aged at 25°C and 45°C, respectively. Moreover, the lifetime prediction model is able to predict the battery end-of-life with mean percentage errors of 6.50% and 5.45% for the batteries aged at 25°C and 45°C, respectively. This work highlights the potential of battery formation data from production lines in quality classification and lifetime prediction.

* Corresponding author.
E-mail address: weihan.li@isea.rwth-aachen.de (W. Li).

https://doi.org/10.1016/j.egyai.2024.100451
Received 15 July 2024; Received in revised form 9 October 2024; Accepted 15 November 2024
Available online 16 November 2024
2666-5468/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Introduction

In recent years, gasoline and diesel vehicles have been recognised as major contributors to air pollution and global warming, prompting a rise in the popularity of electric vehicles. Lithium-ion batteries (LIBs) are favoured for electric vehicles due to their affordability, high power and energy density, and long lifespan [1]. However, concerns about battery quality, particularly regarding energy and power density, lifespan, and safety, remain prevalent. As a result, quality control is a crucial aspect of battery production. Monitoring parameters related to battery quality helps prevent defective or substandard battery products from leaving the production lines and provides essential information for battery scientists and engineers to make necessary adjustments for product optimisation. Currently, state-of-the-art methods, e.g., capacity test and resistance measurement measurements, are widely used during the end-of-line test in battery production [2]. These tests provide insights into the electrical behaviour of batteries but require additional testing equipment and time, and they often do not yield comprehensive information about battery ageing and lifespan. To reduce costs and time associated with end-of-line tests while increasing the accuracy of quality classification and lifespan prediction, there is a growing focus on analysing and utilising data from production processes like battery formation.

There are typically three fundamental processes in battery manufacturing: electrode production, cell production, and cell conditioning. Cell conditioning begins with the formation process, which directly affects the quality of solid electrolyte interphase (SEI) and, consequently, the lifetime and the safety of LIBs [3,4]. During formation, the battery cell is charged for the first time, and several electrical cycles are performed to form the SEI. After formation, the battery cells are stored for several days or weeks before undergoing end-of-line testing and cyclic testing for capacity grading. Zhang et al. [5] identified two stages of SEI formation: the first happens at graphite electrode voltages above 0.25 V, and the second stage occurs between 0.25 V and 0.04 V. A well-formed SEI facilitates smooth electrochemical reactions by allowing lithium ions (Li^+) to pass through while blocking electrons [6]. This selective permeability prevents unwanted reactions between the negative electrode and the electrolyte [7]. Additionally, Zhang et al. noted that the SEI can block the co-intercalation of solvents. An ideal SEI is passivating, electrochemically inert, thin, electronically resistive and Li^+ conductive [8]. SEI growth directly influences battery lifetime and is one of the dominating ageing mechanisms for LIBs in applications, which can be modelled based on kinetic limitations [9,10] and solvent diffusion [11,12]. SEI models can also predict the end of life (EOL) of LIBs [13]. Therefore, the data acquired during the formation process potentially contains crucial information about the quality of LIBs. This data can be used to extract various features to predict the lifetime of LIBs, thereby reducing or eliminating the need for extensive end-of-line testing and preventing further processing of cells already identified as faulty.

To successfully correlate the data collected during formation to battery quality and lifetime performance, it is crucial to understand the ageing processes of battery cells. In general, battery degradation modes can be categorised into three types: loss of lithium inventory (*LLI*), loss of active material in the positive electrode (LAM_{PE}), and loss of active material in the negative electrode (LAM_{NE}) [14]. Many researchers have proposed mechanisms- or physics-based models to predict the lifetime of LIBs. For instance, Safari et al. [15] discussed the potential for lifetime prediction using a mechanical fatigue model, while Yang et al. [16] developed an ageing model that considers both SEI growth and lithium plating. Lui et al. introduced a physics-based approach for remaining useful life (RUL) prediction based on *LLI*, LAM_{PE}, and LAM_{PE} [17]. With the availability of large battery datasets, advancements in computational power, and the mechanism-free characteristics of machine learning, data-driven approaches are increasingly being used for battery lifetime prediction [18]. For example, Xue et al. applied support vector regression (SVR) to predict the RUL of LIBs [19]. Pang et al. combined incremental capacity analysis and Gaussian process regression (GPR) for RUL prediction [20]. Richardson et al. employed GPR to predict the battery lifetime in fast-charging scenarios, demonstrating a reduction in testing time [21]. Other neural networks, such as recurrent neural networks (RNN) [22,23], long short-term memory neural networks (LSTM) [24], convolutional neural networks (CNN) [25,26], sequence-to-sequence models [27] and multi-task learning models [28] have also been widely used for ageing prediction.

To further reduce the time needed for ageing evaluation, some studies aim to predict the battery lifetime using minimal data, such as the first 100 cycles [29], a quarter of the ageing curves [30], or even just one cycle [31]. However, quality control and lifetime prediction using only formation data remain challenging because formation data covers only the initial charge-discharge cycles in battery manufacturing. Weng et al. introduced a method for battery lifetime prediction with demonstrated capabilities using formation data [32]. However, their model primarily relies on resistance signals at low states of charge (SOC), which do not directly originate from the formation process data. Utilising only formation data results in a predictive error of around 15%.

The significance of formation data lies in its potential to serve as an early prognosticator of battery health and lifespan. Unlike studies that focus on data collected well into the battery's life, our work emphasises the predictive value of formation data. This approach is novel and presents a stark contrast to existing literature that predominantly relies on post-manufacture test data for battery assessment. The challenge, however, lies in extracting meaningful features from the formation data that can reliably predict battery performance over its entire lifecycle. To address the research gaps in battery quality control and lifetime prediction using formation data, we propose a data-driven framework to understand the correlation between raw formation data and battery lifetime. This framework includes two machine learning models for quality classification and lifetime prediction. Initially, we extracted various features from raw electrical measurements—voltage, current, capacity, and energy—during the formation process. We then conducted a correlation analysis between these features and battery lifetime to identify those with high correlation, rationalising our findings based on ageing mechanisms, particularly SEI formation. Additionally, we reduced the dimensionality of the selected features using machine learning techniques. Given the limitations in data size and the need to avoid overfitting, we selected Ridge regression to predict battery lifetime, achieving good performance across different formation protocols. For quality classification, we employed the K-Nearest Neighbors (KNN) model, which demonstrated high accuracy. This work demonstrates the potential of data-driven methods to enhance quality assessment in battery manufacturing and reduce end-of-line testing costs.

Methodology

During the formation process, voltage and current are measured, and these measurements typically vary between different batteries due to differences in chemical composition. These measurements indirectly reflect the stability of the SEI film, which is directly related to the battery's quality and lifetime. The overall framework proposed in this work for battery quality classification and lifetime prediction using formation data is illustrated in Fig. 1. To identify the differences between batteries after the formation process, we extracted various features from the raw formation data, which can be roughly divided into three categories: statistical data, differential voltage analysis (widely used in ageing analysis [33]) and electrochemical characteristics. In total, over 100 features were extracted from the raw formation data.

We analysed the correlation between extracted features and battery lifetime, focusing on their relationship with ageing mechanisms and explaining the possible reasons behind high correlations. After the correlation analysis, features with low correlation to battery lifetime were filtered out. Due to the limited size of data samples, it was necessary to

reduce the number of model inputs further to avoid overfitting. Instead of directly filtering out features with low correlation, we used principal component analysis (PCA) to reduce the dimensionality of the model inputs without losing valuable information. Since the battery samples were from the same production batch, their quality and characteristics were similar, making the classification task particularly challenging. We used the KNN model for battery quality classification, which considers the proximity and similarity between data points to assign them to the most similar clusters. For battery lifetime prediction, we used a regularised linear regression model, which achieved high prediction accuracy while avoiding overfitting.

Dataset description

The dataset used in this work, obtained from [32], comprises 40 NMC/graphite LIB pouch cells each with a nominal capacity of 2.36 Ah. These 40 cells were divided into two groups, each subjected to a different formation protocol. The formation process, which typically occurs after the assembly of the cells, is shown in Fig. 2(a). In this study, the two formation protocols are referred to as 'baseline formation' and 'fast formation'. As shown in Fig. 2(b), the baseline formation protocol consists of two C/10-C/10 charge-discharge cycles. In contrast, the fast formation protocol consists of five C/5-C/5 cycles at the beginning, limited to a voltage range of 3.9V to 4.2V. At the end of both formation protocols, a diagnostic cycle is carried out, which includes a six-hour relaxation stage. After the formation process, the battery cells were further divided into two groups and aged with cyclic ageing tests at 45°C and 25°C, as shown in Fig. 2(c). The cycling conditions for the ageing tests included a 1C, constant current (CC) charge to 4.2 V with a constant voltage (CV) hold to 10mA and 1 C discharges to 3.0 V. Diagnostic tests such as reference performance tests (RPTs) and hybrid pulse power characterisation (HPPC) were conducted to assess the cells' low-rate capacity and internal resistances. RPTs involved C/3 and C/20 charge-discharge cycles, while HPPC provided insights into the resistance at various states of charge (SOC). These tests were performed periodically to monitor the cells' health throughout the ageing process. Interruptions in the dataset correspond to these checkup procedures, which are crucial for evaluating the cells' performance and resistance changes over time.

Feature extraction

Before the feature extraction, the prediction targets of the machine learning models must be determined. EOL is a well-known indicator marking the end of safe operation, beyond which batteries may no longer meet requirements and pose higher safety risks [34]. In this study, the prediction targets are set at 80% of the initial capacity (EOL80) [35] and 70% of the initial capacity (EOL70) [36], corresponding to the EOL for the first and second life of batteries, respectively. As illustrated in Fig. 3(a), the life cycles at which batteries reach EOL80 and EOL70 are considered for ageing analysis. In addition to the life cycles reaching EOLs, the remaining capacity at specific life cycles delivers significant information about battery ageing. We defined three retention of cycle (ROC) points at 300, 350 and 400 cycles for further analysis, as shown in Fig. 3(a). Another important ageing indicator is the degradation knee point, which marks the point in the degradation curve where ageing begins to accelerate abruptly [37]. Identification of knee points has been explored in the literature [38,39] and we used the kneed package in Python [40] to identify the knee points in the ageing dataset, as shown in Figs. 3(a) and S1.

Based on the analysis of the raw data characteristics, we extracted features using various approaches. First, we employed an automatic feature extraction method [41] to derive statistical features. For example, we counted the time each battery cell spent within specific voltage ranges during formation, such as the accumulated time in a voltage range greater than 3.5 V during the first charge-discharge cycle. Another key feature is the total formation time spent in a voltage range greater than 3.9 V throughout the entire formation cycle, as shown in Fig. 3(b). Similarly, all possible accumulated times in different voltage ranges were extracted as statistical features to investigate the influence of formation voltage, as summarised in Table S1.

Second, we extracted new features from the raw formation data using differential analysis, as shown in Fig. 3(c). In addition to the commonly analysed relationship between capacity (Q) and charge voltage (V), we expanded our analysis to include charge energy (E) and charge voltage (V) measurements during the final diagnostic cycle of the formation process. This allowed us to calculate dE/dV, representing the rate of change of energy with respect to voltage. By examining the dE/dV versus V curve, we identified peaks and their corresponding locations as a kind of feature using the scipy.signal package for peak detection algorithms [42]. Moreover, we also calculated dV/dE and performed similar smoothing operations to dV/dQ (DV) and dQ/dV (IC) analyses, as shown in Fig. S2.

The peaks and corresponding locations from the IC and DV analyses are recognised as strong indicators for estimating the State-of-Health (SOH) of the battery. Bloom et al. established a connection between DV analysis and physical mechanisms of LIBs, suggesting it provides insights into side reactions at the anode, which can lead to capacity degradation [33]. By incorporating dE/dV analysis as a complementary approach, we aim to gain a more comprehensive understanding of battery degradation processes. This novel approach expands our diagnostic capabilities and potentially offers unique insights into the underlying mechanisms affecting battery health and lifespan.

Finally, we extracted features based on the electrochemical

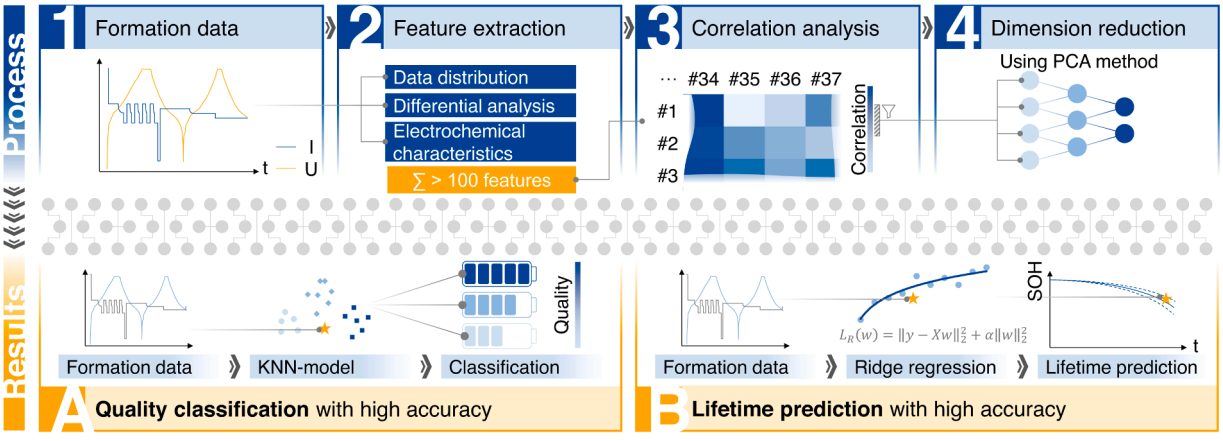


Fig. 1. A data-driven framework to do the quality classification and lifetime prediction.

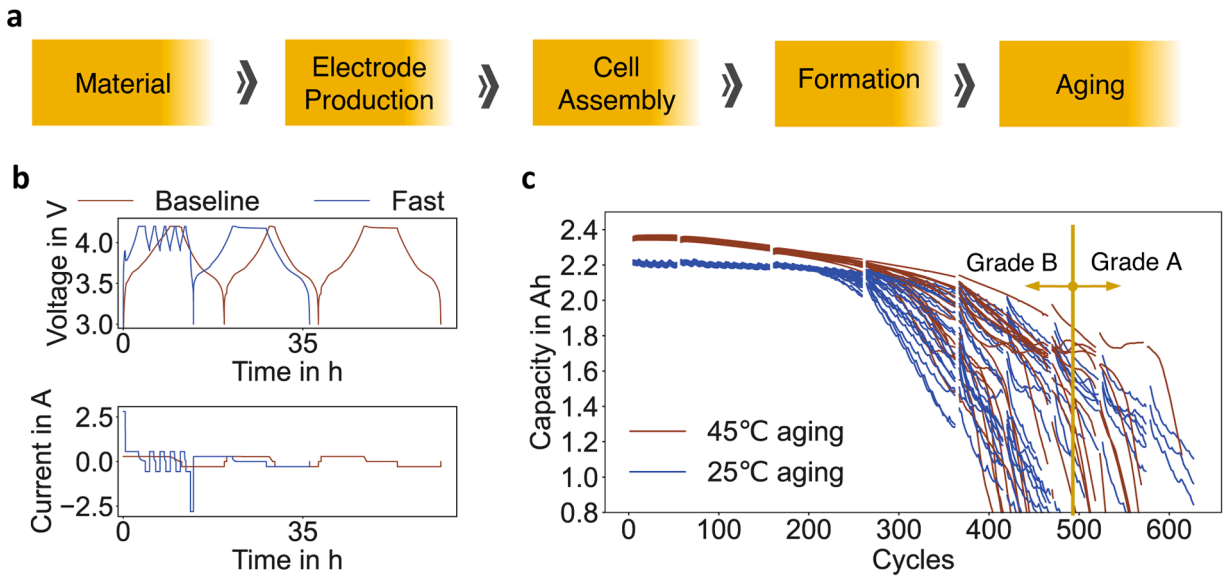


Fig. 2. Battery formation process and dataset description. a) Main process in battery production. b) Two formation protocols in the dataset, i.e., baseline and fast formation protocols. c) Battery ageing performance in 45°C ageing test and 25°C ageing test.

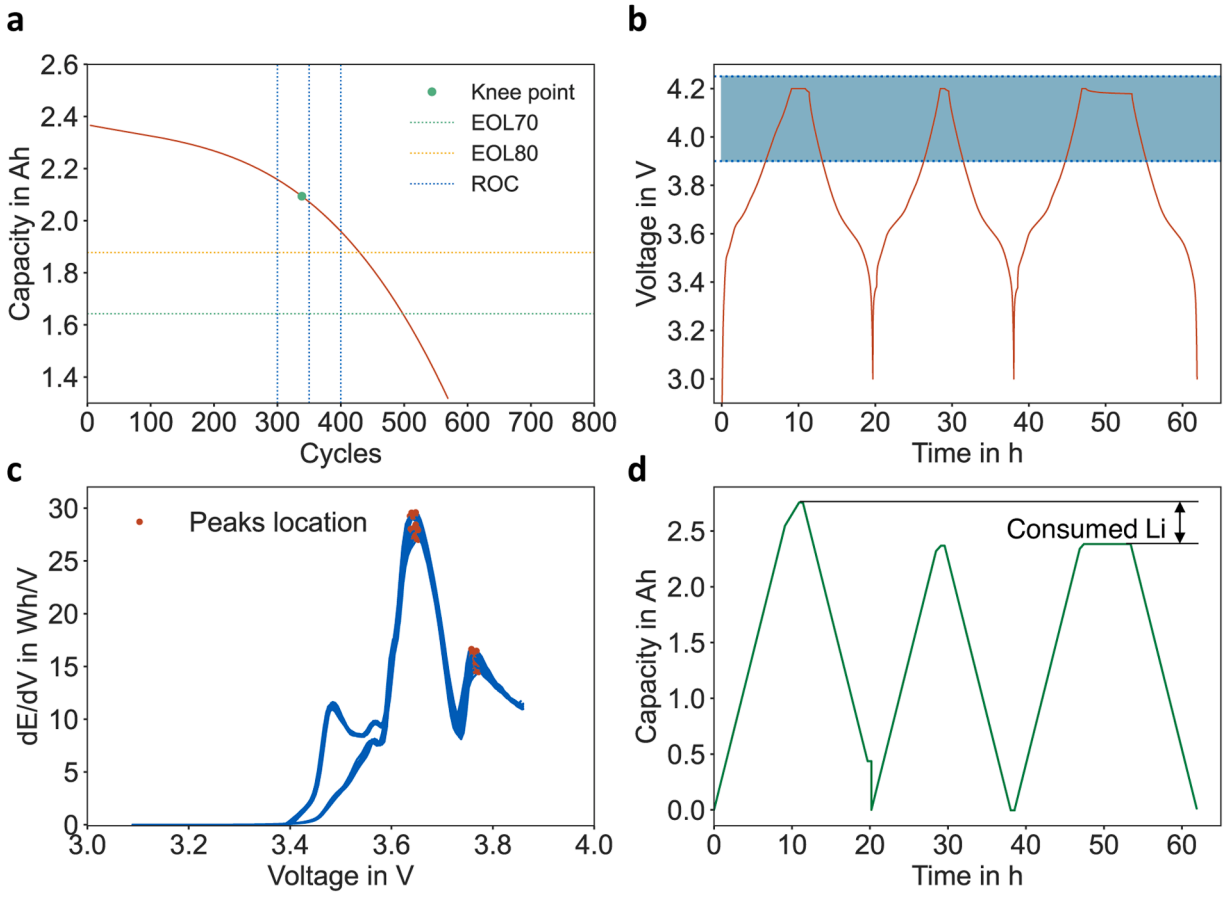


Fig. 3. Typical feature extractions from the raw data. (a) Connotation of prediction targets. (b) Accumulated duration at a specified voltage range in the formation process. (c) Relationship between incremental energy and voltage. Peaks and corresponding locations are recognised. (d) Consumed lithium in the formation process. "ROC" denotes battery retention of a certain cycle number.

mechanisms of LIBs during formation, as shown in Fig. 3(d). This figure illustrates the consumed Li/Li+ as a feature where lithium loss is associated with the growth of the SEI layer [43,44]. SEI growth serves as a model for the battery ageing mechanism, considering kinetic limitation [9,10] and solvent diffusion [11,12]. Assessing the SEI layer offers a measurable parameter to predict the EOL of LIBs, providing valuable insights for battery management systems and replacement strategies. Utilising consumed lithium as a feature of SEI growth enhances

predictive capabilities and deepens the understanding of LIB ageing. This feature is supported by extensive empirical data and theoretical foundations, highlighting the intricate relationship between lithium consumption, SEI development, and overall battery degradation. Additionally, features based on measurements during the diagnostic cycle's relaxation stage, such as the voltage after 1s/10s/60s of the final charge in the formation process, are included. In summary, Table 1 presents the three categories of descriptive features and their extraction methods.

Correlation analysis

After feature extraction, over 100 features are available. Correlation analysis helps prioritise the most significant features, which is crucial for understanding the ageing mechanisms of LIBs. This process also reduces computation costs and enhances prediction performance by selecting key features [45]. Recent studies employ various methods for feature selection, categorised into three main approaches: wrapper, filter, and embedded method [46].

The wrapper method involves multiple training iterations, adjusting the subset of features in each iteration. For example, genetic algorithms are commonly used in this context [47,48]. Despite its effectiveness, the wrapper method's black-box nature limits our ability to explain the correlation between selected features and ageing mechanisms. In contrast, the filter method ranks features, often using heatmaps to display correlation coefficients [49]. This approach allows for comparison among features and provides initial insights into their correlation with ageing mechanisms, as it ranks all features at once. The embedded method combines aspects of both wrapper and filter methods, sharing the wrapper methods's limitations in explaining ageing mechanisms. Thus, this work employs the filter method for correlation analysis of features extracted from formation data, offering a balance between interpretability and effectiveness.

Various coefficients can describe the correlation of features, such as the Pearson coefficient [50–52], Spearman coefficient [53], and grey relation coefficient [54,55]. According to Greenbank et al., the performance of these methods shows little difference [41]. In this work, the Pearson coefficient is chosen for correlation analysis due to its computational efficiency and the availability of the two-tailed P-value. The Pearson coefficient is calculated as follows:

$$\rho = \frac{\sum (F_i - \overline{F})(Z_j - \overline{Z})}{\sqrt{\sum (F_i - \overline{F})^2 \sum (Z_j - \overline{Z})^2}} \quad (1)$$

where F_i is the descriptive feature, Z_j is the target value, $\overline{F}$ is the mean of the descriptive feature, and $\overline{Z}$ is the mean of the target values. This equation represents the covariance of F and Z divided by the product of their standard deviations.

In statistical analysis, the P-value is the probability of observing data as extreme as, or more extreme than, the current data under the null hypothesis [56]. It assesses the statistical significance of the association between a feature and the target variable. Specifically, a low P-value indicates that the observed Pearson correlation coefficient, or one more extreme, would be unlikely if we incorrectly assumed no relationship between the feature and the target. Therefore, a lower P-value provides stronger evidence of a statistically significant association. In this work, a P-value of 0.05 is used as the threshold for statistical significance in feature filtering [57]. This conventional threshold demarcates the boundary between decisive results and random variation. Consequently, any features with a P-value at or below this threshold are considered candidates for further analysis.

Table 1
Categories and definitions of extracted features from the formation cycles.

Source	Subclass	Definition
Raw data (Feature Class A)	All cycles of formation	Cumulative time from V_i to V_j
	First cycle of formation	Cumulative time from V_i to V_j
Differential analysis (Feature Class B)	dV - dQ	Peaks in y vs. x plot (x = Q, V; y = dV/dQ, dQ/dV)
		Locations of peaks in y vs. x plot
	dV - dE	Peaks in y vs. x plot (x = E, V; y = dV/dE, dE/dV)
		Locations of peaks in y vs. x plot
Electrochemical characteristics (Feature Class C)	During relaxation	Voltage 1s/10s/60s after last charge
		Voltage difference between time point i and time point j after the battery starts the relaxation
		Difference between the maximum voltage and the voltage after 6 hours of relaxation during the formation process
		Daily voltage decrease calculated based on the voltage drop over the last 2 hours
		Voltage decrease per second calculated based on the initial 15 minutes
	Others	Total consumed lithium
		First charge capacity
		First discharge capacity
		First discharge capacity / First charge capacity
		Final discharge capacity
		Final discharge capacity / First charge capacity

Dimensionality reduction

Dimensionality reduction is a crucial preprocessing step for modelling, especially given the large number of features remaining after feature filtering. This work employs PCA due to its simplicity and effectiveness in handling the challenges posed by a limited number of samples. PCA reduces the dimensionality of the dataset by computing linear combinations of features while retaining the most relevant information [58]. This method is widely adopted in battery research for feature engineering [59–61]. Given the small dataset size and the potential risk of overfitting, PCA is particularly suitable. By extracting principal components that capture the maximum variance in the original feature space, PCA enables dimensionality reduction while preserving key patterns and structures in the data. This results in a more concise representation without compromising the model's predictive capacity.

While other methods such as Multi-Dimensional Scaling (MDS) [62], t-distributed Stochastic Neighbor Embedding (t-SNE) [63], and Autoencoder [64,65] offer alternative approaches to dimensionality reduction, they have limitations. These include the need for pairwise comparisons, computational complexity, and potential instability due to small sample sizes. In contrast, PCA strikes a balance between simplicity and performance, making it an optimal choice for our research objectives. Therefore, PCA is selected as the primary dimensionality reduction method to extract the most informative features from our dataset. The Python Scikit-learn framework [66] is utilised for the implementation of PCA, enabling efficient processing and analysis of the data.

Machine learning models

In this section, we utilise the dimensionality-reduced features obtained through the aforementioned techniques and apply machine learning algorithms for battery quality classification and lifetime prediction. Due to the limited number of samples, deep learning is unsuitable for this work, as it requires a much larger dataset to train and accurately estimate all predictor parameters. For the classification task, KNN is chosen due to its robustness and suitability for smaller datasets. KNN operates on an instance-based learning approach, where

classification decisions are based on the similarity between new instances and existing data points. This simplicity makes KNN an effective choice for our supervised learning task. Moreover, like many in early-stage battery research, our dataset may not be evenly distributed across all classes, especially since the battery samples were from the same production batch. KNN is less sensitive to such imbalances compared to more complex models that might overfit dominant classes. For the regression task, regularised linear regression is selected. This method imposes constraints on the model complexity, making it well-suited to handle the limitations of a smaller dataset. Regularised linear regression balances capturing relevant patterns in the data while minimising the undue influence of noise, thereby enhancing the model's predictive accuracy. In summary, KNN and regularised linear regression are employed for classification and regression tasks, respectively, to leverage the reduced features and achieve reliable predictions in battery quality and lifetime assessment. As for implementing the models, we employed well-established libraries from scikit-learn, a widely-used Python module for machine learning. Specifically, we utilised sklearn.neighbors for the KNN algorithm and sklearn.linear_model for regularised linear regression.

Quality classification

In the KNN algorithm, we compute the distances between the feature vectors of the test sample and its neighbours, then assign the test sample to the class label most commonly observed among its nearest neighbours. However, determining the optimal value of K, representing the number of nearest neighbours to consider, is not straightforward.

The choice of K is critical as it directly influences the classification results. A low value of K puts more weight on nearby instances and may result in overfitting due to increased sensitivity to noise in the data. Conversely, a high value of K introduces higher computational complexity and may lead to the blending of different classes, compromising the effectiveness of KNN [67]. While there is no universal method to determine the best K value, empirical guidelines suggest using $K = 5$ or the square root of the number of samples.

In this work, we validated our choice of K through the validation results, ensuring that the selected value was optimal for our training dataset and generalisable to unseen data. By evaluating the performance metrics and comparing the classification results, we determined $K = 5$, which yields the most accurate and reliable battery quality classification.

Lifetime prediction

Respecting the combination features from the dimension reduction section as the input, three regularised linear regression models, namely Ridge regression, Lasso regression, and Elastic Net regression, are used for lifetime prediction using formation data. Unlike simple linear regression, the three regularised models add penalty terms to the end of the simple linear predictor. Ridge regression redefines the error function as follows [68]:

$$L_D(\boldsymbol{w}) = \frac{1}{2}(y - X\boldsymbol{w})^T(y - X\boldsymbol{w}) \tag{2}$$

$$L_R(\boldsymbol{w}) = \frac{1}{2}\| \boldsymbol{w} \|_2^2 \tag{3}$$

Then ridge regression considers the regression problem as:

$$\text{argmin}(L_D(\boldsymbol{w}) + \lambda L_R(\boldsymbol{w})) \tag{4}$$

where X, y, and $\boldsymbol{w}$ represent the feature matrix, target vector and weight vector, respectively. $L_D(\boldsymbol{w})$ is the standard error function of the least-squares (LS) problem, while $L_R(\boldsymbol{w})$ is the penalty term. λ represents the regularisation parameter that controls the trade-off between fitting the training data and keeping the coefficients small. A larger λ means greater regularisation, which can prevent overfitting but may lead to underfitting if it is too large.

For Lasso regression [69], the regularised term is changed to:

$$L_R(\boldsymbol{w}) = \frac{1}{2}\| \boldsymbol{w}\|_1 \tag{5}$$

Ridge regression, distinguished by its L2 regularisation term, introduces a penalty proportional to the square of the magnitude of regression coefficients. This regularisation technique, unlike Lasso regression, allows for the simultaneous inclusion of multiple features by penalising their contributions without necessarily driving coefficients to absolute zero. The intrinsic advantage of Ridge regression lies in its ability to mitigate multicollinearity among predictor variables, a common concern in datasets with numerous features. By constraining the coefficients and preventing them from reaching extreme values, Ridge regression effectively stabilises the model against the adverse effects of collinearity, thereby enhancing the generalisation capability of the predictive model.

The Elastic Net [70] combines the L1 and L2 norms into the error function, providing a balance between feature selection and handling multicollinearity. The penalty of the Elastic Net is shown as follows:

$$L_R(\boldsymbol{w}) = \frac{1-\alpha}{2}\| \boldsymbol{w} \|_2^2 + \alpha \| \boldsymbol{w}\|_1 \tag{6}$$

However, given the specific characteristics of our dataset, the additional flexibility offered by Elastic Net to drive some coefficients to absolute zero might not be advantageous. With a relatively small sample size and a considerable number of features, the risk of discarding potentially relevant predictors is a concern. Elastic Net's feature selection capability might excessively prune variables that could hold predictive value, potentially leading to an oversimplified model that disregards important contributing factors. Therefore, in our pursuit of a robust and reliable predictive model for battery life estimation, Ridge regression was deemed more suitable. It maintains a balance between model complexity and predictive accuracy, addressing the multicollinearity issue without risking the exclusion of potentially valuable predictors. Hence, Ridge regression was chosen as the preferred algorithm for the prediction task in our study.

To evaluate the performance of the predictors, we used mean percentage error (MPE) and standard deviation (Std), which can be calculated as follows:

$$MPE = \frac{1}{N}\left|\frac{y_{pred} - y_{true}}{y_{true}}\right| \times 100\% \tag{7}$$

$$Std = \sqrt{\frac{\sum (MPE_i - \overline{MPE})^2}{K}} \tag{8}$$

where y_{pred} and y_{true} represent the predicted and observed values, respectively. N is the number of cells in each temperature case, and K comes from the repeated random subsampling validation. MPE and Std represent the prediction accuracy and robustness, respectively. In this work, repeated random subsampling validation is used rather than K-fold cross-validation because it does not depend on the order of the data, and the results can be more convincing. Specifically, we conducted a comprehensive validation process where we randomly selected 4 out of the 20 available cells as the test dataset in each iteration. Given the small dataset size, we exhaustively considered all possible combinations of selecting 4 cells from 20, which amounts to a total of 4845 unique combinations. In the dataset at room temperature, these 4 batteries were used as test data for a total of 4845 experiments, which is equivalent to generating 19380 samples. As there are only 19 batteries in the data under high-temperature conditions, 4 batteries will be extracted from the 19 batteries each time, resulting in 15504 samples. This result will have the highest reliability in classification problems. This approach allowed us to evaluate the model's performance across all potential test scenarios.

Results and discussion

Feature selection and dimensionality reduction

Fig. 4 presents the three class features and the correlation coefficients with the six prediction targets, using the Pearson coefficient to display the correlation results. The Pearson coefficient ranges from 0 to 1, with greater absolute values, indicating stronger correlations. Fig. 4 (a) illustrates the correlation results for formation features in the 25°C case, while Fig. 4(b) shows the results for the 45°C case. In Fig. 4(a), we specifically focus on the correlation analysis results for 40 features related to feature class A (FA). These selected FA features exhibit similar patterns and have been grouped together for clarity and concise representation. Comprehensive definitions and explanations of the remaining

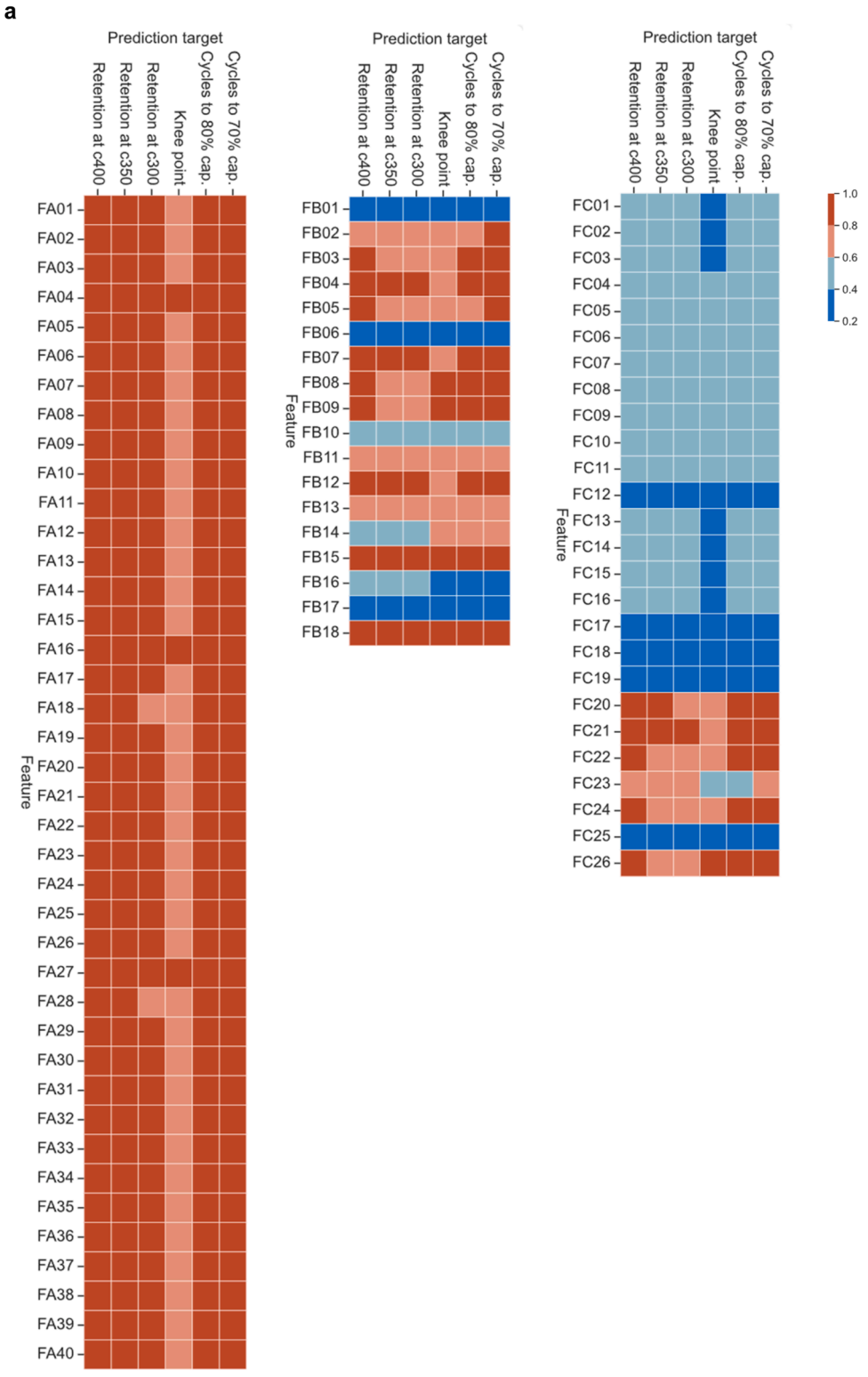


Fig. 4. Correlation results of features extracted from formation data. a) 25°C ageing test. b) 45°C ageing test. FA/FB/FC means feature Class A/B/C and all features extracted from the formation process can be found in Tables S1–S3."c300" denotes cycle 300 and "cap." signifies capacity.

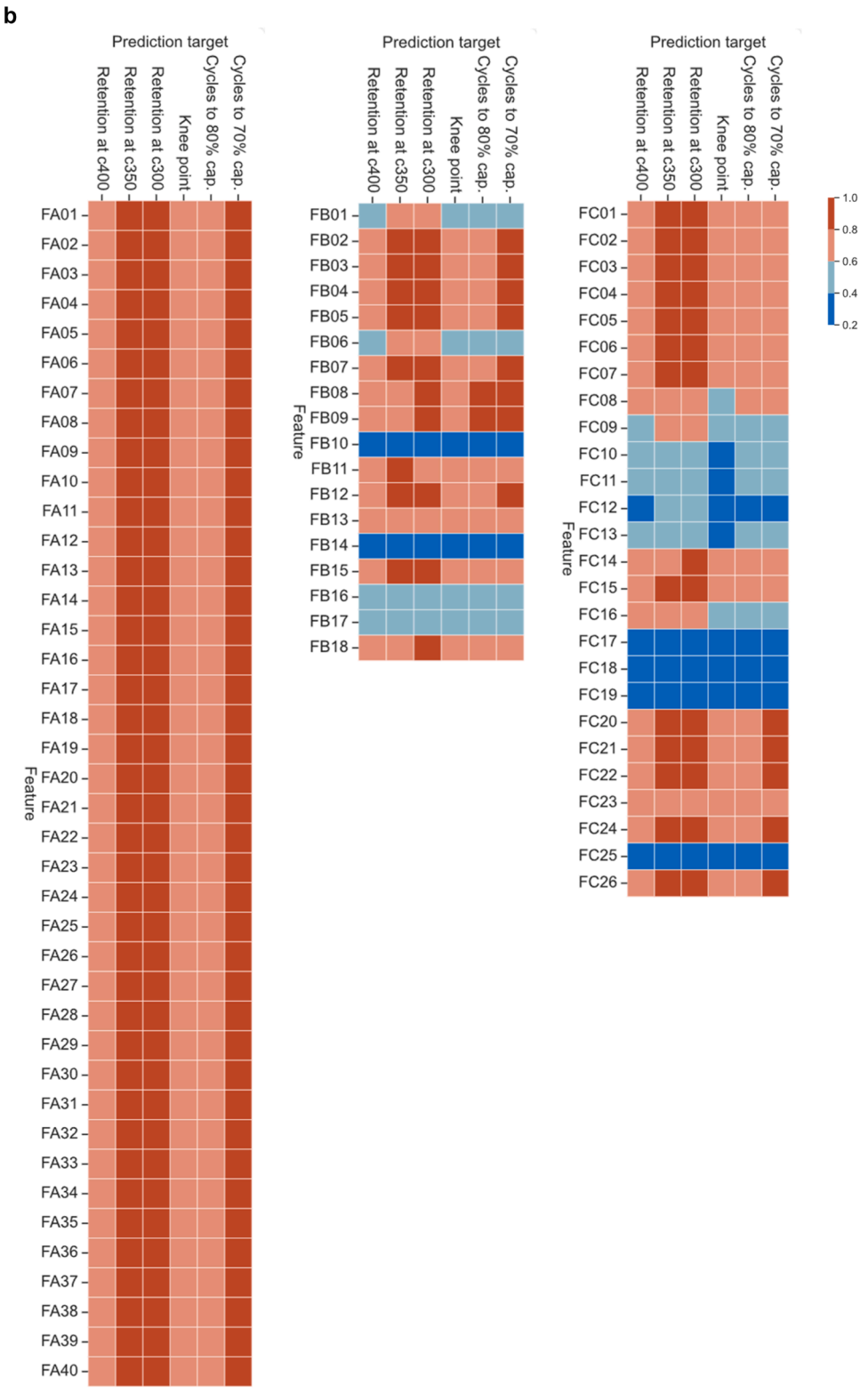


Fig. 4. (*continued*).

FA features are summarised in Table S1. The first two features with the highest correlations with battery ageing, as measured by the Pearson correlation coefficient, are FA75 and FA73. FA75 represents the cumulative time that the battery voltage exceeds 3.9 V, while FA73 denotes the cumulative time spent between 3.9 V and 4.0 V during the formation process. These findings suggest that the duration of high voltage exposure during the formation process is a significant predictor of battery lifespan.

Fig. 4(b) presents a compelling comparison of the correlation analysis at an elevated temperature of 45°C, offering insights into the temperature-dependent dynamics of battery ageing. When juxtaposing the findings from the 45°C scenario with those from the 25°C condition, it is evident that FA features continue to demonstrate robust correlations with ageing indicators. This underscores the resilience of these features

as predictive markers across a range of temperatures. This consistency suggests that FA features, which encapsulate various cumulative time metrics at different voltage thresholds during the formation process, may hold universal significance in predicting battery lifespan.

Additionally, Fig. 4 reveals significant insights into the predictive capabilities of Feature Classes B (FB) and C (FC) on LIBs. These features, derived from differential operation analyses and electrochemical characteristics, show strong potential in forecasting battery ageing behaviour.

The FB features, particularly those related to the peaks observed in the dV/dQ and dV/dE plots, exhibit correlation coefficients ranging from 0.7 to 0.9 with the battery ageing indicators. This suggests a strong linear relationship between these parameters and the ageing process. Notably, the amplitude and coordinates of these peaks at various charge states provide a quantitative measure of the electrochemical reactions occurring within the battery, which are crucial for understanding the underlying ageing mechanisms. For instance, the sharpening of a dV/dQ peak at the LiC_{12} stage is not related to anode degradation but is strongly related to the intercalation content in the graphite anode [71]. The nearly identical results from both dV-dE and dV-dQ analyses imply that the interplay between energy and voltage during the charge and discharge cycles holds equivalent predictive power for LIB lifetime estimation. This finding is significant, as it indicates that either energy or voltage alone can serve as a reliable proxy for assessing the battery's health and RUL.

Furthermore, the FC features, which pertain to voltage measurements taken during the relaxation phase post-charge, show significant correlations with battery ageing. The relaxation phase is a critical period where the battery's internal state stabilises, and any voltage changes during this time can indicate the battery's internal resistance and its

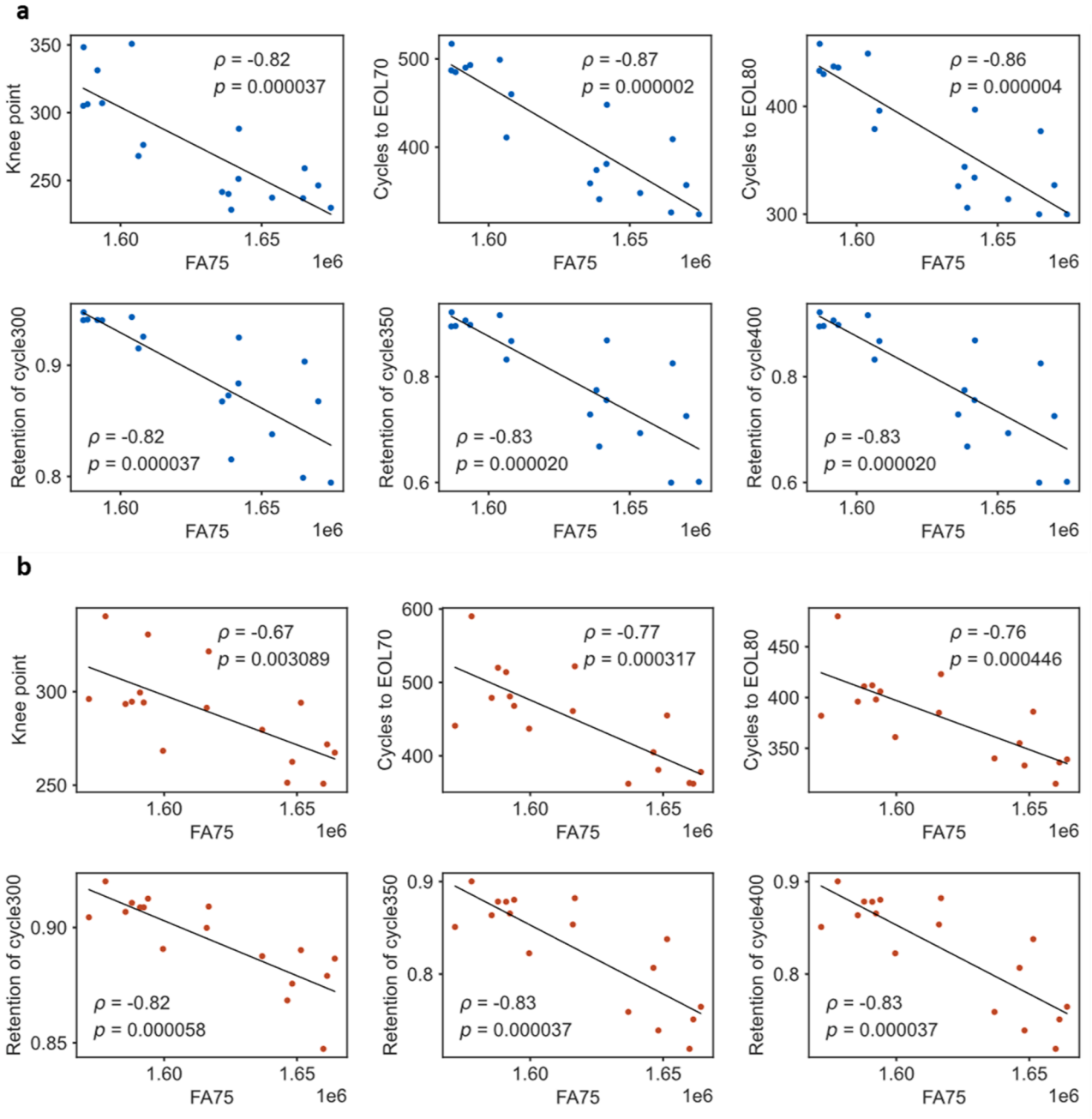


Fig. 5. Correlation analysis of feature FA75. FA75 means the cumulative time greater than 3.9V in the formation process. (a) Correlation analysis of the feature with various targets in 25°C case. (b) Correlation analysis of the feature with various targets in 45°C case.

propensity for lithium deposition (LD). Voltage measurement during this relaxation phase can be considered a form of pseudo-impedance, which, as suggested by Katzer et al. [72], can be related to the detection of LD. Since LD is a known factor that can compromise battery performance and safety, the ability to predict its occurrence through the analysis of FC features is of considerable practical importance.

Fig. 5 shows the correlation results between FA75 and various prediction targets. Two data points were chosen as the validation set and not used to calculate the correlation scores, making the results more statistically convincing. The correlation scores for Class A features are about 0.8 concerning targets of cycles to EOL70 and cycles to EOL80, indicating a strong relationship. For targets related to retention at a specific cycle number, the correlation is slightly lower than for the target cycles at EOL (cycles to EOL70, cycles to EOL80). This is likely due to the influence of long-term ageing factors not fully captured by the formation process data.

Attia et al. found that ethylene carbonate (EC) loss does not passivate the electrode if the negative electrode potential is around 0.5 - 1.2 V during the first charging [73]. This implies the total voltage should be below 3.5 V for the cells, suggesting that extending the high voltage duration can help create a more passivating interface. However, Fig. 5 illustrates a negative correlation between high voltage duration and the prediction targets, suggesting that extended high voltage periods may not necessarily lead to improved passivation. It is important to note that feature FA75 was obtained considering all charge-discharge cycles in the formation, not just the first charge process.

Das et al. modelled the electrochemical kinetics of SEI growth and measured SEI thickness, proposing that SEI growth is restrained during delithiation and promoted during lithiation [74]. Therefore, using the duration of voltage range across all formation cycles as a feature effectively represents the interaction of lithiation and delithiation for SEI growth. As observed in the anode, the non-uniform SEI growth can lead to inhomogeneous lithium-ion flux and stress distribution across the electrode, affecting battery performance and lifespan. Additionally, the mechanical properties of the SEI layer, including its response to volume changes during lithiation and delithiation, can influence its stability and overall battery health.

In light of these insights, the negative correlation between high voltage duration and prediction targets might indicate the SEI layer's vulnerability to over-voltage conditions, which could accelerate its degradation and compromise battery life. Therefore, optimising voltage strategies and understanding the SEI's mechano-electrochemical behaviour are essential for enhancing battery quality and lifespan.

As mentioned earlier, several high-correlation features serve as inputs to the predictor, necessitating further dimensionality reduction. In this work, PCA was used to generate two new features that encapsulate the information from the extracted features. PCA captures the linear relationships between the features while maintaining the robustness required to represent the primary information of the dataset. After applying PCA, two features were obtained, along with two 2×20 matrices, representing the two temperature cases and two different dimensionality reduction methods. The effectiveness of PCA will be evaluated based on the final model prediction results in the next section.

Quality classification

We classified the quality of the batteries based on their lifetime into Class A and Class B using the KNN algorithm, where Class A represents higher quality than Class B. The dataset contained only functional cells; in a real factory setting, an additional category of completely rejected cells would be present, which can be identified without the advanced methods suggested in this work.

The prediction task was treated as a classification task in machine learning. The quality classification results are shown with the confusion matrix in Fig. 6. We used the same rigorous subsampling validation method as previously described, ensuring the robustness of the results

a

		Predicted class: Class A	Predicted class: Class B	
True class	Class A	TP = 5754	FN = 60	5814
	Class B	FP = 1928	TN = 11638	13566
		7682	11698	AC = 89.74%

b

		Predicted class: Class A	Predicted class: Class B	
True class	Class A	TP = 7344	FN = 816	8160
	Class B	FP = 816	TN = 6528	7344
		8160	7344	AC = 89.47%

Fig. 6. Classification results of EOL80 using ridge regression. (a) 25°C ageing test. (b) 45°C ageing test.

despite the small sample size. The values in the table come from the subsampling validation. For instance, out of 20 batteries, 4,845 iterations of subsampling validation were performed, each selecting a subset of 4 batteries for validation. This process culminated in a total of 19,380 predictions. The classification results are composed of four subclasses: TP (True Positive), FN (False Negative), FP (False Positive), and TN (True Negative), where Class A is considered positive and Class B negative. For instance, TP means the true class is Class A and was correctly predicted as Class A, while FP means the true class is Class B but was incorrectly predicted as Class A. The larger the number of TP and TN, the better results. We achieved a classification accuracy of 89.74% for the 25°C ageing case and 89.47% for the 45°C ageing case. This shows a significant performance improvement compared to the control group, which used only the final discharge capacity of the formation as the basis for judgment and obtained accuracies of 55.04% for the 25°C ageing case and 57.75% for the 45°C ageing case.

FN can be understood as an economic waste problem because high-quality batteries (Class A) are used in applications that require only Class B cells. If Class A cells could be sold at higher prices than Class B cells, FN cases lower potential revenue. Conversely, if low-quality batteries (Class B) are assigned to applications requiring higher-quality cells (Class A), performance or warranty issues might emerge, referred to as the specificity in machine learning. In this study, we found the specificity to be 85.79% for the 25°C ageing case and 88.89% for the 45°C ageing case. This means the classifier has high confidence in selecting low-quality batteries among all Class B batteries. These results demonstrate the reliability and high potential of using formation data for battery quality classification.

Lifetime prediction

The prediction results of the battery lifetime using formation data with the regularised regression model are shown in Fig. 7. The true values of EOL80 were measured in both temperature cases, ranging from 300 to 500 cycles. Prediction performance for EOL70, retention at cycle 300/350/400, and knee points can be found in Figs. S4–S8. Table 2 presents the mean prediction percentage error for various prediction targets across two different aging tests at 25°C and 45°C. We utilised common high-correlation features for both temperature cases and formation protocols, reducing the dimensionality to two. We assumed that these two features could represent almost all the information from the extracted dataset features. Notably, the model exhibits a uniformly

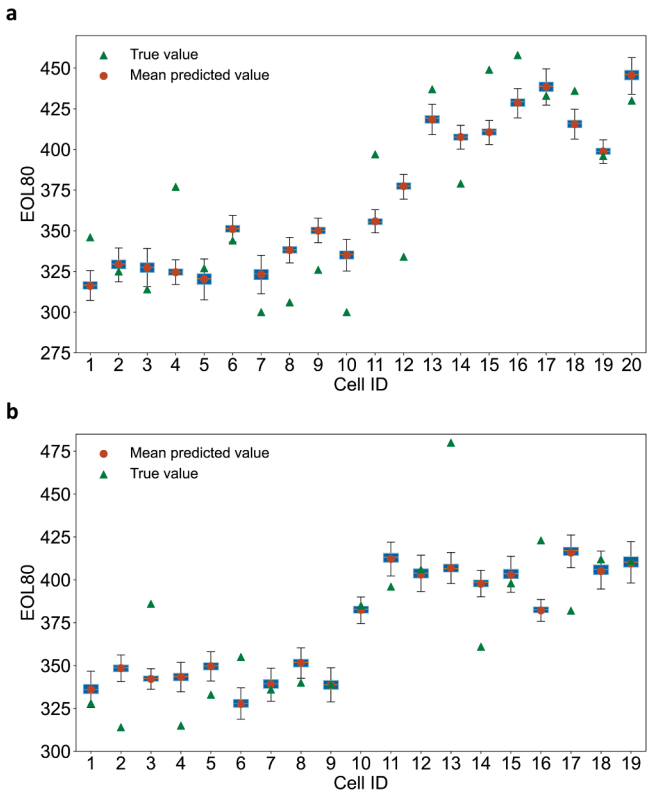


Fig. 7. Prediction results of EOL80 using ridge regression. (a) 25°C ageing test. (b) 45°C ageing test.

Table 2
Mean prediction percentage error of different prediction targets in 25°C ageing test and 45°C ageing test

Ageing temperature	Prediction target	Mean percentage error (%)
25°C	EOL80	6.50
	EOL70	6.91
	Retention at cycle 300	2.74
	Retention at cycle 350	6.79
	Retention at cycle 400	9.65
	Knee point	6.98
45°C	EOL80	5.45
	EOL70	7.79
	Retention at cycle 300	1.05
	Retention at cycle 350	4.29
	Retention at cycle 400	12.64
	Knee point	5.03

lower error rate across all prediction targets in the 45°C case. Interestingly, the Retention at cycle 300 prediction stands out with the lowest error rate of only 1.05%, implying that the model maintains a high level of accuracy even under the accelerated ageing conditions of higher temperatures. The EOL80 prediction also shows a robust performance with an MPE of 5.45%, further underscoring the model's reliability in anticipating end-of-life under thermal stress. However, the retention at cycle 400 is associated with the highest error rate of 12.64%, suggesting that predicting battery capacity at this advanced ageing stage presents a more significant challenge.

Subsampling validation was performed for each combination probability, ensuring maximum reliability but resulting in relatively high computation costs. This rigorous methodology generated an extensive dataset comprising 969 unique prediction values per cell, derived from a comprehensive sampling process. Specifically, out of 20 batteries, a staggering 4,845 iterations of subsampling validation were performed, each selecting a subset of 4 batteries for validation. This process culminated in a total of 19,380 predictions, averaging 969 predictions per battery. The resultant boxplots eloquently map the dispersion of these predictions across the entire spectrum of test groups. This variability may stem from measurement errors in the true values. However, the narrow range of predictions and their close proximity to the actual values underscore the robustness and reliability of our predictive framework. The mean predicted values, nestled within the error bars, consistently approach the true values, reinforcing the validity of our model. This method, which culminated in an MPE of 6.50% for the 25°C case and a slightly lower MPE of 5.45% for the 45°C case, underscores the efficacy of utilising formation data as a pivotal input for crafting battery lifetime prediction models. The consistency and tight clustering of predictions around the true values highlight the model's remarkable ability to generalise and predict battery lifespan with high fidelity, even when subjected to the rigours of the subsampling validation.

Outlook and applications

The formation process, while critical, generates data that is limited in scope, capturing only the initial charge-discharge cycles. This early-stage data may not fully encapsulate the long-term degradation mechanisms that unfold over the battery's lifecycle. Additionally, the variability in formation conditions and the inherent noise in the data can complicate the extraction of robust, predictive features. Despite these limitations, the formation data holds significant value when carefully analysed. Efficient generation and usage of battery production data offer significant potential in the digitalisation of battery production and acceleration of battery ageing evaluation. Looking forward to the future of battery quality control and lifetime prediction in battery production, several key areas present opportunities for exploration and improvement. First, efficient data extraction using various sensing technologies, such as pressure and temperature sensors during the battery formation process, is crucial for enhancing the predictive capabilities of battery lifetime models. Second, integrating formation data representing electrical, thermal and mechanical behaviours will pose challenges for data-driven and physics-based models. Developing a model structure that can effectively handle multi-dimensional datasets from the formation process will further enhance quality classification and lifetime prediction performance. Third, the data-driven approaches proposed in this work can be used in the closed-loop optimisation process to improve formation protocols while reducing the time needed for battery performance evaluation, significantly accelerating the optimisation of formation parameters.

This work provides a proof-of-concept for data-driven battery quality classification and lifetime prediction. Due to the limitations of the dataset size, deep learning models have not been used in this work. However, in large-scale production at giga factories, deep learning models are expected to perform better quality control and lifetime prediction, which need validation with industrial partners. Additionally, the physical interpretation of data-driven models should be further explored to support the understanding and physics-based modelling of the formation process, accelerating the scaling up of the battery production capabilities. Improving the transferability of models will further reduce the time and cost required to develop new models for the new battery cells, including updates in chemistry.

Conclusion

This work proposed a data-driven framework for classifying battery quality and predicting the battery lifetime using only the formation data from 40 NMC/graphite batteries. To the best of our knowledge, this is the first study to achieve high classification and prediction accuracy for battery ageing solely based on formation data. The new framework includes extracting various features from raw formation data, feature engineering to reduce feature dimensionality, and applying two machine learning models for quality classification and lifetime prediction.

The feature extraction combines different methods to extract over 100 features, including statistical features from electrical measurements, peaks and locations in differential analysis, and electrochemical features, covering the most critical information from the formation process. Correlation analysis revealed that high voltage duration is one of the features with the highest correlation with ageing. Using PCA with Ridge regression, the dimensionality of the features was successfully reduced with low computational burden and high robustness. As a result, we obtained an accuracy of 89.74% for the 25°C ageing case and 89.47% for the 45°C ageing case in the quality classification task. For the lifetime prediction task, we achieved a test error of 6.50% in the 25°C ageing case and 5.45% in the 45°C ageing case, validating the high performance of the proposed data-driven approach.

This study highlights the significant potential of using data collected during the formation process in battery production to support quality assurance tasks. By successfully applying machine learning methods, we demonstrate that it is possible to predict the lifetime of LIBs from the same manufacturing batch with high accuracy. The suggested framework can introduce an additional quality gate between formation and ageing in the battery production process, potentially reducing or replacing portions of the costly end-of-line testing procedures.

CRediT authorship contribution statement

Jiayu Zou: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Yingbo Gao:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Moritz H. Frieges:** Writing – review & editing, Visualization, Investigation. **Martin F. Börner:** Writing – review & editing, Validation, Investigation. **Achim Kampker:** Writing – review & editing, Investigation, Funding acquisition. **Weihan Li:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This work has received funding from the research project “SPEED” (03XP0585) funded by the German Federal Ministry of Education and Research (BMBF). Part of the work was done within the research project "Model2life" (03XP0334), funded by the German Federal Ministry of Education and Research (BMBF).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.egyai.2024.100451.

Data availability

Data used in this study is available at Deep Blue Data: https://doi.org/10.7302/pa3f-4w30. The source code is available on request.

References

[1] Schmuch R, Wagner R, Hörpel G, Placke T, Winter M. Performance and cost of materials for lithium-based rechargeable automotive batteries. Nat Energy 2018;3:267–78.
[2] Kampker A, Heimes H, Offermanns C, Wennemar S, Robben T, Lackner N. Optimising the cell finishing process: an overview of steps. Technolog Trend WEVJ 2023;14:96.
[3] Duffner F, Mauler L, Wentker M, Leker J, Winter M. Large-scale automotive battery cell manufacturing: Analysing strategic and operational effects on manufacturing costs. Int J Product Econ 2021;232:107982.
[4] Winter M. The Solid Electrolyte Interphase – The Most Important and the Least Understood Solid Electrolyte in Rechargeable Li Batteries. Zeitschrift für Physikalische Chemie 2009;223:1395–406.
[5] Zhang S, Ding MS, Xu K, Allen J, Jow TR. Understanding solid electrolyte interface film formation on graphite electrodes. Electrochem Solid-State Lett 2001;4:A206.
[6] Wang A, Kadam S, Li H, Shi S, Qi Y. Review on modeling of the anode solid electrolyte interphase (SEI) for lithium-ion batteries. npj Comput Mater 2018;4.
[7] Arora P, White RE, Doyle M. Capacity Fade Mechanisms and Side Reactions in Lithium-Ion Batteries. J Electrochem Soc 1998;145:3647–67.
[8] Wood DL, Li J, Daniel C. Prospects for reducing the processing cost of lithium ion batteries. J. Power Source 2015;275:234–42.
[9] Appiah WA, Park J, Byun S, Ryou M-H, Lee YM. A mathematical model for cyclic aging of spinel $LiMn_2O_4$ /graphite lithium-ion cells. J Electrochem Soc 2016;163.A2757-A2767.
[10] Ramadass P, Haran B, Gomadam PM, White R, Popov BN. Development of first principles capacity fade model for Li-ion cells. J Electrochem Soc 2004;151:A196.
[11] Kamyab N, Weidner JW, White RE. Mixed mode growth model for the solid electrolyte interface (SEI). J Electrochem Soc 2019;166:A334–41.
[12] Tahmasbi AA, Kadyk T, Eikerling MH. Statistical physics-based model of solid electrolyte interphase growth in lithium ion batteries. J Electrochem Soc 2017;164:A1307–13.
[13] Kupper C, Weißhar B, Rißmann S, Bessler WG. End-of-life prediction of a lithium-ion battery cell based on mechanistic aging models of the graphite electrode. J Electrochem Soc 2018;165. A3468-A3480.
[14] Birkl CR, Roberts MR, McTurk E, Bruce PG, Howey DA. Degradation diagnostics for lithium ion cells. J Power Source 2017;341:373–86.
[15] Safari M, Morcrette M, Teyssot A, Delacourt C. Life-prediction methods for lithium-ion batteries derived from a fatigue approach. J Electrochem Soc 2010;157:A713.
[16] Yang X-G, Leng Y, Zhang G, Ge S, Wang C-Y. Modeling of lithium plating induced aging of lithium-ion batteries: transition from linear to nonlinear aging. J Power Source 2017;360:28–40.
[17] Lui YH, Li M, Downey A, Shen S, Nemani VP, Ye H, VanElzen C, Jain G, Hu S, Laflamme S, Hu C. Physics-based prognostics of implantable-grade lithium-ion battery for remaining useful life prediction. J Power Source 2021;485:229327.
[18] Severson KA, Attia PM, Jin N, Perkins N, Jiang B, Yang Z, Chen MH, Aykol M, Herring PK, Fraggedakis D, Bazant MZ, Harris SJ, Chueh WC, Braatz RD. Data-driven prediction of battery cycle life before capacity degradation. Nat Energy 2019;4:383–91.
[19] Xue Z, Zhang Y, Cheng C, Ma G. Remaining useful life prediction of lithium-ion batteries with adaptive unscented kalman filter and optimised support vector regression. Neurocomputing 2020;376:95–102.
[20] Pang X, Liu X, Jia J, Wen J, Shi Y, Zeng J, Zhao Z. A lithium-ion battery remaining useful life prediction method based on the incremental capacity analysis and Gaussian process regression. Microelectron Reliab 2021;127:114405.
[21] Richardson RR, Osborne MA, Howey DA. Gaussian process regression for forecasting battery state of health. J Power Source 2017;357:209–19.
[22] Cheng G, Wang X, He Y. Remaining useful life and state of health prediction for lithium batteries based on empirical mode decomposition and a long and short memory neural network. Energy 2021;232:121022.
[23] Zhao S, Zhang C, Wang Y. Lithium-ion battery capacity and remaining useful life prediction using board learning system and long short-term memory neural network. J Energy Storage 2022;52:104901.
[24] Lu J, Xiong R, Tian J, Wang C, Hsu C-W, Tsou N-T, Sun F, Li J. Battery degradation prediction against uncertain future conditions with recurrent neural network enabled deep learning. Energy Storage Mater 2022;50:139–51.
[25] A. Rastegarpanah, C.A. Contreras, R. Stolkin, Hyperparameter-optimized CNN and CNN-LSTM for Predicting the Remaining Useful Life of Lithium-Ion Batteries, in: 2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS), IEEE, 11212023 pp. 110–115.
[26] He N, Wang Q, Lu Z, Chai Y, Yang F. Early prediction of battery lifetime based on graphical features and convolutional neural networks. Appl Energy 2024;353:122048.
[27] Xie P, Pang X, Wang C, Yang W, Zou H, Zhao W, Chen S, Liu Z. A sequence to sequence prediction model for remaining useful life of lithium-ion batteries with Bayesian optimisation process visualisation. J Energy Storage 2024;87:111346.
[28] Li W, Zhang H, van Vlijmen B, Dechent P, Sauer DU. Forecasting battery capacity and power degradation with multi-task learning. Energy Storage Mater 2022;53:453–66.
[29] Fei Z, Yang F, Tsui K-L, Li L, Zhang Z. Early prediction of battery lifetime via a machine learning based framework. Energy 2021;225:120205.
[30] Zhang Y, Xiong R, He H, Pecht MG. Lithium-ion battery remaining useful life prediction with Box–Cox transformation and Monte Carlo simulation. IEEE Trans Ind Electron 2019;66:1585–97.
[31] Hsu C-W, Xiong R, Chen N-Y, Li J, Tsou N-T. Deep neural network battery life and voltage prediction by using data of one cycle only. Appl Energy 2022;306:118134.
[32] Weng A, Mohtat P, Attia PM, Sulzer V, Lee S, Less G, Stefanopoulou A. Predicting the impact of formation protocols on battery lifetime immediately after manufacturing. Joule 2021;5:2971–92.
[33] Bloom I, Christophersen JP, Abraham DP, Gering KL. Differential voltage analyses of high-power lithium-ion cells. J Power Source 2006;157:537–42.
[34] Li W, Sengupta N, Dechent P, Howey D, Annaswamy A, Sauer DU. One-shot battery degradation trajectory prediction with deep learning. J Power Source 2021;506:230024.

[35] USABC electric vehicle Battery Test Procedures Manual. Revision 2, 1996.
[36] Kim E, Wu B, Shin K, Lee J, He L. Adaptive battery diagnosis/prognosis for efficient operation. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems. New York, NY, USA: ACM; 2019. p. 150–9.
[37] IEEE Recommended practice for sizing lead-Acid batteries for stationary applications. IEEE Std 485–2010 (Revision of IEEE Std 485–1997) 2011;:1–90.
[38] Fermín-Cueto P, McTurk E, Allerhand M, Medina-Lopez E, Anjos MF, Sylvester J, dos Reis G. Identification and machine learning prediction of knee-point and knee-onset in capacity degradation curves of lithium-ion cells. Energy AI 2020;1:100006.
[39] Diao W, Saxena S, Han B, Pecht M. Algorithm to Determine the Knee Point on Capacity Fade Curves of Lithium-Ion Cells. Energies 2019;12:2910.
[40] kneed-PyPI Developers, kneed-PyPI. https://pypi.org/project/kneed/.
[41] Greenbank S, Howey D. Automated feature extraction and selection for data-driven models of rapid battery capacity fade and end of life. IEEE Trans Ind Inf 2022;18:2965–73.
[42] scipy.signal Developers, scipy.signal.savgol_filter and scipy.signal.find_peaks.
[43] Purewal J, Wang J, Graetz J, Soukiazian S, Tataria H, Verbrugge MW. Degradation of lithium ion batteries employing graphite negatives and nickel–cobalt–manganese oxide + spinel manganese oxide positives: Part 2, chemical–mechanical degradation model. J Power Source 2014;272:1154–61.
[44] Yu Y, Yang Z, Liu Y, Xie J. Achieving SEI preformed graphite in flow cell to mitigate initial lithium loss. Carbon 2022;196:589–95.
[45] Xue B, Zhang M, Browne WN, Yao X. A survey on evolutionary computation approaches to feature selection. IEEE Trans Evol Computat 2016;20:606–26.
[46] N.R. Barraza, On the statistical comparison of feature selection methods and the role of experts: the case of Las Vegas strip, 2020, 2020. http://sedici.unlp.edu.ar/handle/10915/116426.
[47] Gheyas IA, Smith LS. Feature subset selection in large dimensionality domains. Patt Recognit 2010;43:5–13.
[48] Jeong Y-S, Shin KS, Jeong MK. An evolutionary algorithm with the partial sequential forward floating search mutation for large-scale feature selection problems. J Oper Res Soc 2015;66:529–38.
[49] Jebli I, Belouadha F-Z, Kabbaj MI, Tilioua A. Prediction of solar energy guided by pearson correlation using machine learning. Energy 2021;224:120109.
[50] Goh HH, Lan Z, Zhang D, Dai W, Kurniawan TA, Goh KC. Estimation of the state of health (SOH) of batteries using discrete curvature feature extraction. J Energy Storage 2022;50:104646.
[51] Kong J, Yang F, Zhang X, Pan E, Peng Z, Wang D. Voltage-temperature health feature extraction to improve prognostics and health management of lithium-ion batteries. Energy 2021;223:120114.
[52] Liu J, Chen Z. Remaining useful life prediction of lithium-ion batteries based on health indicator and Gaussian process regression model. IEEE Access 2019;7:39474–84.
[53] Zhang Y, Wik T, Bergström J, Pecht M, Zou C. A machine learning-based framework for online prediction of battery ageing trajectory and lifetime using histogram data. J Power Source 2022;526:231110.
[54] Xu T, Peng Z, Wu L. A novel data-driven method for predicting the circulating capacity of lithium-ion battery under random variable current. Energy 2021;218:119530.
[55] Li Y, Stroe D-I, Cheng Y, Sheng H, Sui X, Teodorescu R. On the feature selection for battery state of health estimation based on charging–discharging profiles. J Energy Storage 2021;33:102122.
[56] Berman JJ. Principles and practice of big data. Preparing, sharing, and analysing complex information. London: /Jules J. Berman, Academic Press; 2018.
[57] S. Mcleod, What a p-value tells you about statistical significance. https://www.simplypsychology.org/p-value.html.
[58] I.T. Jolliffe, Principal component analysis, Springer International Publishing, Cham, 20.
[59] Sheng H, Xiao J, Wang P. Lithium iron phosphate battery electric vehicle state-of-charge estimation based on evolutionary Gaussian mixture regression. IEEE Trans Ind Electron 2017;64:544–51.
[60] Lipu MSH, Hannan M A, Hussain A. Feature selection and optimal neural network algorithm for the state of charge estimation of lithium-ion battery for electric vehicle application. Int J Renew Energy Res (IJRER) 2017;7:1700–8.
[61] Guo P, Cheng Z, Yang L. A data-driven remaining capacity estimation approach for lithium-ion batteries based on charging health feature extraction. J Power Source 2019;412:442–50.
[62] Borg I, Groenen PJF. Modern multidimensional scaling. New York, London: Theory and applications /Ingwer Borg, Patrick J.F. Groenen, Springer; 2005.
[63] L. van der Maaten, G. Hinton, Visualising Data using t-SNE (2008) 2579–2605.
[64] Kingma DP, Welling M. An introduction to variational autoencoders. FNT Mach Learn 2019;12:307–92.
[65] A. Abid, M.F. Balin, J. Zou, Concrete Autoencoders for Differentiable Feature Selection and Reconstruction, 2019.
[66] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: machine learning in python. J Mach Learn Res 2011.
[67] R. Saxena, KNN Classifier, introduction to K-nearest neighbor algoriithm.
[68] Saleh AKME, Kibria G. Theory of ridge regression estimators with applications. Hoboken, New Jersey: John Wiley & Sons; 2016.
[69] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J Roy Statist Soc: Ser B (Methodolog) 1996;58:267–88.
[70] Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statist Soc B 2005;67:301–20.
[71] Kato H, Kobayashi Y, Miyashiro H. Differential voltage curve analysis of a lithium-ion battery during discharge. J Power Source 2018;398:49–54.
[72] Katzer F, Danzer MA. Analysis and detection of lithium deposition after fast charging of lithium-ion batteries by investigating the impedance relaxation. J Power Source 2021;503:230009.
[73] Attia PM, Harris SJ, Chueh WC. Benefits of fast battery formation in a model system. J Electrochem Soc 2021;168:50543.
[74] Das S, Attia PM, Chueh WC, Bazant MZ. Electrochemical kinetics of SEI growth on carbon black: part II. Modeling. J Electrochem Soc 2019;166. E107-E118.