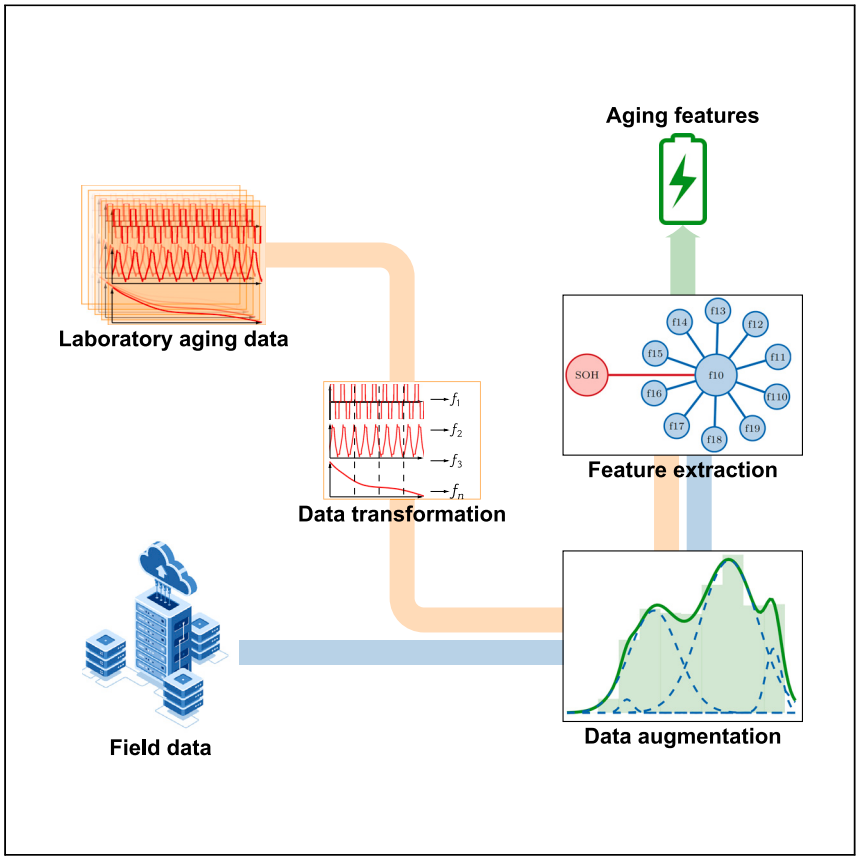
Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Automated feature extraction to integrate field and laboratory data for aging diagnosis of automotive lithium-ion batteries



This research focuses on improving battery aging diagnostics using real-world field data. Compared with laboratory data, field data present challenges in terms of complexity and inconsistent estimations. To overcome these issues, Steininger et al. propose a framework that combines accurate laboratory aging data with driving data from a large customer base. By applying various feature extraction methods and employing statistical data fitting as a data augmentation technique, the methodology achieves a significant 57% increase in aging-estimation accuracy.

Valentin Steininger, Katharina Rumpf, Peter Hüsson, Weihan Li, Dirk Uwe Sauer

valentin.steininger@bmw.de (V.S.)
weihan.li@isea.rwth-aachen.de (W.L.)

Highlights

Investigation of differences between field and laboratory data for lithium-ion batteries

Integration of field and laboratory data for aging-estimation tasks

Investigation of feature extraction methods for histogram data

Data augmentation through the functional fitting of histogram variables

Steininger et al., Cell Reports Physical Science 4, 101596
October 18, 2023
https://doi.org/10.1016/j.xcrp.2023.101596

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Automated feature extraction to integrate field and laboratory data for aging diagnosis of automotive lithium-ion batteries

Valentin Steininger,[1,2,3,4,6,*] Katharina Rumpf,[1] Peter Hüsson,[1] Weihan Li,[2,3,4,*] and Dirk Uwe Sauer[2,3,4,5]

SUMMARY

Battery aging diagnosis using field data readouts presents distinct challenges compared with using laboratory data. These challenges stem from the complexity of the data structure and potential inconsistencies in aging values obtained from variations in battery management system software versions. Consequently, the efficacy of a data-driven approach to identify pertinent aging features from field data becomes susceptible to these factors. In this work, we investigate different feature extraction methods and propose a framework designed to mitigate issues arising from compromised data quality. For this purpose, we leverage the benefits of precise laboratory aging data alongside authentic driving data acquired from a cohort exceeding 600,000 customers to improve the aging diagnosis of vehicle batteries. Moreover, we provide functional fitting of statistical data, addressing the challenges posed by incomplete data structures. We validate our methods by comparing them with state-of-the-art feature extraction techniques, yielding a 57% enhancement in aging estimation accuracy.

INTRODUCTION

Lithium-ion batteries have emerged as the first choice for large-scale energy storage in the automotive sector, effectively meeting demanding technical prerequisites like high power and energy density, coupled with long calendar and cycle life capacity. In particular, the durability of batteries poses a critical challenge for original equipment manufacturers (OEMs), who must ensure sustained battery performance for customers.[1,2] As such, the state of health (SOH), commonly defined by capacity fade and the inner resistance increase, serves as an indicator of the battery's current degradation level.

Accurate measurement of battery capacity and inner resistance necessitates offline implementation, precluding their assessment during battery operation and requiring controlled laboratory conditions. Consequently, estimation methods are inevitable for integration into the battery management system (BMS).[3] However, multifaceted physicochemical reactions within lithium-ion batteries result in highly nonlinear and complex aging behavior, making the estimation procedure a challenging task.[4] These estimation methods fall into categories of experimental, model based, and data driven. This paper focuses on the last. Given the BMS's restricted hardware resources, computationally intensive algorithms like neural networks and electrochemical models are difficult to deploy for serial adoption. Therefore, algorithms striking a balance between computational effort and estimation accuracy are of paramount interest for onboard estimation.[5,6]

[1]BMW AG, Petuelring 130, 80809 München, Germany

[2]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

[3]Center for Ageing, Reliability and Lifetime Prediction of Electrochemical and Power Electronic Systems (CARL), RWTH Aachen University, Campus-Boulevard 89, 52074 Aachen, Germany

[4]Jülich Aachen Research Alliance, JARA-Energy, Templergraben 55, 52056 Aachen, Germany

[5]Helmholtz Institute Münster (HI MS), IEK 12, Forschungszentrum Jülich, 52425 Jülich, Germany

[6]Lead contact

*Correspondence: valentin.steininger@bmw.de (V.S.), weihan.li@isea.rwth-aachen.de (W.L.)

https://doi.org/10.1016/j.xcrp.2023.101596

To monitor aging during operation, OEMs collect battery data from the internal BMS memory and send it to a global data pool for processing and analysis. To minimize BMS memory usage and ensure lifelong data logging, measured signals are often compressed into aggregated formats, such as binned histograms.[7] These data encompass measured variables (temperature, current, and voltage) and estimated states such as SOH and state of charge (SOC), enabling comprehensive customer-specific aging analysis. However, due to a limited storage capacity as well as erroneous readouts, collected data require extensive preprocessing to filter out unreliable information. Moreover, during a battery's life cycle, updated BMS software versions contain different estimation algorithms or an updated parametrization of underlying models, providing inconsistent aging values over a battery's lifetime. In contrast, laboratory data, gathered under controlled testing conditions, provide accurate aging data along with measured input signals over time but lack realistic load patterns. Therefore, laboratory data fail to capture all customer operation modes, providing insufficient information to identify appropriate features for a customer-related data-driven aging estimation.

In this work, we propose a feature extraction framework for statistical field data readouts, automating the selection of aging features for diagnosis tasks. A feature serves as an independent variable for regression and can be derived from the provided driving data format obtained from BMS readouts. This work presents established single-correlation-based methods and introduces new dual-source correlation and elementary function fitting algorithms. These leverage both field and laboratory data to optimize feature selection. We use aging correlation coefficients calculated from accurate laboratory data to select relevant aging features. In addition, we employ these coefficients to guide the mapping of collinear features identified through correlation matrices of the field data, thereby mitigating redundancies within the feature set. This approach employs realistic load patterns from a field database containing BMS readouts from a total of 600,000 customers alongside laboratory aging data of the same cell chemistry. While various feature extraction approaches exist, this is the first work to combine the advantages of laboratory and field data for enhanced diagnostic results.[8–10]

Our approach involves two key steps. First, aging features are statistically derived from accurate SOH values obtained from laboratory data. This ensures that the selected features are statistically meaningful and relevant for aging analysis. Second, features are grouped based on real driving behavior, aiming to eliminate redundancies and avoid duplication of information in the input data. The collected field data from BMS memory are mainly available in binned histograms. For harmonious compatibility, this necessitates that the laboratory data be molded into the same format, thus guaranteeing uniformity in the sets of variables across both data sources. Hence, as a first step, we transform measured time-series data into the driving-data format of the field data. This foundation enables the subsequent computation of Spearman correlation coefficients. Various aggregation methods are then employed, delineating input features based on their correlation with battery aging, while also facilitating the grouping of features according to their collinearity. Furthermore, the dimensionality of each feature group is curtailed by utilizing the partial-least-squares (PLS) method, effectively condensing redundant information. Hence, the framework automatically ascertains the statistically optimal feature set, bypassing the need for manual input construction. Finally, to authenticate the efficacy of our proposed framework, a fully connected neural network is trained using the extracted features from the laboratory data, serving as a validation of its performance. The main contributions of this paper are as follows:

- Correlation analysis and investigation of differences between field and laboratory data of 48 V lithium-ion batteries;
- Integration of field and laboratory data to combine advantages of both data sources for estimation task;
- Investigation of feature extraction methods for histogram data toward aging estimation;
- Data augmentation through the functional fitting of histogram variables.

The rest of the paper is organized as follows: "Field data" provides the collection workflow of the field data as well as the variable structure a readout in the database contains. Furthermore, the section "Laboratory data" covers the laboratory data of the battery-aging tests and the transformation into the driving-data format of collected customer data. In "Feature extraction framework," a comprehensive explanation of the framework is presented. The section "Validation" and the discussion showcase and analyze the superior performance of the framework compared with conventional correlation-based feature extraction methods.

RESULTS

Field data

For this investigation, collected data from a total of 600,000 customers were analyzed to identify estimation capabilities from gathered inputs of the battery SOH on the one hand and to assess differences among BMS estimation algorithms on the other. This section explains the format of the collected field data as well as the variables they contain. As a bridging technology, 48 V batteries are widely used in vehicle applications. In mild-hybrid vehicles (MHEVs), they achieve substantial fuel consumption savings at low system adoption costs by assisting the combustion engine with boosting and recuperating, while in electric vehicles, they support the 12 V battery with high-current peaks, e.g., for roll stabilization systems.[11] The battery is composed of 20 high-power cells connected in series with a nickel-manganese-cobalt (NMC)/lithium titanate oxide (LTO) chemistry and a nominal voltage of 2.4 V and a beginning-of-life (BOL) capacity of 10 Ah. Such cells are designed for high C rates, long cycle life, and low depth-of-discharge (DOD) rates, in contrast to high-energy cells used in traction batteries of electric vehicles, which are operated with lower C rates but high DODs.[12] The BMS controls the battery operation through data acquisition, state estimation, charge and discharge control, balance control, and others.[13] Moreover, data storage facilitates offboard fault diagnosis and is also of great significance in analyzing operation conditions. Considering that data need to be collected over multiple years of operation, compression methods are applied to reduce data storage requirements.[14] A readout from the BMS memory is either event-based or time-based, triggered and sent over the air to a data collection pool. The readout procedure can be found in a previous research paper.[15] A row in the customer database represents a full readout of the customer BMS memory, which contains both measured (current, voltage, temperature) and estimated (SOC, SOH) variables. With an average of 14.2 readouts per customer, the database comprises data from 8.9 million readouts starting from July 2020. The collected variables mainly comprise aggregated data to optimize memory usage and ensure data logging over the entire battery lifetime.

Table 1 gives an overview of available variables from the BMS memory together with their value and bin ranges, where two groups of available data can be distinguished: single values and histogram values. The first represents the instantaneous value of a signal at the time of a readout, such as the energy throughput, which has a maximum

Table 1. Selected memory variables from the BMS comprise single-value and histogram-value variables

Name	Description	Value range
Single values		
SOH	state of health at readout	[0, 100] (%)
SOC	state of charge at readout	[0, 100] (%)
energy_throughput	total battery energy throughput until readout	[0, 8,000] (kWh)
voltage	battery voltage at readout	[0, 70] (V)
current	battery current at readout	[−1,500, 1,500] (A)
temperature	battery temperature at readout	[−126, 126] (°C)
Histogram values		
time_soc_x, $x \in [1, 10]$	time spent in SOC range [0, 10, 20, …, 100] (%)	[0, $2^{32} - 1$] (s)
time_temperature_x, $x \in [1, 6]$	time spent in temperature range [‚0, 0, 20, …, >70] (°C)	[0, $2^{32} - 1$] (s)
(*dis*)*charge_temperature_x*, $x \in [1, 6]$	(dis)charge in temperature range [‚0, 0, 20, …, >70] (°C)	[0, $2^{32} - 1$] (Ah)
number_dod_x, $x \in [1, 7]$	number of DODs in range [0, 1.1, 2.2, …, >9.9] (Ah)	[0, $2^{32} - 1$] (counts

value of 8,000 kWh, approximately equivalent to 9,000 full cycles. Another example is the current with a maximum value of 1,500 A, equivalent to 150 C. Histogram values include counters of binned signal intervals, which are updated during the operation of the vehicle. Consequently, the histogram values are cumulative. For example, time-based histogram bin values refer to the time spent by a battery in a particular state parameter range, such as the variable *time_soc_1*, which stores the time in an SOC range of 0%–10%. To account for calendar aging during parking phases, we updated the histogram data by integrating the start and stop values of the respective variable over the parking time. In addition, every readout entry is associated with a unique readout ID, a hashed vehicle ID, and the current software version of the BMS. In the following, we will call this set of variables, representing a snapshot of the battery history, the driving data format. The mileage of monitored vehicles in the customer database ranges from new vehicles with 50 km up to older ones with more than 130,000 km, and the number of equivalent full cycles of the 48 V battery ranges from 0 to over 3,000.

Figure 1A illustrates the average SOH values over the equivalent full cycles of the collected data for different climate conditions and software versions during battery development. Every software version corresponds to a specific configuration of the onboard aging model, which has to be kept confidential.

Here, the SOH is defined as follows:

$$SOH_C = 100 \cdot \left(1 - \frac{C_0 - C(t)}{C_0}\right) = 100 \cdot \frac{C(t)}{C_0}, \quad \text{(Equation 1)}$$

$$SOH_R = 100 \cdot \left(1 - \frac{R(t) - R_0}{R_0}\right) = 100 \cdot \left(2 - \frac{R(t)}{R_0}\right), \quad \text{(Equation 2)}$$

$$SOH = 0.8 \cdot SOH_R + 0.2 \cdot SOH_C, \quad \text{(Equation 3)}$$

with SOH_C as the capacity degradation and SOH_R the inner resistance increase. Since the battery is mainly exposed to high C rates together with low DODs, resistance increase plays a higher role in the total SOH calculation. Accordingly, the SOH_R in Equation 3 weighs 80% and the SOH_C represents the remaining 20% of the overall battery SOH. As can be seen in Figure 1A, the calculated mean SOH trajectories show quite strong oscillation, which is mainly due to the uneven sampling time of vehicle readouts. Eventually, the SOH happens to increase once vehicle

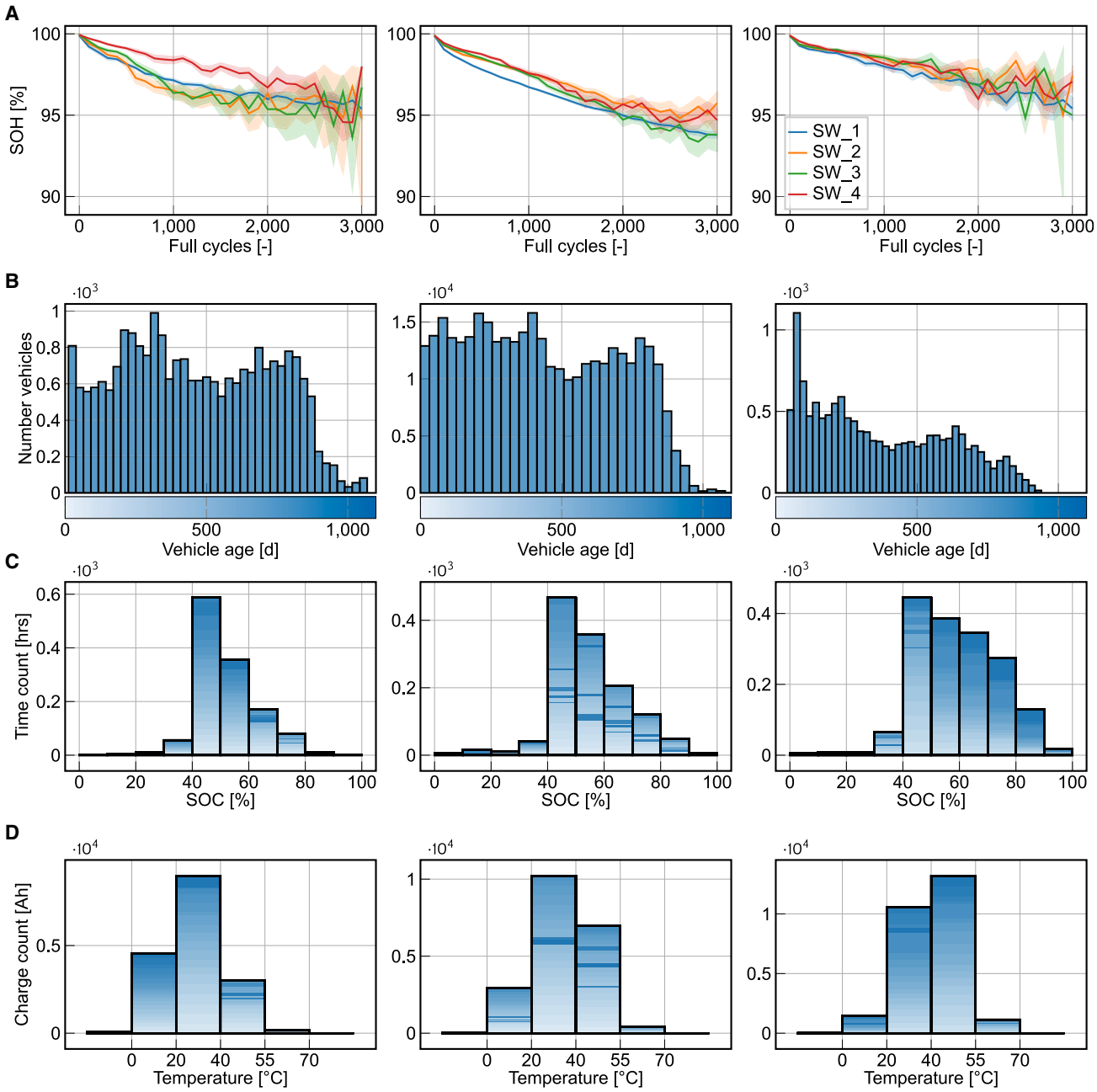


Figure 1. Collected vehicle diagnostic data for three different climate regions

Left, cold climate; middle, moderate climate; right, hot climate.

(A) Mean SOH values and 95% confidence interval for collected vehicle data from different software versions.

(B) Histogram distribution of vehicle ages.

(C) Time histogram distributions in SOC ranges over vehicle age.

(D) Charge histogram distributions under different temperatures over vehicle age.

readouts with lower SOH values are not available for the respective time step. Moreover, a discrepancy between the estimated aging behavior of different software versions is noticeable. Under cold climate conditions (left), software 4 (SW_4) estimates a slower battery degradation than the remaining software versions, whereas, under moderate conditions (middle), software 1 (SW_1) predicts a faster aging of the

battery. The best congruence among software versions can be found under hot climate conditions (right), providing similar aging trajectories. As a result, inconsistent SOH values over the collected field data make a data-driven analysis of aging features impractical. Figure 1B illustrates the distribution of vehicle ages in the respective climate regions. Current throughputs from charging the battery under different operating temperatures are illustrated in Figure 1D, where a clear shift of the main temperature operating window among climate regions can be noticed. Since cycling of the 48 V battery frequently alters due to boosting, recuperating, and supporting the 12 V onboard electrical system, the accumulated currents of charging and discharging are in very close range for every temperature bin.

To assess the estimation quality of a respective software version from collected field data, accurate aging data are needed to provide reference values from accelerated aging tests. Moreover, to understand better how to apply aggregated vehicle diagnostic data for offboard aging estimation, we investigate feature correlations from laboratory data in the following section.

Laboratory data

The laboratory data utilized for studying battery aging in various operating modes comprise measurements obtained from calendar and cyclic aging tests. These tests have been specifically designed to simulate high-power applications that experience high C rates while maintaining a low discharge depth.

The laboratory data utilized for studying battery aging in various operating modes (temperature, SOC, DOD, etc.) comprise measurements obtained from calendar- and cyclic-aging tests. These have been specifically designed to simulate high-power applications that face high C rates together with a low discharge depth.[16] The investigated cells contain an LTO/NMC chemistry with a nominal voltage of 2.4 V and compose a battery from the field data with 20 cells connected in series to provide a 48 V voltage level. Tables S1 and S2 show the corresponding test matrices of the calendar- and cycle-aging tests for a total of 25 individual cells. Main operation points comprise tests at 60°C with an SOC range from 5% to 95% for the calendar tests, as well as 5 C (dis)charging rates at 40°C for cyclic tests. Further information about test design and analysis can be found in Bank et al.,[17] including an in-depth investigation of aging implications for 48 V battery systems.

After every 30 days for calendar tests and 500 equivalent full cycles for cyclic tests, a checkup at 25°C is conducted to determine the aging progress. Therefore, a capacity test is performed with a constant current of 1 C and also a pulse test at 2, 5, 10, 20, and 25 C in discharge and charge directions to evaluate the inner resistance of the cell. The cell capacity is determined by the total current count in the discharge direction and the cell inner resistance is computed by the measured resistances at 2 C discharge pulses after 3 s. For a test duration range of 120–550 days and an equivalent full-cycle range of 7,500–23,000 cycles, respectively, this yields 450 computed values of capacity and inner resistance. Figure 2A shows the voltage curve of a checkup test with the markings to determine the aging factors. Afterward, the final SOH is computed using Equation 3 from the field data definition. In addition, to increase the number of resulting data points, we apply a piece-wise linear interpolation between two neighboring checkup tests and insert 10 query points for the SOH. Although we do not increase the information content, this step is required to provide enough data to a data-based model in the validation step later.

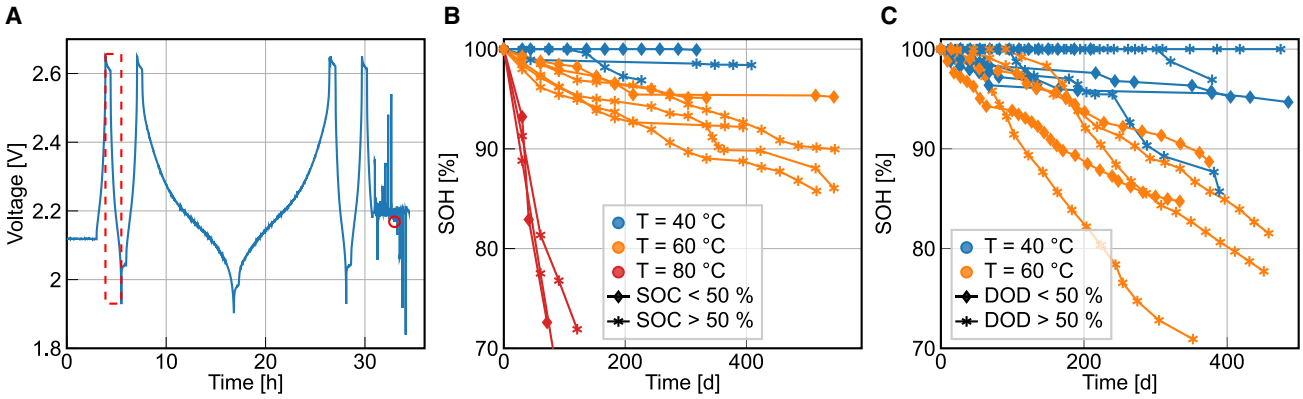


Figure 2. Laboratory cell aging tests

(A) Cell voltage profile of checkup test to determine the aging factors. Capacity is measured at a 1 C full discharge (dashed box) and the inner resistance at a 2 C discharge pulse after 3 s (circle).

(B) Resulting SOH values from calendar-aging tests after interpolation.

(C) Resulting SOH values from cyclic-aging tests after interpolation.

Figures 2B and 2C show the aging trajectories of each cell of the calendar- and cyclic-aging tests. The markers indicate the time steps of a checkup test, which also represent the sample points of the linear interpolation. As illustrated, higher temperatures lead to a faster degradation of all cells. Moreover, higher storage SOCs foster the aging of the calendar test cells. The measured data from the laboratory tests represent accurate aging data together with time-series input signals. In contrast, the collected field data comprise single and histogram values of real customer behavior but, therefore, include inconsistent SOH values. Hence, as a first step to identify possible aging factors from the field data, the laboratory data will be transformed into the driving data format presented in the "Field data" section. Therefore, at every checkup time step, current i, voltage u, and temperature T measurements from all previous time steps are collected to compute the drive data format, including all single and histogram values of the concatenated signals. In addition, the SOC is calculated by accumulating charge and discharge currents (ampere-counter method) together with voltage measurements during the checkup tests. Therefore, at every checkup test, we assume a fully charged battery at 100% SOC previous to the full discharge cycle to determine the battery capacity and construct the SOC profile from the current measurements to the next checkup test. With the measured laboratory time-series data, histogram- and single-value variables from Table 1 can be computed for each cell together with an associated SOH value from the corresponding checkup test or interpolation query point. With the sampling time vector Δt, the charge vector Q^+, and the discharge vector Q^-,

$$\Delta t = \begin{bmatrix} t_1 - t_0 \\ t_2 - t_1 \\ \vdots \\ t_n - t_{n-1} \end{bmatrix}, \quad \text{(Equation 4)}$$

$$Q^+ = i \cdot \Delta t, \text{where } i > 0, \quad \text{(Equation 5)}$$

$$Q^- = i \cdot \Delta t, \text{where } i < 0, \quad \text{(Equation 6)}$$

the following logical conditions apply to calculate the histogram variable values from the time-series data[18]:

$$time\_soc\_x \rightarrow \sum \Delta t_{SOC_l < SOC < SOC_u}, \quad \text{(Equation 7)}$$

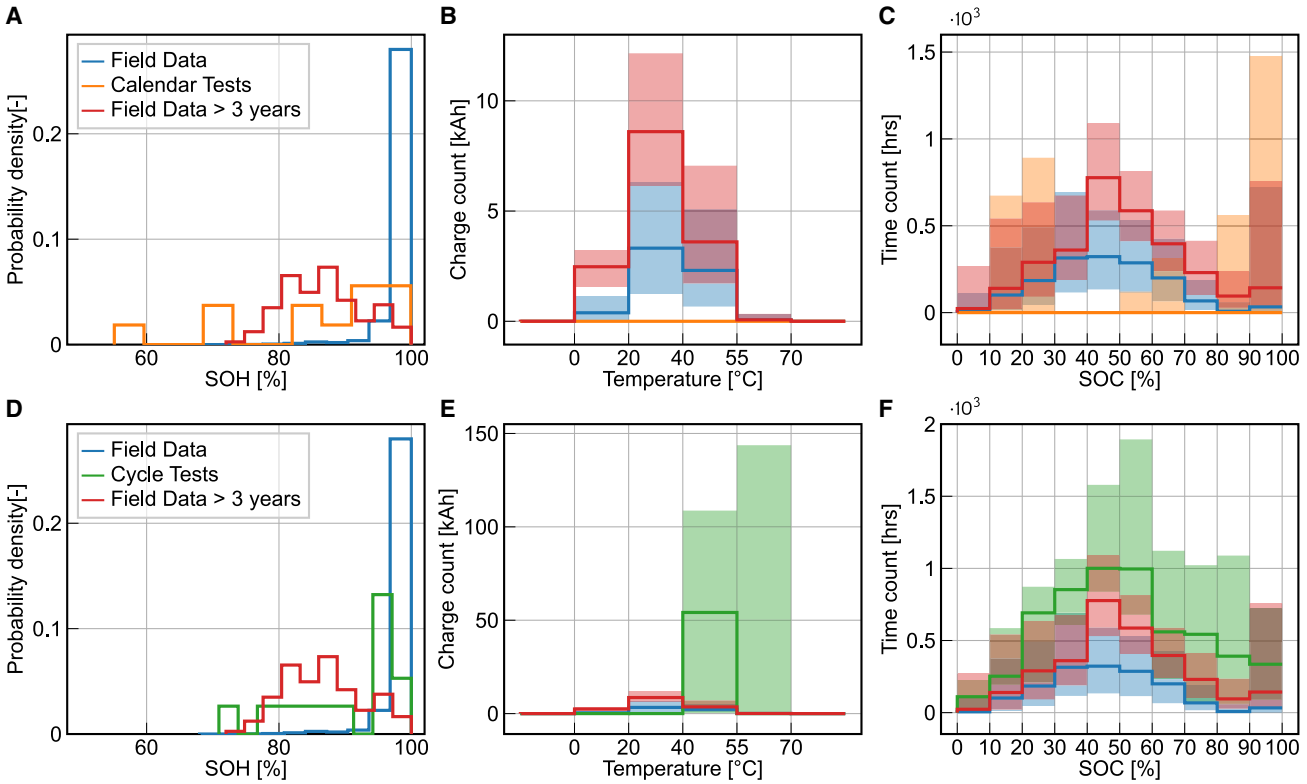


Figure 3. Distributions of the SOH and selected histogram features of the field and laboratory data

Solid lines represent mean values of histogram variables, the filled areas indicate 50% of the available data. See Figure S1 for the individual distribution plots.

(A) SOH distribution of field data and calendar test data.

(B) Charge count over temperature bins of field data and calendar test data.

(C) Time count over SOC range of field data and calendar test data.

(D) SOH distribution of field data and cycle test data.

(E) Charge count over temperature bins of field data and cycle test data.

(F) Time count over SOC range of field data and cycle test data.

$$time_{temperature_x} \rightarrow \sum \Delta t_{T_l < T < T_u}, \quad \text{(Equation 8)}$$

$$charge_{temperature_x} \rightarrow \sum Q^{+}_{T_l < T < T_u}, \quad \text{(Equation 9)}$$

$$discharge\_temperature\_x \rightarrow \sum Q^{-}_{T_l < T < T_u}. \quad \text{(Equation 10)}$$

The lower and upper parameter limits l and u define the bin range where a certain condition is met and can be found in Table 1. Note that the sum of all the times spent in each parameter range (SOC and temperature) must equal the total time elapsed within that load pattern, and the charge and discharge bin values must equal the total current throughput at the respective checkup test. Finally, assuming a similarly distributed cell degradation within a battery, we scale the resulting drive-data variables from cell level up to a 48 V battery level by considering the connection topology. To account for a series connection, the energy throughput and voltage variables are multiplied by the number of connected cells.

Figure 3 illustrates the distribution of the SOH of the field data and laboratory tests, as well as distribution diagrams of histogram data. In that respect, the solid lines

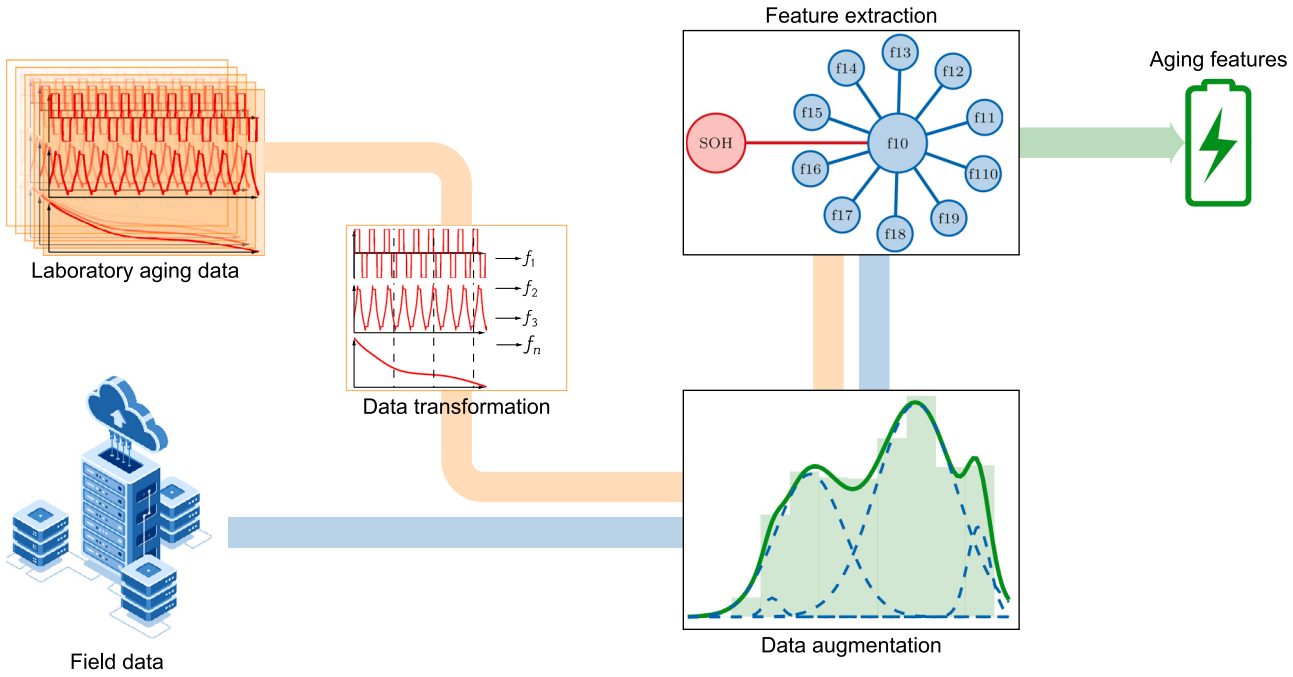


Figure 4. Feature extraction framework

At first, the collected laboratory aging data are transformed into the driving data format of the field data. After data augmentation through functional fitting, correlation matrices are merged for the feature extraction process to generate the final aging features.

represent the mean of available data, and the filled areas indicate the range where 50% of considered data points reside. As shown, the majority of field data vehicles have, in general, high SOH values due to a high percentage of new vehicles in the fleet. The same applies to SOC and temperature counter histograms, where cycle tests show higher distribution values. In contrast, selected vehicles with more than 3 years in the field show significantly lower SOH values, which peak at around 85% SOH. However, the charge throughputs are still relatively low compared with the cycle test data, but therefore, the time counts of SOC histograms are in comparable value ranges. Thus, calendar aging has been dominant for the battery degradation of those vehicles. The separate distribution plots for each respective group are shown in Figure S1.

Feature extraction framework

The feature extraction framework is designed to combine the advantages of field and laboratory data. For this, it employs field data readouts from 600,000 customers (see "Field data") and laboratory data from cell-aging tests. The workflow of the framework is illustrated in Figure 4. To join benefits from both data sources, the laboratory data (see "Laboratory data") are transformed into the driving data format first. Afterward, a data augmentation step is applied to improve the information content of compressed histogram formats by approximating the true histogram distribution. Finally, the feature extraction procedure selects aging features based on the aging correlations of the laboratory data and additionally identifies redundancies among input features from correlations of the field data. For validation, we compare a state-of-the-art feature filter for the given data with our improved filter and also analyze performance increase by the functional fitting of the histogram data.

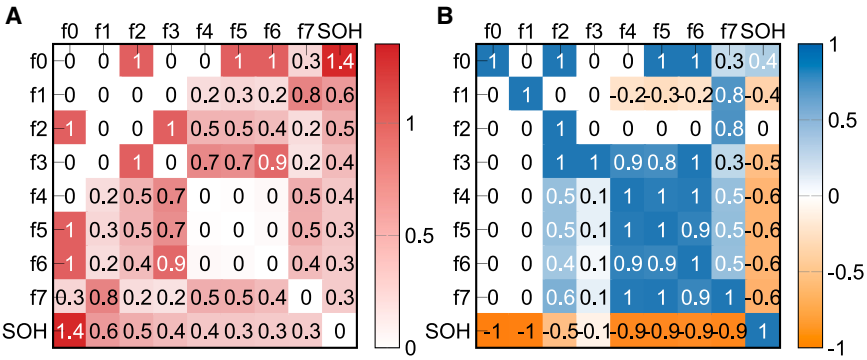


Figure 5. Correlation analysis

(A) Difference in correlation matrices of field and laboratory data for the top eight features, f0–f7, with the highest difference with regard to the SOH correlation.

(B) Correlation matrices for laboratory (botthom-left triangle) and field (top-right triangle) data. See Table 2 for the feature descriptions.

As a starting point, we compute the Spearman correlation matrix from Equation 14 to analyze dependencies between features and the SOH and also among the features themselves. The resulting correlation coefficients shall provide insight into the statistical distinction between the field and the transformed laboratory data since both data sources fundamentally differ by operating mode and SOH accuracy. On one hand, field data are characterized by realistic load points from customer behavior along with onboard SOH estimations of varying BMS software, while laboratory data are measured under controlled conditions and specific test designs. Therefore, the dependencies between features, as well as between the SOH and the features, are likely to be different. To contrast the correlations of field and laboratory data, we compute the absolute difference of both matrices, which is illustrated in Figure 5A for the top eight features, f0–f7, with the highest difference with regard to the SOH correlation. The correlation coefficients of the same features can be seen in Figure 5B for the laboratory data in the bottom-left triangle and for the field data in the top-right triangle. Further, a description of features f0–f7 is listed in Table 2.

It shows that for features f0–f3 correlation differences can be found in high temperature ranges above 70°C because such temperatures very rarely occur in the field, and thus a valid statistical dependency on SOH is not present. However, the laboratory test matrix covers a wider temperature window and, therefore, contains the required information to draw the statistical influence on the SOH. From the upper triangle in Figure 5B, it can be seen that the correlation coefficient between f0 and SOH in the field data even has a positive value, 0.4, which would imply an increasing SOH when the battery resides between 90% and 100% SOC with a temperature over 70°C. In contrast, the lower triangle of Figure 5B shows a clear negative correlation between f0 and the SOH from the laboratory data, which matches battery degradation theory. Considering dependencies among features, the correlation from f0 and f2 shows a difference of 1 between the field and the laboratory data. Since charging of the battery (f2) happens very likely when the battery is between 90% and 100% (f0) in the same temperature window, it can be concluded that the missing dependency in the laboratory data is due to a lack of operating modes. In conclusion, the correlation matrices show on one hand inaccurate correlation coefficients between features and the SOH in the field data and on the other missing dependencies among features due to a limited range of battery usage patterns.

Table 2. Top eight features with the highest SOH correlation difference between field and laboratory data

Label	Feature	Δ correlation
f0	time (s) when SOC is between 90% and 100% and temperature is above 70°C	1.35
f1	time (s) when SOC is between 60% and 70% and temperature is above 70°C	0.64
f2	charge count (Ah) when the temperature is above 70°C	0.51
f3	time (s) when SOC is between 30% and 40% and temperature is above 70°C	0.43
f4	time (s) when SOC is between 10% and 20% and temperature is between 55°C and 70°C	0.35
f5	time (s) when SOC is between 20% and 30% and temperature is between 55°C and 70°C	0.35
f6	time (s) when SOC is between 0% and 10% and temperature is between 55°C and 70°C	0.33
f7	charge count total (Ah)	0.33

To automatically select the most relevant features for a data-driven aging estimation task, correlation-based methods are common in the literature. Therefore, a high absolute correlation between SOH and a feature indicates a strong influence of that feature on battery aging, while a high correlation between a pair of features indicates information redundancy (collinearity). Hamar et al.[19] employ the Pearson correlation coefficient as a measure to select important aging features but also to mitigate collinearity by selecting highly correlated features with the SOH and discarding one feature from each collinear pair. However, this method captures the statistical information of only the most relevant feature within a correlated group and ignores potentially crucial data from the discarded features. As a remedy, we apply a correlation-based feature-grouping algorithm (see experimental procedures) where we select the highest correlated aging feature from the laboratory data correlation matrix and, in turn, allocate features that are correlated to the aging feature from the field data correlation matrix. Afterward, we compress the data of each feature group using the PLS method and thereby extract the most significant data from the entire feature group with regard to the aging estimation task. Moreover, we account for valid SOH correlations from the laboratory data and also integrate real customer behavior operating modes from the field data.

The clustering process produces a total of 12 feature groups, each containing a varying number of features, ranging from 1 to 21. As an example, Figure 6A shows three selected feature groups, f0x–f2x, of the field data from the correlation clustering process. The given correlation coefficients quantify the relationship between a feature and the SOH in red and the collinearity among features in blue. As can be seen, the number of correlated features *n* can vary between feature groups. A description of each feature, f0x–f2x, can be found in Table S3. The relevant information concerning aging degradation can now be compressed by applying the PLS method on each feature group and transforming the *n*-dimensional feature matrix into a one-dimensional significant feature vector. Figure 6B shows the reduced feature groups pls0–pls2 after applying the PLS method. Since only important information regarding the regression problem has been extracted from each feature group, the resulting aging correlation coefficients are equal to or higher than the correlation of the original feature groups. From Figure 6B it can be seen that the aging correlation increases between 0% and 9% from the initial features f0x–f2x to the transformed pls0–pls2. Furthermore, an increased number of features within a feature group does not necessarily result in a higher aging correlation, as observed between the compressed features pls0 and pls2 in Figure 6B. While the four features of group f2x do not offer additional information to better explain the SOH, the two features in f0x contribute to a higher aging correlation after the PLS. Hence, the method not only reduces the overall feature dimension of the input data but, at the same time, ensures preservation of relevant data.

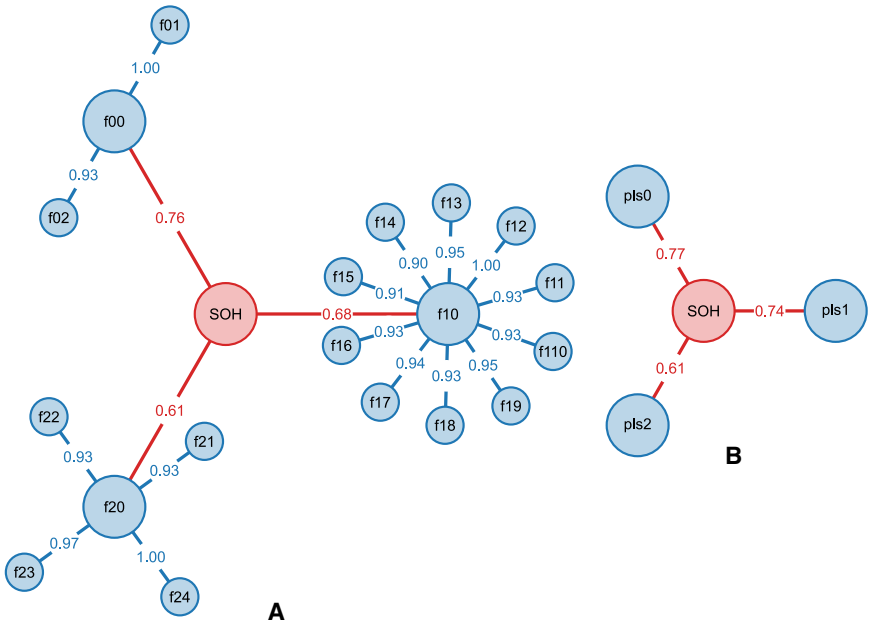


Figure 6. Improved correlation filter

See Table S3 for a description of each feature f0x–f2x.

(A) Three selected correlated feature groups were extracted from the field data. Correlation coefficients in blue quantify the collinearity among features and coefficients in red determine the relationship to the SOH.

(B) Resulting features after applying the one-dimensional PLS method on the feature groups. Correlation coefficients to the SOH are equal to or higher than the original coefficients of the respective feature group.

In the case of employing aggregated data for regression problems, it is important to note that histogram bin values have limitations in accurately capturing the true underlying distribution of the actual histogram data. Instead, they provide discrete values of accumulated counters. More importantly, wide bin ranges cause a loss of input data, which might be crucial for the regression problem. As a remedy, we apply a functional fitting algorithm to characterize the underlying distribution of histogram data using the superposition of elementary distribution functions as provided in the experimental procedures. Therefore, we superimpose as many elementary pseudo-Voigt functions as the number of peaks that have been detected in the linear histogram distribution. Since a histogram variable with n bins can have a maximum of $n - 1$ peaks and the elementary pseudo-Voigt (pV) function takes four arguments, we obtain a maximum of $3n - 4$ augmented features from the fitting process.

Figure 7 illustrates three histogram variable distributions with different numbers of detected peaks and elementary pV functions. As can be seen, to capture accurately the underlying distribution of the *time_soc_x* histogram, four pV functions are required, whereas for the *time_temperature_x* as well as the *charge_temperature_x*, one pV function is sufficient due to only one global maximum in the distribution. Since histogram distributions are not symmetrical around peaks, fit function means do not necessarily match their exact position. Each individual fitting computation requires 0.03 s, leading to a cumulative computation time of 74 h for a total of 8.9 million readouts. For our study, we use distributed executors in a cluster architecture to apply the data augmentation step, reducing the computation time to 12 h.

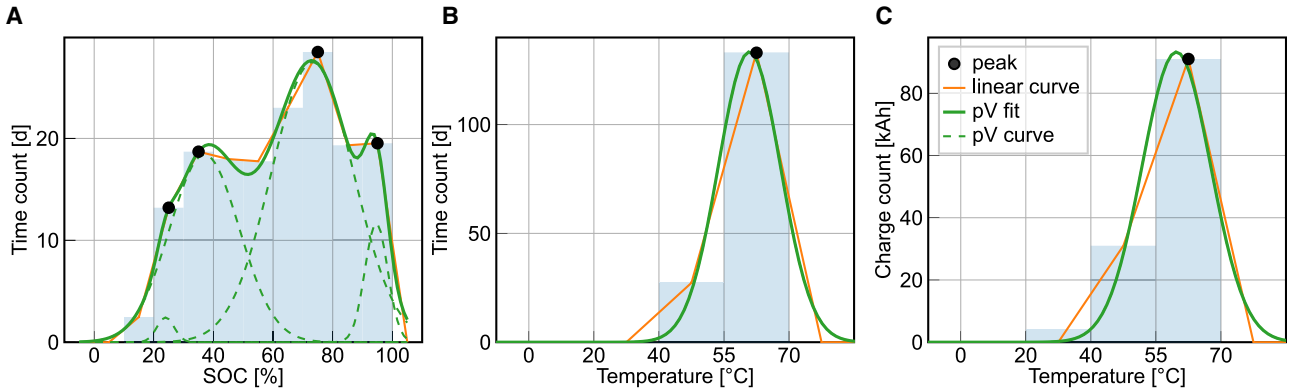


Figure 7. Pseudo-Voigt fit for three histogram variables

The number of elementary pV curves is determined by the number of local maxima (peaks) in the linear histogram distribution curve.

(A) Functional fit of SOC histogram using four pV curves.

(B) Functional fit of temperature histogram using one pV curve.

(C) Functional fit of charge histogram using one pV curve.

Validation

We apply the above feature extraction framework on the field and laboratory data and train a fully connected benchmark neural network, which consists of three hidden layers, to validate our methods. Moreover, we apply a basic correlation filter, which selects features with the highest correlation to the SOH, to assess the performance of our framework. For the network training, we split the laboratory data into 80% training and 20% test data and employed the Adam algorithm to optimize the weights of the network using a constant learning rate of 10^{-3} over 1,000 epochs. Moreover, we define an early stop criterion to abort the training process once the validation loss has not increased over 15 epochs to prevent overfitting. For each feature extraction method, we ran the training procedure five times and selected the best and worst results with regard to the test mean squared error (MSE). Figure 8 illustrates the network estimations of the laboratory SOH when using the basic correlation filter versus the improved correlation filter directly on the histogram variables and also when applying the functional fitting data augmentation approach.

In Figures 8A–8C, the estimation error is particularly high in lower SOH ranges, mainly because there are fewer data points available in those regions. As can be seen, the functional fitting data augmentation leads to a better model performance for both correlation filter methods, with a best-case improvement of 60% for the basic and 20% for the improved correlation methods. However, the functional fitting of selected features based on the basic correlation filter shows the highest test MSE (8.16) from all worst-case training runs, indicating a higher sensitivity of the model toward parameter initialization. Applying the improved correlation filter decreases the MSE by 46%, from 3.99 to 2.16. The augmented features selected by the improved correlation filter show satisfying results for both best- and worst-case scenarios. Furthermore, these features exhibit a reduced estimation error in lower SOH regions.

DISCUSSION

In this work, we propose and examine a feature extraction framework designed for statistical field data readouts, which systematically automates the selection of aging-related features for diagnostic tasks. Our approach seamlessly integrates data from laboratory aging tests and field readouts, synergistically harnessing the strengths of both data sources. On the one hand, we leverage the precision of SOH values derived from rigorously

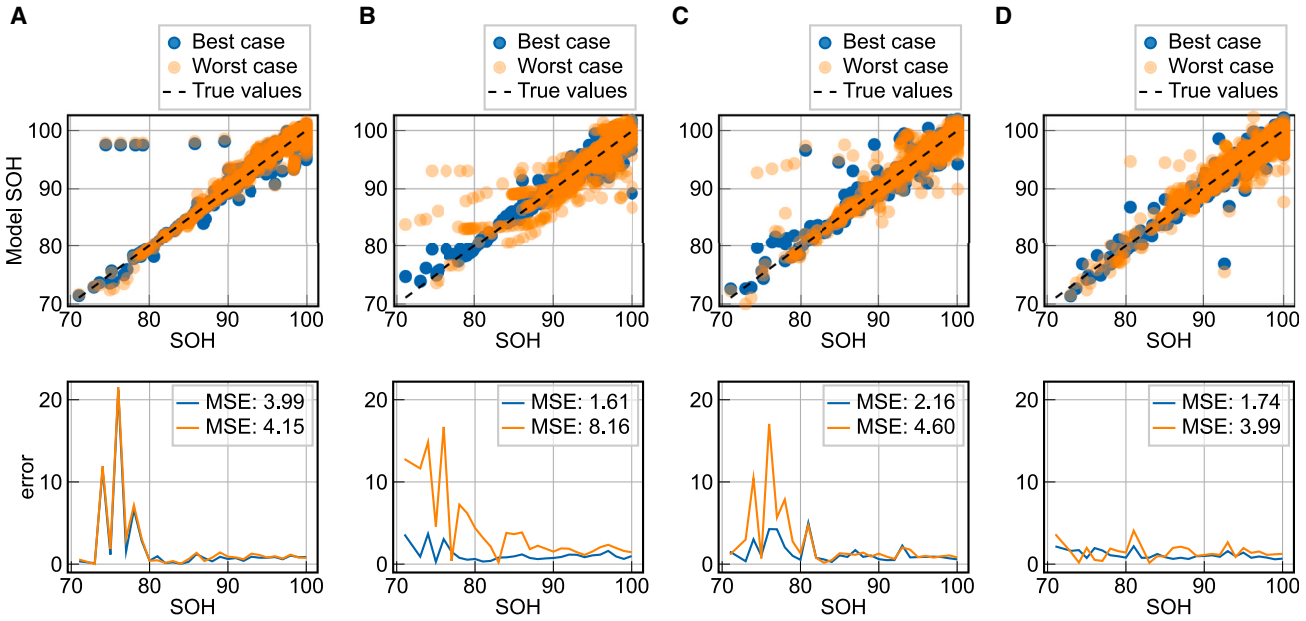


Figure 8. Feature extraction framework validation with best and worst test MSEs of five training runs

(A) Network estimation using a basic correlation filter directly on the histogram variables.

(B) Network estimation using a basic correlation filter on the functional fitting parameters.

(C) Network estimation using the improved correlation filter directly on the histogram variables.

(D) Network estimation using the improved correlation filter on the functional fitting parameters.

controlled aging test conditions. On the other hand, we tap into the richness of real-world operational modes extracted from the driving behaviors of a substantial cohort of 600,000 customers. Hence, our strategy exploits laboratory data to identify relevant aging features while concurrently harnessing customer behavior correlations inherent in field data to cluster akin features and mitigate redundancies in the input information. In achieving this synthesis, we employ the transformation of measured time-series data from laboratory tests into the statistical data format characteristic of field data readouts. The heart of our proposed feature extraction framework hinges on a correlation-based feature-grouping algorithm complemented by the PLS technique to condense the statistical essence of akin features concerning SOH. Remarkably, the application of PLS to each feature group yields elevated aging correlations for compressed features compared with the original feature group. Furthermore, we introduce an innovative data augmentation strategy, employing elementary pV functions to aptly characterize the authentic underlying distribution of histogram variables. The final framework decisively enhances model accuracy by 57% compared with a basic correlation filter. With the rising importance of field data analysis and high demand for customer-oriented component design, we firmly believe that our framework not only adeptly tackles estimation challenges posed by statistical field data but also bridges the gap between field and laboratory data in the realm of battery aging regression.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and materials should be directed to and will be fulfilled by the lead contact, Valentin Steininger (valentin.steininger@bmw.de).

Materials availability

This study did not generate any unique materials.

Data and code availability

The laboratory and field data are kept confidential to protect customer-related sensitive data. All original code has been deposited at Zenodo under https://doi.org/10.5281/zenodo.8228280 and is publicly available as of the date of publication.

Spearman correlation

After data preparation, a correlation analysis is used to investigate the influence of input features on the SOH and also the collinearity between the features themselves to avoid redundancies among inputs. Therefore, we compute the Spearman rank correlation coefficient $r_{x,y}$ to quantify the monotonic dependency of features x and y as:

$$r_{x,y} = \frac{cov(R(x), R(y))}{\sigma_{R(x)} \cdot \sigma_{R(y)}}, \quad \text{(Equation 11)}$$

where $cov(\cdot)$ denotes the covariance function and σ the standard deviation of the rank variables $R(x)$ and $R(y)$.[20] The empirical covariance function of two variables x and y with sample size n is defined as:

$$cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu_x) \cdot (y_i - \mu_y), \quad \text{(Equation 12)}$$

with μ as the sample mean of the respective variable. From an illustrative point of view, the covariance expresses whether two variables vary in the same direction from their mean values for all samples i. Now, the correlation coefficient gets normalized using the product of both sample variances and evaluates to 1 only for perfectly correlated variables where the covariance equals the product of the sample variances. With that, the correlation coefficient represents the degree of monotonic relationship between all collected values of two variables. For n features of a dataset, the correlation matrix P contains the correlation coefficients from all pairwise feature combinations:

$$\mathrm{P} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{bmatrix}. \quad \text{(Equation 13)}$$

Considering the fact that, e.g., some batteries age faster due to higher loads or more extreme temperatures than others, a monotonic dependency over time gets distorted as a consequence of different aging rates. For example, Figure 2 shows the aging degradation over time of the laboratory data and also the relationship between the SOH and time. Despite a clear dependency for each individual cell, the overall correlation evaluates to merely 0.6. Hence, the relationship would not be adequately quantified. For that reason, we determine the coefficients for each battery separately and then calculate the mean over all m resulting correlation matrices:

$$\mathrm{P} = \frac{1}{m} \sum_{i=1}^{m} \mathrm{P}_i. \quad \text{(Equation 14)}$$

Partial least squares (PLS)

PLS is a statistical method designed to deal with the challenges of multicollinearity and feature reduction for regression problems. With regard to feature reduction, the more popular approach of principal-component analysis (PCA) is applied to map

high-dimensional data into the low-dimensional space through linear transformation and maximize the variance of the new data during the projection to preserve as much of the original data as possible.[21] Hence, it is an unsupervised technique that does not consider the predictability of the target variable using the transformed features. Consequently, in most cases, PCA is not an appropriate feature reduction method for regression problems, particularly when input features with a low variance correlate highly to the target variable.[22] As a remedy, PLS transforms the input features by maximizing the covariance between the target variable and the new features. In the case of a one-dimensional target variable, the underlying model of PLS is defined as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}, \quad \text{(Equation 15)}$$

where $\mathbf{X}$ is the $(n \times m)$ feature matrix, $\mathbf{T}$ the $(n \times l)$ projection of $\mathbf{X}$ for l transformed features, $\mathbf{P}$ the $(m \times l)$ so-called loading matrix containing the weights for the linear transform of $\mathbf{X}$, and $\mathbf{E}$ the error matrix. To compute the weights in the loading matrix $\mathbf{P}$, different algorithms exist, which all aim to maximize the covariance between the target variable $\mathbf{Y}$ and the feature transform $\mathbf{T}$.[23] In this work, we group input features that have a collinearity of greater than 0.9 and apply the PLS method to obtain one resulting input feature. To transform the grouped feature matrix $\mathbf{F}_g$ into one feature vector $\mathbf{f}_g$, we compute:

$$\mathbf{f}_g = \mathbf{F}_g\mathbf{p}, \quad \text{(Equation 16)}$$

with $\mathbf{p}$ as the $(m \times 1)$ loading matrix computed from the feature group. With regard to aging estimation, we apply the following algorithm to structure our input features:

(1) Select the feature with the highest absolute Spearman correlation coefficient (>0.8) to the SOH from the laboratory correlation matrix.
(2) Allocate remaining features that have strong collinearity (>0.9) to the aging feature from the field correlation matrix.
(3) Compress the feature group using the PLS method.
(4) Repeat for all remaining input features.

With this algorithm, we reduce the number of collected input features from customer data on one hand and create a stable, uncorrelated regression input feature set on the other. Since we take correlations from the laboratory data to identify feature correlations to the SOH from accurate measurements and also correlations from customer behavior in the field data to detect collinearity among input features, we combine advantages from both data sources to construct features from the input data. Moreover, since PLS accounts for the regression problem, we compress the most relevant information from redundant feature groups without discarding important data for the aging estimation process.

Functional fitting

To describe the underlying distribution of histogram variables, we apply a functional fitting algorithm using the superposition of elementary distribution functions. Therefore, we use the pV function, a convolution of the Gaussian function $G(x)$ and Lorentzian function $L(x)$ with a weighting factor α given as follows:

$$pV(x) = (1 - \alpha)\cdot G(x) + \alpha \cdot L(x). \quad \text{(Equation 17)}$$

Both functions share three parameters, comprising the amplitude A, the mean μ, and the standard deviation σ. The equation of the pV is thus defined as:

$$pV(x, A, \mu, \sigma, \alpha) = \frac{(1-\alpha)\cdot A}{\sigma\cdot\sqrt{2\pi}}\cdot e^{\frac{-(x-\mu)^2}{2\cdot\sigma^2}} + \frac{\alpha\cdot A}{\pi}\cdot\frac{\sigma}{(x-\mu)^2+\sigma^2}. \quad \text{(Equation 18)}$$

Figure 3 shows that histogram variable distributions can have multiple local maxima. Hence, the number of pV curves required to fit a histogram-based functional curve equals the number of local peaks detected in the histogram curve. To detect peaks in the functional curve, the gradient is evaluated and checked for a plateau followed by a strict fall, indicating a local peak in the curve. With that, the final functional curve to fit the histogram distribution with n peaks is defined as follows:

$$f_{fit} = \sum_{i=1}^{n} pV(x, A_i, \mu_i, \sigma_i, \alpha_i). \quad \text{(Equation 19)}$$

The typical approach for performing the fitting operation to determine the $4n$ parameters is through the least-squares method, where the curve-fit solution of overdetermined systems is approximated by minimizing the sum of the squares of the residuals for every equation. In this work, functional components must be under certain constraints on mean and amplitude coefficients to avoid overshooting or overlapping between curve components. As a result, this work uses the trust region reflective (TRR) method to minimize the least-square error in successive iterations to achieve the best quality fit.

Neural networks

Neural networks are data-based models that have been originally inspired by the human brain, specifically the input-output structure of their smallest entity, the neuron. The network training, an iterative optimization algorithm where the neuron parameters (weights) are adapted based on the estimation error, resembles the process of learning from experience. The layers of a network serve as the topologic units, and their structure significantly determines the level of abstraction and complexity the network is able to provide. Therefore, the input layer takes the preprocessed data and forwards it to the hidden layers whose neurons apply a transformation function on the layer inputs. Mathematically, for a neuron $j \in \{0, 1, \ldots, J\}$ with n inputs in a hidden layer $l \in \{0, 1, \ldots, L\}$, the transformation behind the propagation from one unit to another can be formulated as:

$$z_{j,l} = \sum_{i=1}^{n} w_{j,l}^{(i)} \cdot h_{j,l-1}^{(i)} + b_{j,l}, \quad \text{(Equation 20)}$$

$$h_{j,l} = a_l\left(z_{j,l}\right), \quad \text{(Equation 21)}$$

where w, b, and a are the weight of the neuron, the bias factor, and the activation function, respectively. The sensitivity of the neuron toward an input is determined by the value of the respective weight $w^{(i)}$. To ensure universal function approximation of the network, a nonlinear activation function retransforms the output $z_{j,l}$ of the linear combination. The output $h_{j,l}$ is then passed to the neurons of the next layer, which again apply the transformation using their weights and bias values. In this work, we use a fully connected network, where each neuron of a layer is connected to all neurons of the next layer and a rectified linear unit (relu) as the activation function. The model architecture consists of three hidden layers with 16, 64, and 1 neuron, respectively.

For model training and validation, we compute the MSE to quantify the deviation between model estimation $\hat{y}$ and actual values y as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad \text{(Equation 22)}$$

where y_i represents an SOH value from the laboratory data and $\hat{y}_i$ an estimation from the network given a respective feature extraction method.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrp.2023.101596.

ACKNOWLEDGMENTS

This work was financially and technically supported by BMW AG. Part of the work was supported by the research project "COBALT-P" (16BZF314C), funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK).

AUTHOR CONTRIBUTIONS

Conceptualization, V.S., K.R., P.H., and W.L.; methodology, V.S. and W.L.; investigation, V.S., K.R., P.H., W.L., and D.U.S.; writing – original draft, V.S.; writing – review & editing, V.S., K.R., P.H., W.L., and D.U.S; supervision, W.L. and D.U.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 2, 2023
Revised: June 23, 2023
Accepted: September 4, 2023
Published: September 22, 2023

REFERENCES

1. Masias, A., Marcicki, J., and Paxton, W.A. (2021). Opportunities and Challenges of Lithium Ion Batteries in Automotive Applications. ACS Energy Lett. *6*, 621–630. https://doi.org/10.1021/acsenergylett.0c02584.

2. Börner, M.F., Frieges, M.H., Späth, B., Spütz, K., Heimes, H.H., Sauer, D.U., and Li, W. (2022). Challenges of second-life concepts for retired electric vehicle batteries. Cell Reports Physical Science *3*, 101095. https://doi.org/10.1016/j.xcrp.2022.101095.

3. Ren, H., Zhao, Y., Chen, S., and Wang, T. (2019). Design and implementation of a battery management system with active charge balance based on the SOC and SOH online estimation. Energy *166*, 908–917. https://doi.org/10.1016/j.energy.2018.10.133.

4. Kim, S., Yi, Z., Kunz, M.R., Dufek, E.J., Tanim, T.R., Chen, B.-R., and Gering, K.L. (2022). Accelerated battery life predictions through synergistic combination of physics-based models and machine learning. Cell Reports Physical Science *3*, 101023. https://doi.org/10.1016/j.xcrp.2022.101023.

5. Topan, P.A., Ramadan, M.N., Fathoni, G., Cahyadi, A.I., and Wahyunggoro, O. (2016). State of Charge (SOC) and State of Health (SOH) estimation on lithium polymer battery via Kalman filter. In International Conference on Science and Technology-Computer (ICST), 2nd, ed., pp. 93–96. https://doi.org/10.1109/ICSTC.2016.7877354.

6. Kumar, B., Khare, N., and Chaturvedi, P.K. (2018). FPGA-based design of advanced BMS implementing SoC/SoH estimators. Microelectron. Reliab. *84*, 66–74. https://doi.org/10.1016/j.microrel. 2018.03.015.

7. Zhang, Y., Wik, T., Bergström, J., Pecht, M., and Zou, C. (2022). A machine learning-based framework for online prediction of battery ageing trajectory and lifetime using histogram data. J. Power Sources *526*, 231110. https://doi.org/10.1016/j.jpowsour.2022.231110.

8. Aitio, A., and Howey, D.A. (2021). Predicting battery end of life from solar off-grid system field data using machine learning. Joule *5*, 3204–3220. https://doi.org/10.1016/j.joule.2021.11.006.

9. Greenbank, S., and Howey, D. (2022). Automated Feature Extraction and Selection for Data- Driven Models of Rapid Battery Capacity Fade and End of Life. IEEE Trans. Ind. Inf. *18*, 2965–2973. https://doi.org/10.1109/TII.2021.3106593.

10. Paulson, N.H., Kubal, J., Ward, L., Saxena, S., Lu, W., and Babinec, S.J. (2022). Feature engineering for machine learning enabled early prediction of battery lifetime. J. Power Sources *527*, 231127. https://doi.org/10.1016/j.jpowsour.2022.231127.

11. Geringer, D., Hofmann, P., Girard, J., Trunner, E., and Knefel, W. (2021). Aging investigations and consideration for automotive high power lithium-ion batteries in a 48 V mild hybrid operating strategy. Automot. Engine Technol. *6*, 219–234.

12. Lain, M.J., Brandon, J., and Kendrick, E. (2019). Design Strategies for High Power vs. High Energy Lithium Ion Cells. Batteries *5*, 64. https://doi.org/10.3390/batteries5040064.

13. 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.

14. Zhou, L., He, L., Zheng, Y., Lai, X., Ouyang, M., and Lu, L. (2020). Massive battery pack data compression and reconstruction using a frequency division model in battery management systems. J. Energy Storage *28*, 101252. https://doi.org/10.1016/j.est.2020.101252.

15. Steininger, V., Hüsson, P., Rumpf, K., and Sauer, D.U. (2023). Customer-centric aging simulation for 48 V lithium-ion batteries in

vehicle applications. eTransportation *16*, 100240. https://doi.org/10.1016/j.etran.2023.100240.

16. Bank, T., Klamor, S., and Sauer, D.U. (2020). Lithium-ion cell requirements in a real-world 48 V system and implications for an extensive aging analysis. J. Energy Storage *30*, 101465. https://doi.org/10.1016/j.est.2020.101465.

17. Bank, T., Feldmann, J., Klamor, S., Bihn, S., and Sauer, D.U. (2020). Extensive aging analysis of high-power lithium titanate oxide batteries: Impact of the passive electrode effect. J. Power Sources *473*, 228566. https://doi.org/10.1016/j.jpowsour.2020.228566.

18. Richardson, R.R., Osborne, M.A., and Howey, D.A. (2019). Battery health prediction under generalized conditions using a Gaussian process transition model. J. Energy Storage *23*, 320–328. https://doi.org/10.1016/j.est.2019.03.022.

19. Hamar, J.C., Erhard, S.V., Canesso, A., Kohlschmidt, J., Olivain, N., and Jossen, A. (2021). State-of-health estimation using a neural network trained on vehicle data. J. Power Sources *512*, 230493. https://doi.org/10.1016/j.jpowsour.2021.230493.

20. Makowski, D., Ben-Shachar, M., Patil, I., and Lüdecke, D. (2020). Methods and Algorithms for Correlation Analysis in R. J. Open Source Softw. *5*, 2306. https://doi.org/10.21105/joss.02306.

21. Xing, J., Zhang, H., and Zhang, J. (2023). Remaining useful life prediction of – Lithium batteries based on principal component analysis and improved Gaussian process regression. Int. J. Electrochem. Sci. *18*, 100048. https://doi.org/10.1016/j.ijoes.2023.100048.

22. Liu, C., Zhang, X., Nguyen, T.T., Liu, J., Wu, T., Lee, E., and Tu, X.M. (2022). Partial least squares regression and principal component analysis: similarity and differences between two popular variable reduction approaches. Gen. Psychiatr. *35*, e100662. https://doi.org/10.1136/gpsych-2021-100662.

23. Kvalheim, O.M. (2010). Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots. J. Chemometr. *24*, 496–504. https://doi.org/10.1002/cem.1289.