Energy and AI 22 (2025) 100575

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Semi-supervised battery state of health estimation for field applications

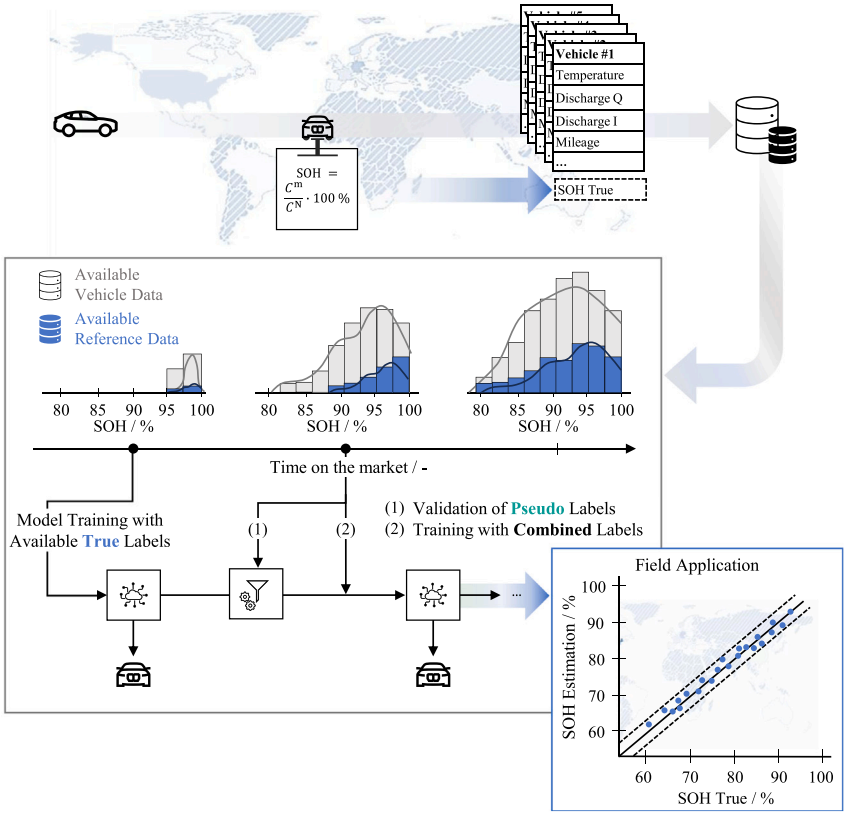
Nejira Hadzalic [a,b,c,*], Jacob Hamar [c], Marco Fischer [b], Simon Erhard [c], Jan Philipp Schmidt [a]

[a] Chair of Systems Engineering for Electrical Energy Storage (SysEE), University of Bayreuth, Bavarian Center for Battery Technology (BayBatt), Universitätsstraße 30, Bayreuth, 95447, Bavaria, Germany
[b] Chair of Electrical Energy Storage Technology, Technical University of Munich, Karlstraße 45, Munich, 80333, Bavaria, Germany
[c] BMW Group, Petuelring 130, Munich, 80809, Bavaria, Germany

HIGHLIGHTS

- Real-world operational complexities can be captured more effectively by integrating both labeled and unlabeled data.
- Semi-supervised outperforms supervised methods in dynamic field conditions.
- The proposed framework employs multi-view learning with a confidence-driven pseudo-labeling strategy.
- Validation data consists of 3000 EVs with standardized capacity measurements.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Lithium-ion battery

ABSTRACT

Battery electric vehicles are exposed to highly diverse operating conditions and driving behaviors that strongly influence degradation pathways, yet these real-world complexities are only partially captured in laboratory

* Corresponding author at: Chair of Systems Engineering for Electrical Energy Storage (SysEE), University of Bayreuth, Bavarian Center for Battery Technology (BayBatt), Universitätsstraße 30, Bayreuth, 95447, Bavaria, Germany.
E-mail addresses: nejira.hadzalic@uni-bayreuth.de, nejira.hadzalic@bmw.de (N. Hadzalic).

https://doi.org/10.1016/j.egyai.2025.100575
Received 7 March 2025; Received in revised form 24 June 2025; Accepted 26 July 2025
Available online 21 August 2025

State of health estimation
Semi-supervised learning
Field data
Machine learning

aging tests. This study investigates a semi-supervised learning approach for robust estimation of battery state of health, defined as the ratio of remaining to nominal capacity. The method integrates a multi-view co-training algorithm with a rule-based pseudo labeling mechanism and is developed and validated using field data from 3000 BMW i3 vehicles with battery capacity of 60 Ah, collected since 2013 across 34 countries. The available data comprises standardized full charge capacity measurements, which serve as ground truth labels. The proposed training and validation pipeline is designed to address challenges inherent in real-world data generation and is particularly advantageous during early deployment of new battery technologies, when labeled data is scarce. By incrementally incorporating newly available labeled data into both evaluation and retraining, the model adapts to heterogeneous aging patterns observed in the field. Comparative analysis demonstrates that, relative to a supervised benchmark, the proposed method reduces estimation error by 28 % under limited-label conditions and by 6 % under optimally labeled scenarios, highlighting its robustness for field applications.

1. Introduction

In recent years, advancements in lithium-ion battery technology have significantly improved the safety, longevity, and performance of electric vehicles (EVs). Yet, accurately monitoring and diagnosing battery aging in the field remains a critical challenge [1]. The main difficulties arise from the inherently complex and nonlinear nature of battery degradation processes, further compounded by the highly variable loads, irregular charge–discharge patterns, and fluctuating ambient environments in which EVs operate [2,3]. In addition, measuring key state of health (SOH) indicators, such as capacity fade and resistance growth, in the field is often impractical, limiting the ability to map operational variability directly to the underlying aging behavior [4].

SOH estimation techniques fall into two broad categories: model-based and data-driven methods [5]. Model-based methods rely on detailed electrochemical or mathematical representations of cell behavior, providing interpretability at the cost of extensive parameterization and limited adaptability to new chemistries [6]. In contrast, data-driven methods infer SOH by identifying intricate dependencies and complex relationships in operational data, such as voltage, current, temperature, and usage logs, without explicit physicochemical assumptions [7]. To date, most data-driven studies rely on well-annotated, laboratory-derived datasets, representing a limited number of cells aged under tightly controlled conditions. While these datasets provide valuable insights, they fail to fully represent the wide range of degradation mechanisms and operational variability encountered by EVs in real-world settings [8–11]. Consequently, resulting models are seldom applicable to large-scale field data, where labels are scarce, sensor signals are noisy, onboard computing resources are limited, and often only sensor signals of groups of connected cells are available [12–14]. As an example, Pan et al. [8] estimate capacity-based SOH from ohmic and polarization resistance measured in three commercial nickel manganese cobalt (NMC) cells with a nominal capacity of 2 Ah. The cells are operated at 25 °C until a capacity loss of 20 % with constant-current constant-voltage (CCCV) charging at 0.5C and a discharge protocol that includes constant-current (CC) discharge, driving cycle, and repeated pulses. Xu et al. [9] propose a capacity-based SOH prediction model employing a convolutional neural network (CNN) and a long short-term memory (LSTM) architecture, trained on the NASA and Oxford open-source datasets. The NASA dataset contains cycling data from four 2 Ah lithium-ion cells aged at 24 °C under CCCV charging at 0.75C and CC discharging at 1C with three different cutoff voltages, until a 30 % capacity loss [15]. The Oxford dataset provides aging histories of eight 740 mAh lithium-ion cells subjected to CCCV charging at 2C and an ambient temperature of 40 °C, followed by repeated drive cycle discharges [16]. Across three experimental settings utilizing 60 %, 70 %, and 80 % of the available data, the model shows a threefold performance drop under the lowest data availability, underscoring its limited robustness to aging patterns insufficiently represented in the training data. Roman et al. [17] demonstrate the effectiveness of an ensemble learning pipeline for SOH estimation by combining multiple open-source laboratory datasets. The training data comprises a total of 179 cells, including 33 cells from the NASA datasets [15], 8 cells from the Oxford dataset [16], 14 cells from the CALCE datasets [18], and 124 cells from the MIT dataset [19]. All datasets characterize cell-level degradation under controlled experimental conditions. The CALCE dataset documents the aging of 2 Ah lithium-ion cells cycled at 25 °C combining CCCV charge at 0.5C with CC discharge at 0.5C and 1C, until a 20 % capacity fade [18]. The MIT dataset captures cyclic aging of 2 Ah lithium-ion cells subjected to various fast-charging regimes with constant discharge at 4C [20]. The ensemble approach achieves RMSE values between 0.1 % and 5.6 % across different charge protocols and cell populations, demonstrating that generalization challenges persist even with high-quality laboratory datasets characterized by low measurement noise and complete charging profiles.

To reduce dependence on exhaustive full-cycle data and large volumes of labeled data, recent works explore transfer learning (TL) approaches to generalize knowledge across chemistries and operating regimes, as well as semi-supervised learning (SSL) frameworks that exploit the abundance of unlabeled measurements [21–23]. Shen et al. [24] introduce a TL-ensemble approach with cycling data from eight cells in the source domain, and 20 cells from the NASA dataset in the target domain. The method achieves RMSE of 1.5 %, significantly outperforming other supervised learning (SL) baselines. However, effective adaptation requires the source domain to sufficiently reflect conditions encountered in the target domain [5,21]. Fan et al. [25] propose a shared-encoder architecture that jointly optimizes self-supervised and supervised objectives on the Oxford and MIT datasets [16,20]. A comparative analysis of models trained with sufficient and insufficient labels and validated on data of five battery cells shows fairly constant RMSE values at approximately 0.5 % for both approaches when there are sufficient training samples. In scenarios with limited labels, RMSE values of the SSL method range from 0.6 % to 1.0 %, while the RMSE of the SL method ranges from 0.9 % to 1.4 %. Despite the success in reducing data requirements, the generalization capability remains uncertain on validation sets consisting of a few battery cells aged under constant conditions.

Given these challenges, field-sourced data is increasingly recognized as essential for capturing the full complexity and variability of battery degradation [5,7,26–29]. Although modern EVs generate large volumes of low-cost, unlabeled operational data, conventional data-driven approaches cannot fully capitalize on this resource, primarily due to the lack of reliable ground truth reference labels [30]. Accurate assessment of a battery's true aging state mandates comprehensive diagnostic procedures, which are resource-intensive and impractical to perform at scale [4]. Existing field-based studies generally rely on small vehicle fleets observed over limited time periods and often assume uniform data fidelity. For example, Zhao et al. [7] present a framework for pack-level SOH estimation using one year of field data. The proposed model extracts internal health features from partial charging segments and external features, such as ambient temperature and accumulated mileage. An extreme learning machine combined with particle swarm optimization is used to predict capacity-based SOH, achieving RMSE of 0.13 %, but only for two vehicle samples. Furthermore, training is based

on reference labels derived from charging events with a depth of charge exceeding 40 %, which may introduce estimation bias. Given the narrow evaluation scope, questions remain about the model's scalability and robustness when applied to extended aging horizons, noisy and erratic charging segments, or different field clusters. Similarly, Qi et al. [26] estimate battery SOH using field data from two EVs, with one vehicle used for training and the other for testing. Reference labels are inferred from selected charging segments using an inverse form of Coulomb counting enhanced by OCV- and resistance-based corrections. While this method improves estimation accuracy in a constrained setting, its reliance on carefully curated data segments, validation on a minimal dataset, and high computational complexity limit its scalability and applicability to broader vehicle populations. Wang et al. [5] propose a self-supervised pipeline that utilizes two years of unlabeled charging data from 20 EVs to extract health indicators from short windows centered around the main peak of the incremental capacity curve for battery SOH prediction. Similar to prior works [7,26], reference labels are extracted from charging segments, however, to mitigate label uncertainty, monthly SOH references are computed by averaging multiple estimated capacity values. While this approach marks a meaningful step forward, its generalization capability and robustness remain to be validated, particularly under realistic conditions involving sensor noise, irregular or truncated charging events, and the computational overhead associated with feature extraction, model pre-training and deployment.

To bridge the gap between real-world data constraints and the growing need for accurate battery SOH estimation, this study proposes a scalable, field-adapted framework grounded in a semi-supervised paradigm. The approach is explicitly designed to accommodate the challenges of field data, including limited data availability, sparse and temporally delayed ground truth labels, and the dynamic nature of battery aging. In contrast to methods that require densely sampled, high-throughput charging or discharging profiles, the framework is optimized for computational efficiency and is suitable for deployment in resource-constrained onboard environments. It demonstrates robust performance across application-relevant scenarios, particularly during the early deployment phases of new battery technologies, where labeled data is scarce and traditional supervised methods typically fail.

The method leverages a rule-based, multi-view SSL architecture with pseudo labeling to extract insights from unlabeled operational data, enabling SOH estimation under even extreme label scarcity. It is developed using field data from BMW's first fully electric vehicle, the BMW i3, collected globally since 2013 under diverse environmental and operational conditions. The aging history of 3000 vehicles is complemented by standardized full charge capacity measurements, widely regarded as the most reliable proxy for the true EV battery SOH. These measurements, performed under controlled conditions that include temperature conditioning and a charge–discharge procedure at a low current rate, enable robust mapping between operational variability and actual aging states, and serve as high-quality reference labels. Empirical results from two case studies, focusing on battery age and data availability, demonstrate that the proposed method outperforms SL baselines by up to three percentage points in RMSE and significantly reduces the incidence of tolerance band violations, underscoring its potential for scalable and adaptive SOH estimation. More broadly, the framework is well-suited for all application domains in which the acquisition of ground truth labels is costly, labor-intensive, or logistically impractical, yet large volumes of unlabeled data are readily available.

2. Theory

The proposed SSL method for battery SOH estimation in field applications employs a co-training framework with a self-supervised pseudo labeling mechanism. This section outlines the fundamental principles of SSL and describes the core components of the developed framework.

2.1. Fundamentals of semi-supervised learning methods

SSL occupies an intermediate position between supervised and unsupervised paradigms, leveraging both labeled and unlabeled data to enhance predictive performance and reduce labeling costs [31]. Let $L = \{(x_i, y_i)\}_{i=1}^{l}$ be the labeled dataset and $U = \{x_j\}_{j=l+1}^{l+u}$ be the unlabeled dataset, where x denotes the input feature vector and y the corresponding labels, l the number of labeled samples, and u the number of unlabeled samples [31]. Unlike purely supervised methods, which rely exclusively on the labeled dataset L, and unsupervised methods, which rely exclusively on the unlabeled samples in U, semi-supervised methods integrate ground truth information from L with structural patterns in U. By simultaneously leveraging information from both L and U during training, SSL aims to construct a better predictor and achieve a lower generalization error than would be possible using only labeled data L, especially when $l \ll u$ [32–35].

The effective use of unlabeled data in SSL relies on three interrelated assumptions: smoothness, cluster, and manifold. These assumptions constrain the data structure in ways that allow unlabeled instances to inform the model and improve learning [31,34]. Under the smoothness assumption, the target function varies continuously with the input, so points close in feature space are expected to have similar labels or function outputs in regression settings [34]. The cluster assumption posits that the data distribution decomposes into distinct, high-density clusters sharing the same label. Accordingly, decision boundaries, or large changes in the target function, should lie in regions of low data density between clusters [35]. The manifold assumption asserts that high-dimensional observations concentrate near a lower-dimensional manifold embedded in the ambient space, and that points connected along this manifold tend to share a label [32].

Common SSL approaches can be categorized as self-training, co-training, graph-based, and generative models [34]. Self-training generates pseudo labels for unlabeled data by training a base model on a small labeled dataset and including high-confidence predictions, pseudo labels, in the following iterations to progressively improve model performance. Key design choices include pseudo labeling selection criteria, strategy for incorporating pseudo labels in subsequent iterations, and stopping criteria. This approach is particularly effective when the model produces high-confidence predictions early in training, enabling incremental gains without explicit supervision [31,36].

Co-training extends self-training by employing two or more supervised models trained on complementary views of the data or heterogeneous algorithms. Each model generates pseudo labels for unlabeled data and shares them with others for retraining. This approach assumes model diversity, either in architectures, leading to diverse decision boundaries, or in input features, to reduce the risk of reinforcing the same errors. By exchanging pseudo labels, the models collaborate to correct mistakes, leading to improved performance over iterations. Early co-training implementations relied on strictly distinct views of the data, typically corresponding to disjoint subsets of features. However, in real-world scenarios, truly independent views of the data are rarely available. Recent variants encourage diversity through heterogeneous learning algorithms rather than strictly disjoint feature sets [37,38].

Graph-based methods represent both labeled and unlabeled data samples as nodes in a graph, with edges connecting similar data points based on some distance metric or kernel [31]. Label information propagates through the graph using algorithms such as label propagation or graph neural networks, which learn representations encoding both features and graph structure. This approach excels in domains with inherent relational structures, though constructing an accurate graph can be computationally expensive and may require domain-specific definitions of similarity metrics [39].

Generative methods assume a joint distribution in which each class generates data from a mixture model. Unlabeled samples help estimate the input distribution, revealing the structure of data clusters so that a

few labeled samples suffice to assign cluster labels. In essence, the unlabeled data helps the model to expand its understanding of where the high-density regions are [40,41]. Hybrid approaches integrate multiple SSL techniques to leverage individual model strengths and are often employed in real-world settings with diverse, multimodal data [37].

In the battery domain, SSL offers a promising pathway to address the mismatch between the demand for labeled data in SL and the abundance of unlabeled operational data available in real-world battery systems, provided that the model robustness is maintained. By exploiting this unlabeled data, SSL can accelerate the deployment of data-driven methods in practice, where generating labeled samples is both time-consuming and costly. Regardless of the chosen SSL technique, however, establishing a reliable supervised baseline remains essential for quantifying the true benefits of incorporating unlabeled data [31].

2.2. Pseudo labeling techniques

Pseudo labeling describes the process of assigning labels to previously unlabeled data based on model predictions in order to augment the training set. Pseudo labeling techniques are often based on statistical principles and provide probabilistic estimates or predictions with associated confidence intervals, but they cannot provide a formal guarantee of correctness because they are inherently dependent on the quality and reliability of the model's initial predictions and the underlying data distribution [42]. Confidence-based, entropy-based, consistency-based, and temporal pseudo labeling represent widely used pseudo labeling techniques [37,43].

2.3. Performance metrics

The accuracy of the developed methods in estimating SOH is evaluated using the root mean squared error (RMSE) and mean absolute error (MAE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{1}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{2}$$

where y_i represents the true value and $\hat{y}_i$ the estimated value of the prediction model.

In addition, a $\pm 5\,\%$ tolerance band is introduced as an auxiliary performance metric to reflect prospective government regulations mandating accuracy monitoring of all metrics displayed in vehicles, analogous to requirements for certified energy and certified range [44–46]. Following the definition,

$$\mathrm{Tolerance}_{5\%} = \frac{\left|\mathrm{abs}\left(y_i - \hat{y}_i\right) \geq 5\%\right|}{N} \cdot 100\% \tag{3}$$

quantifies the percentage of samples violating the $\pm 5\,\%$ accuracy requirement, where the numerator counts the instances exceeding the tolerance and N denotes the total number of samples.

3. Materials and method

To enable effective and scalable implementation of data-driven techniques, it is essential to exploit the full potential of the available data, including its distributional characteristics. The following section details the design and implementation of the proposed SSL approach.

3.1. Data

This study is based on empirical data collected from the BMW i3 fleet, equipped with NMC batteries with a nominal capacity 60 Ah, delivering a range of 130–160 km and an energy consumption of 0.20–0.25 kWh/km under city and highway test conditions, respectively [47]. The dataset predominantly comprises unlabeled field data from approximately 55 000 vehicles operated in 34 countries since 2013.

A vehicle anonymized number (VAN) facilitates the mapping of unique measurements collected, on average, on a weekly basis. The dataset includes meta-information such as vehicle age, mileage, and average ambient temperature, alongside measurements collected by high-voltage battery system sensors and processed within the electronic control unit. These include variables such as charge throughput, C-rates, cell temperatures, and cell voltages.

Several variables are provided in the form of binned histograms, with records of the cumulative time in each bin. Feature extraction using binned histograms involves discretizing time-series data into predefined intervals tailored to the characteristics of each variable. For each bin, the duration of battery operation is accumulated, and a time-weighted average across the distribution is computed, yielding a feature that represents the vehicle's average operating condition up to the full-charge reference measurement.

The variables subject to binning are C-rate, SOC, and ambient temperature. C-rate is derived from current measurements normalized by the nominal capacity and grouped into five bins for both charge and discharge directions. SOC is binned in 5 % increments. Although the SOC estimation method employed here is proprietary, it is conceptually aligned with standard industry approaches such as OCV-based lookup tables and Coulomb counting. Ambient temperature is similarly discretized into five intervals to capture the vehicle's thermal exposure. As an example, a vehicle with an average ambient temperature of 10 °C has been operated at both lower and higher temperatures, but the value of 10 °C reflects the operating average over the entire vehicle lifetime, extending up to the measurement date.

Other variables used to develop the model include age, total mileage, and charge throughput, and they are recorded as a cumulative amount up to the time of measurement. Additional variables such as minimum, maximum, and mean cell voltage and cell temperature represent readings of the cells with the lowest and highest values in the pack, and the mean across all cells at readout.

Further insight into the data is provided by Fig. 1 and Table 1. Fig. 1 shows the distribution of the collected data, describing the average conditions observed in the field. The vehicle samples cover a wide range of conditions, including vehicles up to ten years old, with an average vehicle age of six years. The average total mileage driven is 68 404 km, but vehicles with mileage greater than 135 000 km – where the value given represents the 95th percentile and the maximum recorded mileage is 256 000 km – are also included in the dataset. The average discharge C-rate ranges between 0 and 1.07 h^{-1}, with an average of 0.52 h^{-1}. The charge C-rate is less spread out compared to the discharge counterpart and is centered around 0.32 h^{-1}, which is most likely dominated by standard charging procedures. The average ambient temperature at which vehicles operate ranges from 7 to 31 °C.

In addition to the data sent regularly from the vehicle, some vehicles undergo a full charge reference measurement. The standardized full charge reference measurement uses a low current discharge and charge regime at a constant temperature to ensure accurate and consistent results for all measurements, as introduced in the works of Hamar et al. and Hofmann et al. [4,48]. The test is performed at the request of the customer and there is no correlation between the vehicles with a full charge reference measurement and any specific vehicle behavior. In addition, the measurements are completely independent of the age of the vehicle. The measured capacity, C^m, is compared to the nominal

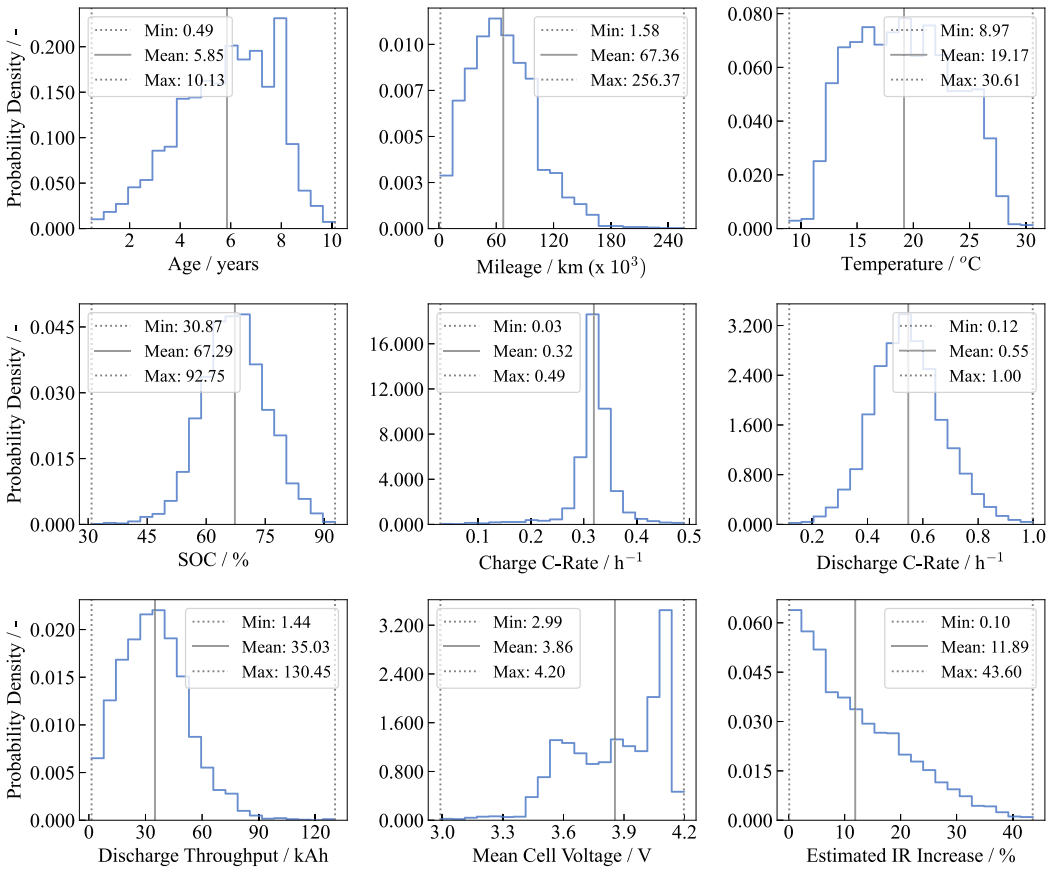


Fig. 1. Distribution of variables collected from the BMW i3 fleet relevant for this study. The variables include vehicle age, total mileage up to the time of measurement, average ambient temperature, average discharge depth, average charge and discharge C-rates, average discharge charge throughput, and estimated average rate of increase in internal resistance. The range of variables represents the variety of average environmental and operating conditions.

capacity, C^N, of the battery at the beginning of its life according to the following equation:

$$\text{SOH} = \frac{C^{\text{m}}}{C^{\text{N}}} \tag{4}$$

The resulting SOH value represents the most reliable proxy for the true state of EV batteries that have been exposed to various environmental and driving conditions, including aging patterns that may not have been anticipated during battery testing and development [4]. Thus, the available data is considered a ground truth reference from the field and provides a unique opportunity to explore the potential of data-driven approaches to battery state estimation under field conditions. Of all available vehicle samples in the BMW i3 fleet, over 3000 are vehicles with existing reference labels, i.e. full charge reference measurements. The dataset in this study is limited to the labeled vehicle samples, allowing for the design of a realistic training and evaluation procedure, with a particular emphasis on the limited data availability in the early stages of a new generation release, followed by a gradual increase in reference labels over time as the fleet and technology ages and more and more vehicles receive a full charge reference measurement.

Data is collected under nominal vehicle operating conditions and may contain missing values or erroneous measurements. Potential causes of such signals include onboard controller flash updates and sensor noise. Before being used for model development, all available data undergoes a preprocessing pipeline, as data quality greatly affects the predictive power of machine learning models [49]. The dimensions of data quality include: consistent representation, completeness, feature accuracy, target accuracy, uniqueness, and balance of target classes [50]. Data is cleaned by removing duplicates, removing samples with missing values, encoding categorical values, and addressing implausible sensor values. Implausible signals are defined as measurements outside the physical sensor limits, such as SOC values below 0 % or above 100 %. Samples containing such values are excluded, as they typically result from sensor malfunctions or battery management unit conversion errors and can negatively affect the learning process.

The data is normalized in order to scale the features to a common range, which improves the numerical stability and convergence speed. A correlation analysis is performed to select learning features with the highest correlation with the target variable and the lowest multicollinearity [33]. Here, features with an absolute Pearson correlation coefficient, r, below 0.5 are considered to have low correlation with the target variable and are excluded from the training data. The Pearson correlation coefficient is derived from Eq. (5), where x_i, y_i represent the individual points of two variables, and $\bar{x}$, $\bar{y}$ represent the sample means of the respective variables. Among the features with a correlation coefficient above the threshold, there are features with a coefficient factor above 0.9, indicating high collinearity. For such pairs, only one of them is included in the model training and the selection is based on the higher correlation with the target variable. This reduces redundancy in the input feature space and mitigates prediction instability that can result from highly correlated independent features [51]. The correlation coefficient values are presented in the heatmap in Appendix A.1, Fig. A.8.

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{5}$$

After preprocessing, a hold-out validation set of 565 vehicle samples is created from the remaining 2271 samples. To ensure proportional representation across all SOH groups, the data is segmented into 5 % SOH steps, with 20 % of each subset allocated to the hold-out validation set. This dataset is used to ensure the integrity of the model's performance

Table 1
Overview of the relevant variables collected from fleet vehicles. The dataset includes vehicle meta-information, such as age, mileage, average ambient temperature, and high-voltage battery measurements, recorded on average on a weekly basis.

Name	Description	Unit
Age	Age of the high-voltage battery, calculated as the difference between the readout date and the vehicle production date	years
Mileage	Total kilometers driven until readout	km
Temperature	Average ambient temperature readings, accumulated over the battery's aging history	°C
Charge Throughput	Accumulated positive (charging) and negative (discharging) charge throughput	Ah
C-rate	Cumulative average current normalized by nominal battery capacity and reported separately for charge and discharge	h^{-1}
SOC	Weighted average of the binned SOC histogram, obtained by discretizing the time-series SOC data into predefined intervals or bins	%
Internal Resistance Increase	Deterioration of internal resistance over time determined as the ratio $\frac{R_i}{R_{BOL}}$, where resistance values are estimated by a separate onboard algorithm	%
Cell Temperature	Mean cell temperature calculated by averaging the measurements across all cells equipped with temperature sensors	°C
Cell Voltages	Minimum, mean, and maximum cell voltage measurements, where the minimum and maximum voltages represent the lowest and highest measured value, respectively, and the mean voltage represents the mean of all the recorded cell voltage measurements at the time of the readout	V
Battery Voltage	Average battery voltage measured at battery terminals	V

on previously unseen data, regardless of the model or test design being evaluated. A preview of the final data frame used as input is provided in Appendix A.2, Table A.6. Available input features include vehicle age, total mileage, average ambient temperature, average charge and discharge throughput, average SOC, average discharge C-rate, and average increase in internal resistance.

3.2. Experimental design

The training and validation pipeline for semi-supervised battery SOH estimation is designed to address the challenges of data generation in real-world applications. Fig. 2 illustrates the time-dependent increase in data volume and the change in the SOH distribution of available samples over time, which directly impact the ability of the data-driven estimation methods to capture the aging behavior in the field.

The training process starts with initially available labeled samples from the ordered training set for both the benchmarked SL models and the novel SSL model. After the initial training, the model predicts SOH values for the unlabeled samples of that iteration. Some predictions are converted to pseudo labels and added to the training set. In each subsequent iteration, the observation window is shifted, expanding the pool of available data to new samples consisting of a limited number of labeled samples and a larger number of unlabeled samples. The length of the observation window is user-defined and can vary depending on the application and the expected rate of change of the data patterns. Here, the observation window is defined to represent battery aging of one year, allowing for a battery lifetime simulation and providing a deeper understanding of the model's ability to capture changing aging behavior.

The limited number of labeled samples introduced into the training process in each iteration represents the naturally incoming true labels resulting from the increasing number of vehicles undergoing a reference measurement. The larger set of unlabeled samples represents the remaining vehicles in the fleet without reference measurement. In each new iteration, the newly added labeled samples are first used to validate the accuracy of the model from the previous iteration. After the evaluation, they are added to the training set to improve the model's coverage of existing trends. Such a framework allows the model to capture patterns that describe different stages of aging and field trends that may not have been part of the test environment. It also accounts for the complexity and variability of the field patterns, allowing for a realistic evaluation of the performance of data-driven estimation methods in the field. Additional steps that leverage unlabeled data and distinguish the novel SSL from the traditional SL approaches are introduced in detail in Section 3.4.

This framework is applicable to any use case where labeled data is initially limited but increases over time. The experimental design is based on the two assumptions regarding the average distribution of the field data and the field aging, i.e., the rate of change of the observed patterns. First, battery aging is a slow and gradual process that unfolds over time [52], hence there are no extreme shifts in the feature distribution over brief periods. Consequently, the changes that indicate capacity loss in the field may not be visible over short time intervals, such as a month. Also, a model that is exclusively trained on data comprising the aging behavior of fleet vehicles up to, e.g., three years may not accurately predict SOH for vehicles in operation for a significantly longer time, e.g., eight years. Second, the labeled samples, i.e., the reference test results, increase over time and reveal additional aging patterns. This assumption is linked to the nature of reference data generation. As the fleet grows, there are more vehicles that undergo a full charge reference test and the number and diversity of SOH labels increases. If generating reference labels through full-charge reference measurements under a standardized protocol is not possible, reference labels may be obtained via alternative approaches, provided that they satisfy the application's precision requirements and characteristics, by defining strict guidelines for naturally occurring field events.

3.3. Supervised learning (reference) model development

The SSL method developed in this study is based on two independent SL models. The most commonly used data-driven battery state estimation methods are: artificial neural networks, random forest and gradient boost algorithms, support vector machines, and k-nearest neighbors [53–55]. The methods listed operate on different principles to predict the output of the regression fit. This diversity allows for a comprehensive investigation of the strengths and limitations of data-driven approaches in the context of SOH estimation using fleet data. A total of five SL models are run through a benchmarking pipeline that includes hyperparameter tuning, training, and evaluation on the hold-out validation set. This ensures a fair comparison and accurate evaluation of the predictive performance of the models. Moreover, the two best performing models are selected for SSL method development. The supervised model with the highest evaluation score is also employed in the final performance evaluation of the newly developed approach.

Hyperparameter tuning is performed through a Bayesian optimization process, an iterative approach that uses probabilistic models to predict model performance based on prior evaluations. This is achieved by constructing a surrogate model of the objective function to identify the subsequent set of hyperparameters for evaluation, as described by Koehrsen et al. [56]. In this study, the scikit-learn optimization library [57] is used with 100 iterations, 10 cross-validation loops, 5

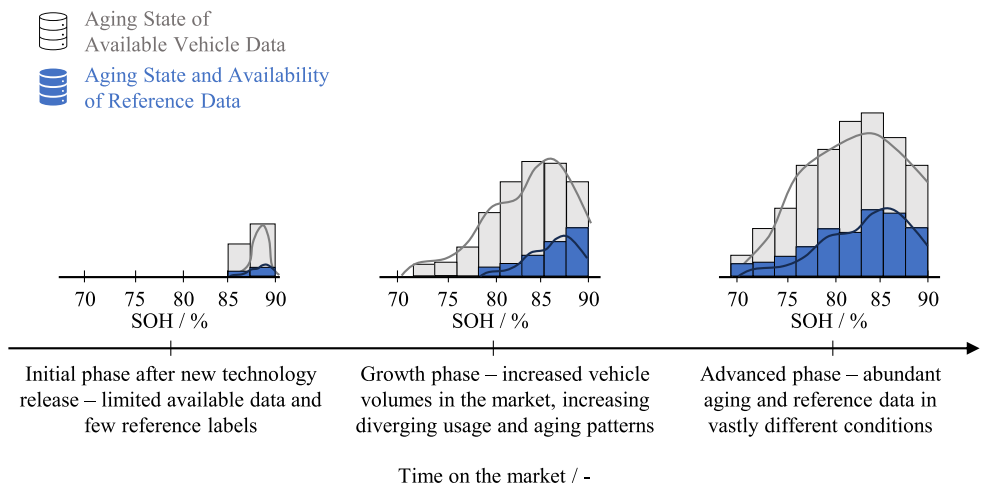


Fig. 2. Illustration of the fleet data generation and its time dependency characteristics. Along with the data volume, the number of ground truth reference labels rises with time, since more vehicles undergo a capacity measurement. Moreover, with time, the amount of diverging aging patterns increases as the vehicles experience diverse conditions, such as extreme weather conditions and driving behavior.

Table 2
Summary of the defined search space for each of the five supervised methods, including the optimization results which denote the optimal set of configuration parameters for that method.

Method	Search space	Result
Artificial Neural Network	Number of Layers: [3, 4, 5],	3
	Size of Hidden Layers: [10, 300],	100
	Batch Size: [10, 500],	10
	Maximum Number of Iterations: [50, 250],	246
	Validation Fraction: [0.1, 0.5]	0.12
Random Forest	Number of Estimators: [10, 300],	151
	Maximum Tree Depth: [1, 10],	8
	Minimum Number of Split Samples: [0.01, 1],	0.01
	Maximum Number of Features: 'sqrt', 'log2'	'sqrt'
Gradient Boost	Number of Estimators: [10, 300],	55
	Maximum Depth: [10],	9
	Learning Rate: [0.01, 0.70],	0.24
	Minimum Number of Split Samples: [0.01, 1],	0.14
	Minimum Leaf Weight Factor: [0.01, 0.50],	0.22
	Subsample: [0.01, 1]	0.74
Support Vector Machine	Epsilon Tube: [0.01, 10],	4.29
	Regularization Parameter C: [1e−3, 1e+4],	120
	Degree: [1, 5],	4
	Kernel Function: {'poly', 'rbf'},	'rbf'
	Kernel Coefficient: {'scale', 'auto'},	'auto'
	Kernel Independent Term: [−5, 5]	3.5
K-Nearest Neighbor	Number of Neighbors: [1, 250],	11
	Weights: {'uniform', 'distance'},	'distance'
	Algorithm: {'ball_tree', 'kd_tree'},	'kd_tree'
	Power Parameter: [1, 2]	1

different hyperparameter combinations per iteration, and RMSE as the evaluation metric. The final search space for each of the five models and the corresponding optimization results are summarized in Table 2.

Each model is initialized with the individual set of hyperparameters and trained on the same training samples. The total available labeled data contains 450 samples, split in a 70/30 ratio for training and testing, respectively. Generalization performance, i.e. the ability of the models to produce consistent results on unseen data, is evaluated using the validation hold-out set, which contains 565 vehicle samples.

Table 3
Results of the benchmarking analysis of the five supervised methods, showing the RMSE, MAE scores, and the violation of the 5 % tolerance band. The benchmark is a result of validation under the hold-out validation set.

Model	RMSE	MAE	$\text{Tolerance}_{5\,\%}$
Artificial Neural Network	3.2 %	2.4 %	11 %
Random Forest	3.4 %	2.6 %	13 %
Gradient Boost	3.1 %	2.3 %	10 %
Support Vector Machine	3.5 %	2.6 %	13 %
K-Nearest Neighbor	3.6 %	2.8 %	15 %

The results of the SL benchmark are presented in Table 3. Among the models considered, the gradient boost model shows the best performance with RMSE of 3.1 %, MAE of 2.3 %, and 10 % of test samples outside of the tolerance band. The second-ranked supervised model, the multi-layer perceptron, achieves RMSE of 3.2 %, MAE of 2.4 %, and 11 % of test samples outside of the tolerance band. These two models are selected as the base models for the SSL approach.

Overall, the benchmarking results show that supervised approaches can capture the predominant aging patterns present in the field, given a sufficient number of labeled samples in the training set. However, a large violation of the tolerance band indicates challenges with the prediction accuracy, possibly due to the presence of aging patterns that are not well represented in the training set and/or the need to adapt the training procedure to the data curation specific to the application. Therefore, supervised models may not always be appropriate at all times in applications where the availability of ground truth is limited and time-dependent, as outlined in the problem statement of this study. For such applications, it is imperative to explore alternative approaches to data-driven SOH estimation.

To highlight the importance of labeled data in the supervised framework, Fig. 3 shows the prediction results for two different scenarios: one with a gradient boost model trained on a substantial amount of labeled samples, namely 400 samples, and another trained with only 60 training samples. In both scenarios, the models are tested on the same validation dataset, and their SOH estimates are compared to the SOH true values. The dashed lines represent the ± 5 % tolerance bands. It can be concluded that the second model fails to generalize under field variability and is limited to predicting conditions similar to those seen during training, as evidenced by the nearly horizontal lines in the figure.

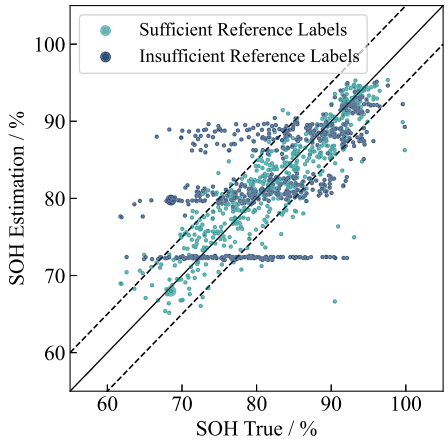


Fig. 3. Predictions of a supervised model for two scenarios: sufficient and insufficient reference labels, where SOH true represents the full charge reference measurement results, the SOH estimate represents the model estimates, and the dashed lines represent the ± 5 % tolerance band. Sufficient reference labels contain a total of 400 vehicle samples, while insufficient reference labels contain 60 vehicle samples. The second scenario shows poor performance due to the lack of labeled data, demonstrating the impact of training data on the model's prediction and generalization performance.

3.4. Semi-supervised learning model development

The implementation of the semi-supervised co-training method developed in this study is illustrated in Fig. 4. A co-training self-supervised algorithm is based on the premise that two or more supervised models can leverage their different views on available data, based on the independent architecture or unique data segments, increasing confidence in the identified overlapping patterns [35,58]. The first implementation of the co-training approach involves training identical models on separate data segments, cross-evaluating their predictions, and augmenting each model's training set by incorporating all estimated labels from the other in subsequent iterations [35]. While effective in domains such as natural language processing, this strategy can induce a relatively high rate of labeling errors, which is often tolerable due to the inherently large data throughput and high error tolerance of such applications [38]. In contrast, the data throughput and error tolerance in this study are considerably lower than in big-data settings. Consequently, the distinction between models' views is established primarily through differences in model architectures rather than through partitioning of the data.

The method involves two supervised models and a rule-based pseudo label selection (PLS) that controls the expansion of the training set for subsequent iterations. The training procedure begins with the supervised training of both models based on initially available vehicle data, typically generated during the test phase of the vehicle or battery technology under development. The first model is a gradient boost model with a learning rate of 0.24, 55 estimators, and a maximum tree depth of 9. Input features include vehicle age, total mileage, ambient temperature, discharge current throughput, average SOC, average increase in internal resistance, and discharge rate. The second model is a multilayer perceptron with 100 neurons in a single hidden layer, a batch size of 10, and a maximum of 246 iterations. Input features for this model are vehicle age, total mileage, ambient temperature, discharge current throughput, and discharge rate. The input feature sets are distinguished by slight variations to enhance the distinction between the two views, however, they belong to the same unique vehicle samples. After training, the models predict SOH labels for the unknown samples. Unknown samples represent all vehicles in the field whose data is available in the given iteration, but which do not have a full charge reference measurement. The predictions are then fed into the PLS mechanism described in Algorithm 1.

Algorithm 1 Co-Training Validation

Ensure: $y_t - \hat{y}_t \approx 0, y_{t+1} - \hat{y}_{t+1} \approx 0$ where
y = true reference, $\hat{y}$ = model prediction

Define pseudo label candidates:
$\hat{y}_t = \frac{\hat{y}_{\text{model1}} + \hat{y}_{\text{model2}}}{2}$ where $|\hat{y}_{\text{model1}} - \hat{y}_{\text{model2}}| \leq 1\%$

In parallel branch:
(1) Expand the training set $(X, \mathbf{y})_t, (X, \hat{y})_t$
(2) Validate models' learning using $(X, \mathbf{y})_{t+1}$

if $|y_{t+1} - \hat{y}_{t+1}| \leq 1\%$ **then**
  Expand training set with $(X, \hat{y})_t$ and $(X, y)_{t+1}$
else
  Reject added pseudo labels in parallel branch
  Strengthen training set of previous iteration with new true labels $(X, y)_{t+1}$
  Retrain & re-evaluate
end if

PLS compares the predictions of both models and marks the predictions that fall within the tightened tolerance band as pseudo label candidates, assuming that if different models compute a similar estimate for an unknown sample, then the battery aging characteristics of that sample are well represented in the current training set and thus the confidence in the prediction accuracy of the current models is high. The tightened tolerance band represents a ±1% deviation from the true labels, but it can be loosened for a higher turnover of pseudo labels, depending on the application's tolerance for error. Here, the tightened tolerance band, which is responsible for controlling the conversion of the pseudo labels, is kept at a low level to minimize the error probability in the extended training set induced by pseudo labels of the given iteration. The pseudo label candidates are evaluated in the learning validation step, which starts a separate branch and trains both models with the extended training set consisting of previous training data combined with pseudo labels. Depending on the specifics of the use case, the PLS mechanism can be extended with additional rules.

When new true references become available, they are first used to validate the learning and performance of the models trained on the extended training set in the parallel branch. If the absolute prediction error of the models in the parallel branch is below 1% threshold, the validation is successful and the branches are merged, i.e., the pseudo labels used in the parallel branch are accepted into the training set for the next iteration. The proposed approach is designed to minimize the risk of extrapolation by prioritizing prediction consistency across the two model views and incorporating pseudo labels only when prediction confidence exceeds a defined threshold. This selective mechanism ensures that pseudo labels are drawn primarily from regions of the data space that are already well-represented, thereby encouraging interpolation rather than extrapolation. To further minimize the risk of precision losses due to averaging, if the validation is successful, only the first model's predictions are retained as pseudo labels. In the case of unsuccessful validation, the parallel branch is discarded and a subset of the available new labels is added to the existing training set of the parallel branch to increase the coverage of the unlabeled data distribution in the training. The models are then retrained and the learning validation step of the PLS is repeated with the remaining newly available labeled samples.

4. Results and discussion

The present study investigates the potential of SSL for battery state estimation in real-world EV applications, with particular emphasis on the challenges arising from limited availability of ground truth

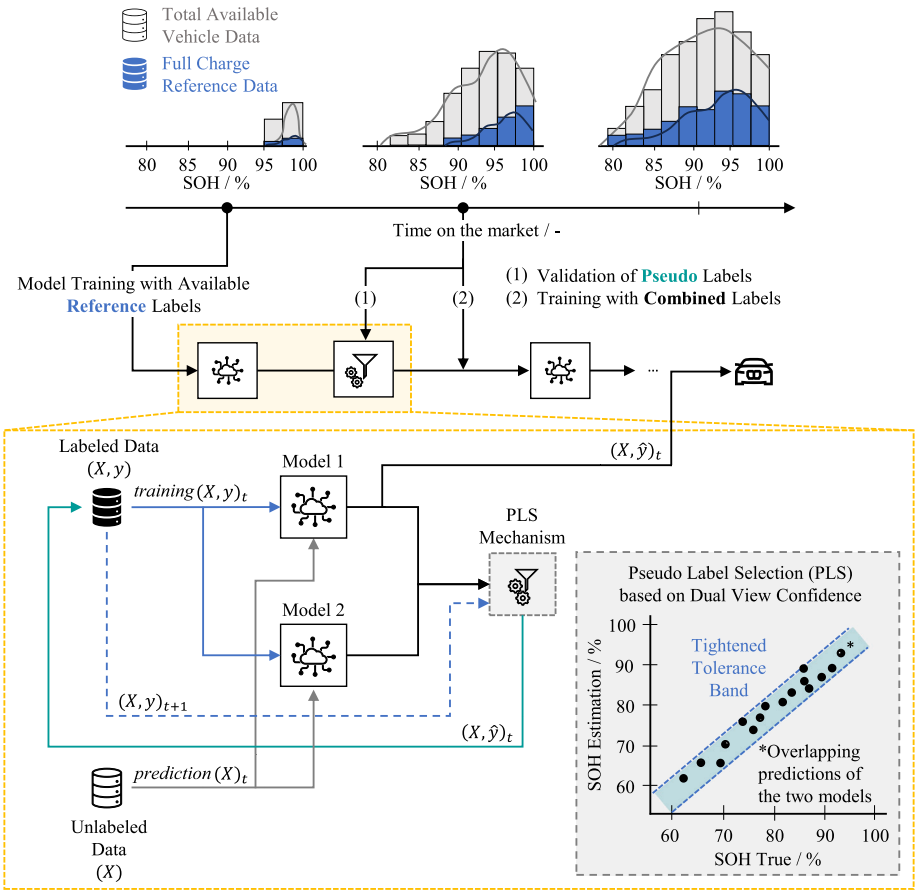


Fig. 4. Illustration of the SSL algorithm for SOH estimation in field applications. The training process is designed to maximize the use of available reference labels by employing a rule-based multi-view pseudo labeling technique. Initially, the available reference labels are used for model training and initialization of the pseudo labeling mechanism, which compares the results of the two learning models and considers those with small difference as pseudo labels. As additional reference labels become available, they are first used to validate the pseudo label selection, and only then are they added to the combined set of true and pseudo labels in the next iteration of model training. With such a training design, the algorithm can better capture the changing dynamics in the field, despite the small number of full charge reference samples available for training.

measurements and the temporal dependencies inherent in aging processes. Building on state-of-the-art supervised approaches, the proposed method introduces a controlled training set expansion strategy that integrates co-training, a multi-view learning paradigm, with a pseudo labeling mechanism, thereby enabling effective utilization of unlabeled data. The training and evaluation pipelines are explicitly designed to mirror real-world field conditions, enabling a realistic assessment of model performance in application-relevant scenarios. The central discussion focuses on the performance comparison of the SL and SSL methods, analysis of the accuracy trade-offs introduced by pseudo labeling, the sensitivity to the quantity of reference labels that initialize and support the SSL training process, and an analysis of the recurrent outliers. The comprehensive analysis is facilitated by the availability of SOH labels of all vehicle samples included in this study, which are withheld during training.

4.1. Benchmark

Performance benchmark is conducted against the optimal SL model identified in Section 3.3, using a test pipeline over eight successive one-year iterations on a fixed hold-out set. The hold-out set used for model evaluation is defined prior to any training and also employed both for SL benchmarking and for all subsequent case studies. Fig. 5 visualizes the effectiveness of the proposed method in real-world conditions. Evaluation metrics, along with the number of training samples, distinguished by true labels and pseudo labels, and test samples used in each iteration are provided in Table 4.

The early iterations underscore the benefit of SSL under extreme label scarcity. In the first iteration, shown in first subplot of Fig. 5, the reference SL model fails to capture underlying patterns, as evidenced by the nearly linear point distribution. By contrast, SSL approach augments the four true labels with six high-confidence pseudo labels, setting RMSE to 1.5 % and MAE to 1.1 %. SL metrics are over 1 percentage point higher. In iteration two, with 49 true reference labels, SL model improves, scoring 1.7 % in RMSE and 1.1 % in MAE, with only one sample on the borderline of the tolerance band, yet SSL maintains a marginal edge in overall accuracy.

Mid-stage iterations three and four underscore the resilience of the SSL framework as degradation trajectories broaden and lower SOH states enter the validation set. In iteration three, the SL model's RMSE degrades to 3.8 %, with 2.5 % in MAE and 14 % outside the tolerance band, reflecting its inability to capture the expanded SOH span. The third subplot in Fig. 5 shows predicted points located above the 90 % SOH threshold, indicating the absence of lower-SOH examples in its training set and the model's consequent inability to capture those degradation states. Meanwhile, pseudo labels rise from 35 to 123 and 312 across iterations, demonstrating PLS validation and SSL's use of unlabeled data, but the growing fraction of tolerance-band violations suggests that too many pseudo labels can undermine accuracy. The prediction evaluation shows 3.7 % and 2.7 % in RMSE, and 2.1 % and

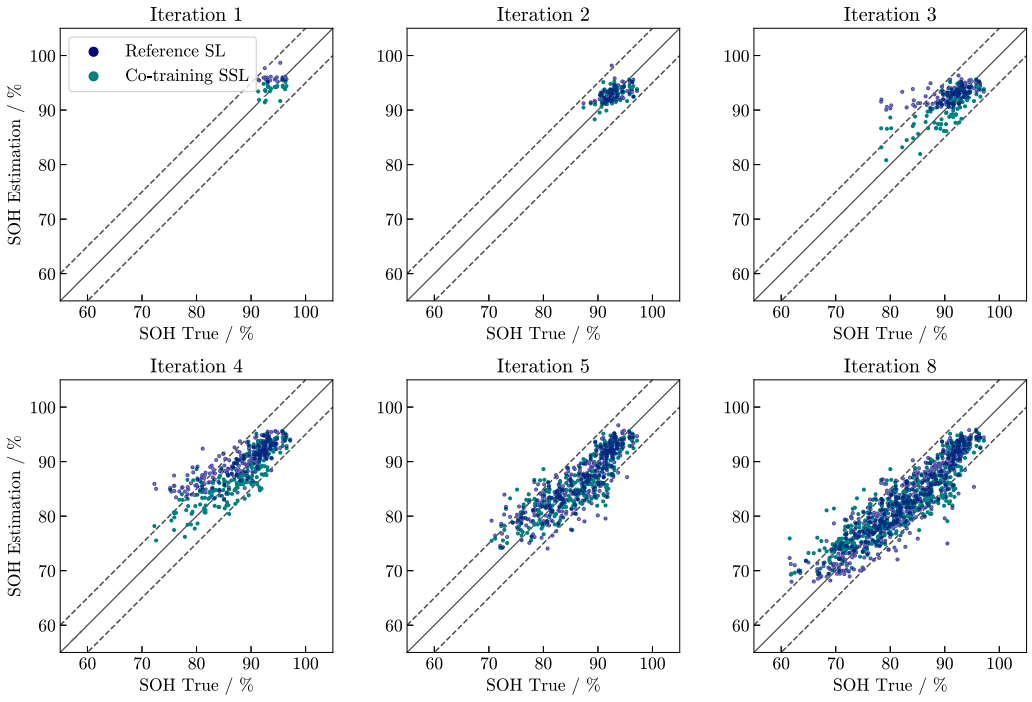


Fig. 5. Performance comparison between the reference SL, represented by blue points, and the proposed co-training SSL, represented by green points, focusing on the first iterations and the turning points dominated by the increase in reference labels from full charge measurements. The points outside the dashed lines represent the predictions that violate the ± 5 % tolerance band. Iteration frequency and segmentation can be easily adjusted to suit application characteristics. Here, they represent one year of battery aging relative to the vehicle sample.

Table 4
Overview of the number of samples included in the training and validation of each iteration, SOH range of samples in the validation set, and relevant evaluation metrics for the two methods. While true labels are accessible to both approaches, pseudo labels contribute exclusively to the semi-supervised learning process. In the presented case study, the iterations represent one year of battery aging relative to the vehicle sample.

Iteration	Method	Training samples		Test samples	SOH Range	RMSE	MAE	$\text{Tolerance}_{5\,\%}$
		True labels	Pseudo labels					
1	SL Reference	4	–	18	91.3–97.1 %	2.6 %	2.2 %	0 %
	SSL Co-Training		6			1.5 %	1.1 %	0 %
2	SL Reference	49	–	42	90.5–97.1 %	1.7 %	1.3 %	1 %
	SSL Co-Training		35			1.6 %	1.3 %	0 %
3	SL Reference	133	–	100	84.5–97.1 %	3.8 %	2.5 %	14 %
	SSL Co-Training		123			2.3 %	1.8 %	3 %
4	SL Reference	275	–	181	72.5–97.1 %	3.7 %	2.7 %	18 %
	SSL Co-Training		312			2.1 %	1.5 %	7 %
5	SL Reference	525	–	290	70.6–97.1 %	3.2 %	2.4 %	12 %
	SSL Co-Training		584			3.1 %	2.5 %	10 %
6	SL Reference	808	–	388	70.1–97.1 %	3.3 %	2.5 %	13 %
	SSL Co-Training		897			3.3 %	2.6 %	11 %
7	SL Reference	1129	–	491	63.5–97.1 %	3.3 %	2.5 %	12 %
	SSL Co-Training		1163			3.5 %	2.7 %	13 %
8	SL Reference	1421	–	565	62.2–97.1 %	3.2 %	2.3 %	10 %
	SSL Co-Training		1330			3.5 %	2.7 %	13 %

1.5 % in MAE for SL and SSL, respectively, when models are trained on 275 true reference samples and 312 pseudo labels. The upward bias of the SL method toward the higher SOH labels, again evident in the fourth subplot of Fig. 5, can be attributed to the increased SOH range and divergent aging patterns in the field that were not part of the dataset with the true labels. The overall advantage of the SSL approach is most pronounced in these early stages of deployment, where scarce true labels make pseudo label augmentation essential for capturing emerging degradation patterns.

The fifth iteration marks the point at which true reference label abundance begins to rival the benefits of pseudo labeling. With 525 true and 584 pseudo labels, SL and SSL RMSE converge to 3.2 % and 3.1 %, and tolerance violations narrow to 12 % and 10 %, respectively. This plateau effect continues in the last three iterations. RMSE of the SL model stabilizes near 3.2 %–3.3 %, MAE at about 2.5 %, while SSL RMSE hovers between 3.3 % and 3.5 % with MAE around 2.7 %. The tolerance violations of SSL slightly exceed those of SL in the last two iterations.

These results confirm that SSL co-training confers substantial accuracy and coverage advantages during the early deployment phase, when true SOH labels are scarce and aging behaviors evolve rapidly, by leveraging unlabeled data to fill coverage gaps. As true labels accumulate, the relative benefit of pseudo labels diminishes, and a purely supervised regime attains parity. Crucially, no evidence of uncontrolled

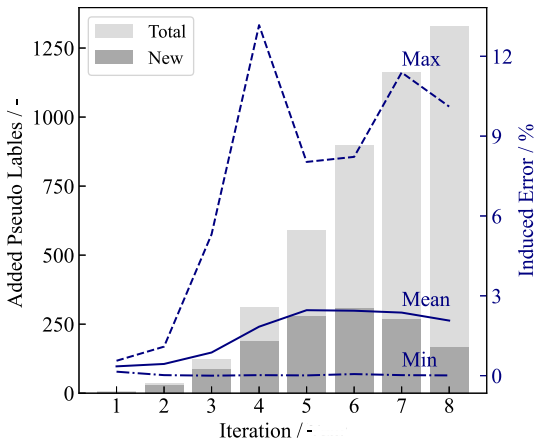


Fig. 6. Breakdown of the number of added pseudo labels in the training process of the SSL approach, showing the minimum, average, and maximum labeling error of newly added pseudo labels in each iteration.

pseudo label drift was observed across eight years of simulated operations, validating the robustness of the pseudo label selection mechanism and underscoring the practicality of SSL frameworks for large-scale, field-based SOH monitoring.

4.2. Induced error

The pseudo labeling approach inherently introduces a degree of error into the model training process. By design, the training set is augmented with pseudo labels that represent the model's current best predictions, which may compromise accuracy. Even when the error rate of pseudo labels is kept to a minimum, a disproportionate number of pseudo labels in the training set can lead to significant error propagation, i.e., the model learns from an inaccurate training set, thereby propagating errors in subsequent iterations.

Fig. 6 shows the minimum, maximum, and mean error of pseudo labels added to the SSL training process. The figure also provides a split between the pseudo labels added to the current iteration step and the total accumulated pseudo labels up to that step. The lowest mean of pseudo label accuracy error of 0.5 % is achieved in the early stages when the conversion rate is at its lowest. Starting from the fourth iteration, with over 300 pseudo labels added to the training process, the mean error starts to increase with higher pseudo rate conversion, indicated by the rising error marker in Fig. 6. Furthermore, the maximum error curve shows that there is at least one pseudo label with an error above 10 %. A high number of the converted pseudo labels, coupled with an increased labeling error rate, may explain why the SL starts to outperform the SSL after the fourth iteration. The pseudo labels that are accepted into the training set in this iteration show their impact in the following iterations. The maximum mean error is reached at 2.9 % in the sixth iteration, which may be the consequence of the pseudo labels starting to dominate the training set after the fourth iteration.

The use of pseudo labels can still provide a significant advantage, as demonstrated in Figs. 3 and 5. However, the pseudo labeling mechanism needs to be extended with supplementary rules that consider the availability of the true labeled samples in the training loss function. This will enable the optimization of the allowed pseudo label conversion rate according to the error sensitivity of the application. The benefits of SSL with a PLS mechanism over the SL approach in critical scenarios are analyzed in the following section.

4.3. Limited reference availability

To investigate the effect of the induced errors on the model performance, the reference SL model and the co-training SSL model are compared in two scenarios. In the first scenario, both models are trained under optimal training conditions for data-driven estimation methods. The training set consists of 1500 labels that are gradually added to the training process while still simulating real-world field conditions. Training starts with 60 labeled samples, with the number of labeled samples doubling in each successive iteration. In addition to this, the SSL model extends the training set with its pseudo labels. This represents a best-case scenario for training data-driven methods in a given application, with respect to the availability of reference labels. Moreover, in such a setting, the risk of pseudo labels dominating the training set can be eliminated. In contrast, the second scenario represents a case of insufficient availability of reference samples, which is a critical scenario for data-driven methods. In total, 600 labeled samples are used in simulating iterative model training under field conditions over an eight-year period. The training begins with six samples and proceeds at a steady, low rate, with 20 to 40 samples added in each iteration. In both scenarios, an evaluation is performed against the same hold-out validation set with samples from all SOH groups as used in previous evaluations. The objective here is to assess the merits of the SSL approach while considering the challenges it faces with error caused by pseudo labeling over time. While six samples in the second scenario proved sufficient to characterize the dominant aging behavior in the first year of our dataset, we do not claim that six labels will suffice at all stages or for all vehicle populations, especially as aging signals diverge over time or across operating conditions. An in-depth investigation of the minimum number of training samples per aging path or customer group falls beyond the scope of the present work, however, such analysis represents a critical direction for future research.

Fig. 7 shows a central tendency of the estimation accuracy of the compared methods: the reference SL model is visualized with blue boxes, the co-training SSL is visualized with green boxes, and their counterparts trained with insufficient data are visualized with dashed boxes of the same color. The dashed line at the 5 % accuracy marker highlights for which feature segments the methods tend to exceed the tolerance band. The figure highlights the error distribution over the true SOH and three other features, namely total mileage, battery age, and average ambient temperature. In addition, the box plots also indicate which feature segments may be underrepresented during the training process, and thus not accurately predicted. This information can be used when selecting labels to augment the training set as a weighting factor or prediction uncertainty indicator for specific vehicle groups.

When trained on sufficient data, the SSL method demonstrates an advantage for the data segments that are at the tails of the distribution of all available samples, as shown in Fig. 1, such as customer groups with SOH below 70 %, total mileage above 150 000 km or vehicles operating in regions with an average ambient temperature below 10 °C. This shows that SSL can compensate for the lack of samples describing specific customer groups, as long as there are some samples to provide guidelines for the pseudo labeling mechanism. In addition, the SSL shows better performance in data segments centered around the mean for features with flattened distributions, such as battery age or temperature, where aging patterns are expected to diverge. The SL method has access exclusively to the aging patterns covered by the vehicles with full charge reference labels, which may not fully represent the predominant patterns observed in the field, but the SSL can capture additional patterns observed in the unlabeled field data. Overall, the SSL shows a one percentage point improvement in both RMSE and MAE. With regard to the total number of predictions that fall outside the tolerance band, SSL shows an improvement of 6 %.

The methods differ significantly when trained on insufficient data. The dashed box plot in blue represents the performance of the SL method in the second worst-case field scenario, where the SSL approach outperforms it in almost every feature segment. The prediction errors of the reference SL trained on insufficient training data extend beyond the

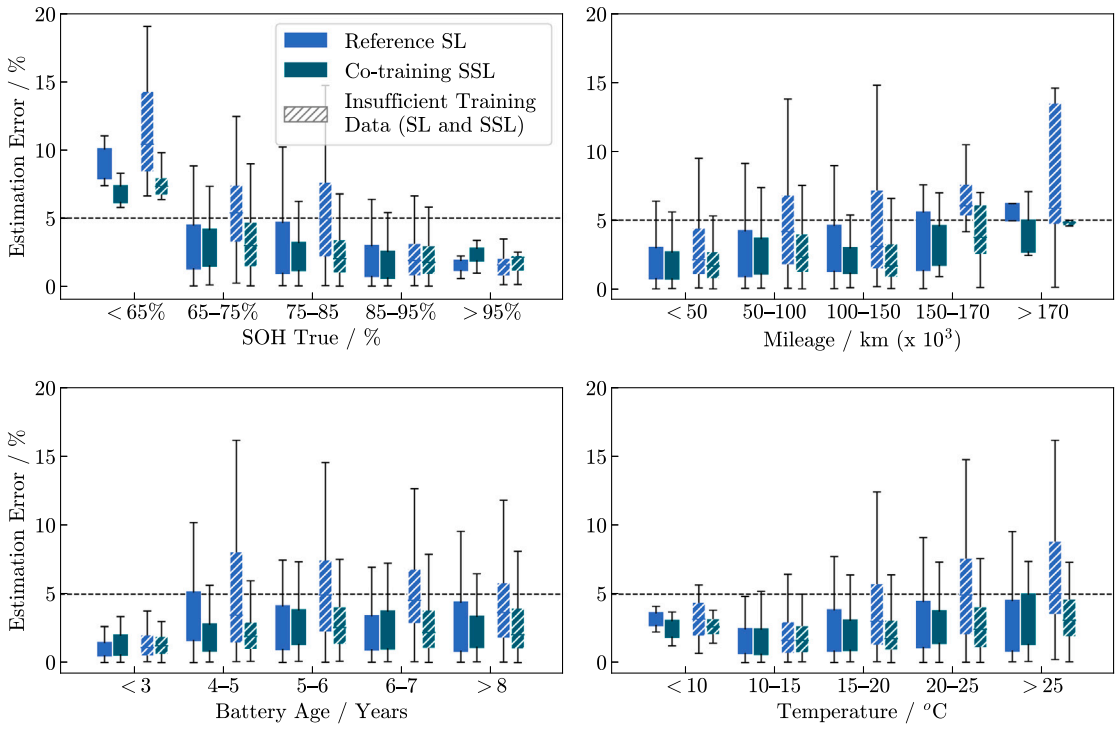


Fig. 7. Analysis of the impact of the limited reference availability and pseudo labeling on model performance in two different scenarios: one with sufficient ground truth labels for robust data-driven estimation, and another with insufficient ground truth labels for a data-driven estimation based on limited field data.

Table 5
List of extreme outlier examples, including battery age, total mileage, average ambient temperature, average discharge throughput, true SOH value, and absolute estimation errors of both SL and SSL methods. A positive sign indicates overestimation, meaning the models predict a higher SOH value than is measured, whereas a negative sign denotes underestimation, with the models predicting a lower SOH value than is measured.

VAN	Age	Mileage	Discharge throughput	Temperature	SOH true	SL error	SSL error
d17…	4 years	75 976 km	44.2 kAh	15.1 °C	89.7 %	−5.2 %	−5.5 %
ca9…	5 years	51 533 km	26.8 kAh	27.3 °C	87.5 %	−2.3 %	−7.2 %
hk8…	6 years	92 888 km	41.0 kAh	25.6 °C	70.3 %	9.3 %	5.5 %
au7…	6 years	108 051 km	59.3 kAh	26.0 °C	82.1 %	−10.3 %	−6.7 %
ap4…	7 years	126 391 km	72.6 kAh	20.0 °C	82.6 %	−4.2 %	−5.4 %
a91…	7 years	26 679 km	13.5 kAh	19.8 °C	91.5 %	−11.2 %	−8.3 %

tolerance limit for almost all data segments, reaching a peak absolute error of almost 20 percentage points. In this scenario, the SSL method achieves an improvement of at least 28 % in terms of the number of samples violating the tolerance band. Another signal of poor coverage of aging patterns in the field is the comparatively tall estimation error box plots for samples in the worst-case scenario, which indicate a larger spread and increased uncertainty of the model in the predictions. This highlights the need for alternative approaches in field applications where reference data is scarce and underlying patterns diverge over time, as otherwise data-driven methods developed with limited data, such as those generated in short-term and controlled laboratory experiments, may fail. To further explore the limitations of the proposed method, the potential causes of the outliers, defined here as predictions that violate the tolerance band, are discussed in more detail in the following section.

4.4. Outlier analysis

The purpose of outlier analysis is twofold: first, to understand the limitations of the method, and second, to investigate possible patterns of inaccurate predictions. Outliers are defined as samples with SSL SOH estimates that violate the ±5 % tolerance band. Table 5 shows the relevant features, as well as the true SOH label resulting from the capacity measurements, and the SL and SSL estimation errors, for the most critical outlier examples. The negative sign in the error columns denotes underestimation, meaning that the models predicted a lower SOH value than was actually measured, and vice versa.

Most of the outliers overlap between methods, indicating that some aging patterns are not well described by the data or have conflicting information in the strongest features, leading to over- or underestimation by both methods. For example, an underestimation of 11.2 or 8.3 percentage points for a vehicle that is seven years old and has less than 30 000 km may indicate that calendar aging is not well represented in the provided data and that more similar samples or more data features are needed during the training process to capture this type of aging. Similarly, examples of vehicles with low age but high mileage and current throughput, or samples operating at higher temperatures are equally challenging for both methods.

4.5. Outlook

Future work can extend the proposed framework in two main directions: alternative implementations of PLS to extend the control over the pseudo label conversion and advanced feature engineering to identify divergent aging patterns with higher precision. The PLS refinements include rule extensions to regulate the number of added pseudo labels and continuous re-evaluation of the accepted labels. This can be accomplished by distinguishing between true labeled samples and

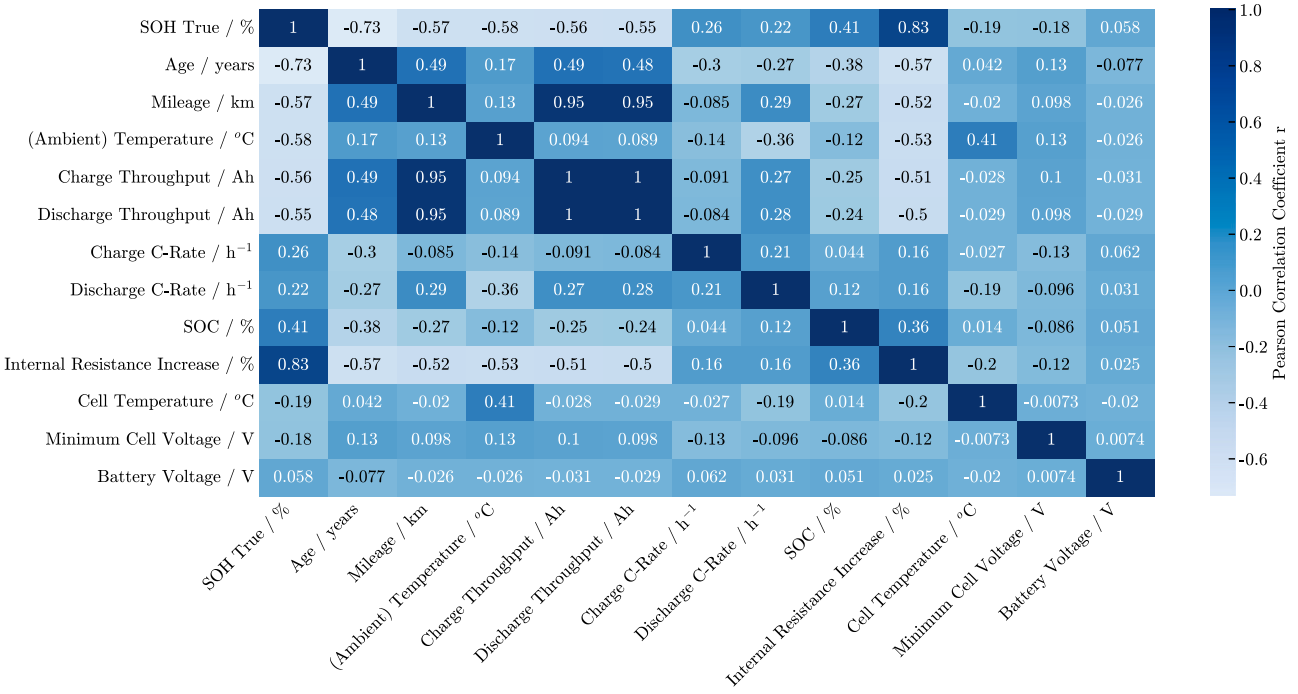


Fig. A.8. A correlation heatmap, illustrating the relationship between variables relevant for battery performance. Darker shades correspond to higher Pearson correlation coefficients, indicating stronger linear dependencies. Recall: Variable 'SOH True / %' represents SOH value in % calculated after the reference measurement; 'Age / years' represents battery age in years; 'Mileage / km' represents cumulative distance driven at the time of readout; '(Ambient) Temperature / °C' represents average ambient temperature in which the vehicle has been operated; '(Dis)Charge Throughput / Ah' represents accumulated charge throughput over the battery lifetime; '(Dis)Charge C-rate / h^{-1}' represent average C-rate in both charge and discharge direction in h^{-1}; Internal Resistance Increase %' represents estimated increase of battery internal resistance relative to the BOL value in %; 'Cell Temperature / °C' represents mean value across available cell temperature sensors; 'Minimum Cell Voltage / V' represents lowest measured cell voltage within the pack; 'Battery Voltage / V' represents average battery voltage measured at battery terminals.

Table A.6
Example of a data frame head used as input for model training.

VAN	Age	Mileage	Temperature	Discharge throughput	SOC	Discharge C-rate
a57…	5 years	33 060 km	13.2 °C	17.7 kAh	78.6 %	0.6 h^{-1}
ak7…	7 years	75 490 km	26.4 °C	36.4 kAh	63.0 %	0.5 h^{-1}
a6o…	7 years	30 330 km	16.4 °C	20.2 kAh	70.43 %	0.4 h^{-1}
an1…	9 years	142 260 km	23.1 °C	82.2 kAh	52.8 %	0.4 h^{-1}
a7t…	9 years	77 030 km	20.6 °C	42.9 kAh	68.7 %	0.5 h^{-1}

those added by pseudo labeling in the training set, and adjusting the previously assigned pseudo labels in each iteration, as new true samples become available. A certain degree of error is inevitable in practice, since there is no guarantee that a reference label for the vehicle sample with a pseudo label will ever become available. However, it is possible to minimize the risk of induced error by introducing additional rules and limiting the number or the proportion of pseudo labels allowed in the training set. In addition, adding new variables to model training can help capture aging behavior that differs from the average battery aging patterns. Additional variables can be constructed by computing rates of change, extracting higher-order statistical descriptors from histogram data, or integrating event-triggered time-series signals. Furthermore, repeated hyperparameter optimization is recommended for emerging vehicle and battery technologies, since the volume of available field data and the underlying aging patterns evolve over time. Adapting the model architecture online to these shifts can significantly improve predictive accuracy.

5. Conclusion

This study investigates the potential of the semi-supervised approach to address the challenges posed by the limited availability and temporal dependency of labeled data for data-driven SOH estimation. The experimental design reflects real-world conditions of data availability, characterized by initially scarce labeled data that accumulates over time. In critical scenarios when the scarcity of available reference data is most acute, SSL leverages the abundant unlabeled data to augment the training set. In two case studies, one focusing on battery age and the other on data availability, SSL outperforms SL by reducing RMSE and MAE by more than one percentage point and the absolute number of samples violating the tolerance band by up to 28 %.

Overall, the results show that the rule-based, multi-view, semi-supervised method can outperform the traditional supervised methods in critical scenarios and can achieve performance levels at least as good as the supervised reference model, provided that the pseudo labeling mechanism is optimized to the error sensitivity and general characteristics of the field application. The proposed method is most advantageous in the early stages of new technology deployment, when the use of a traditional supervised approach is not feasible. In later stages, the proposed method demonstrates the potential to reduce the dependence on true reference labels and to cover a wide range of existing aging patterns that would otherwise be underrepresented. Nevertheless, effective regulation of error propagation and the pseudo label conversion rate is imperative to prevent the training set from being dominated by the pseudo labels.

CRediT authorship contribution statement

Nejira Hadzalic: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Jacob Hamar:** Writing – review & editing, Supervision, Data curation, Conceptualization. **Marco Fischer:** Writing – review & editing, Supervision. **Simon Erhard:** Writing – review & editing. **Jan Philipp Schmidt:** Writing – review & editing.

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.

Acknowledgments

Funded by the Open Access Publishing Fund of the University of Bayreuth. This work was performed in cooperation with the University of Bayreuth - Chair of Systems Engineering for Electrical Energy Storage, Technical University of Munich - Chair of Electrical Energy Storage Technology and BMW Group.

Appendix A. Field data

A.1. Correlation analysis

See Fig. A.8.

A.2. Input data frame

See Table A.6.

Data availability

The authors do not have permission to share data.

References

[1] Fotouhi A, Propp K, Auger DJ, Longo S. State of charge and state of health estimation over the battery lifespan. In: Behaviour of lithium-ion batteries in electric vehicles: battery health, performance, safety, and cost. 2018, p. 267–88.
[2] Wang Z, Feng G, Zhen D, Gu F, Ball A. A review on online state of charge and state of health estimation for lithium-ion batteries in electric vehicles. Energy Rep 2021;7:5141–61.
[3] Zhao J, Feng X, Tran MK, Fowler M, Ouyang M, Burke AF. Battery safety: Fault diagnosis from laboratory to real-world. J Power Sources 2024;598:234111.
[4] Hamar JC, Erhard SV, Canesso A, Kohlschmidt J, Olivain N, Jossen A. State-of-health estimation using a neural network trained on vehicle data. J Power Sources 2021;512:230493.
[5] Wang Q, Ye M, Celik S, Deng Z, Li B, Sauer DU, et al. Unlocking the potential of unlabeled data: Self-supervised machine learning for battery aging diagnosis with real-world field data. J Energy Chem 2024.
[6] Hofmann T, Hamar JC, Li Jiahao, Erhard S, Schmidt JP. The ΔQ-method: State of health and degradation mode estimation for lithium-ion batteries using a mechanistic model with relaxed voltage points. J Power Sources 2024;596:234107.
[7] Zhao X, Hu J, Hu G, Qiu H. A state of health estimation framework based on real-world electric vehicles operating data. J Energy Storage 2023;63:107031.
[8] Pan H, Lü Z, Wang H, Wei H, Chen L. Novel battery state-of-health online estimation method using multiple health indicators and an extreme learning machine. Energy 2018;160:466–77.
[9] Xu H, Wu L, Xiong S, Li W, Garg A, Gao L. An improved CNN-LSTM model-based state-of-health estimation approach for lithium-ion batteries. Energy 2023;276:127585.
[10] Zhou L, Lai X, Li B, Yao Y, Yuan M, Weng J, et al. State estimation models of lithium-ion batteries for battery management system: status, challenges, and future trends. Batteries 2023;9(2):131.
[11] Zhang C, Luo L, Yang Z, Zhao S, He Y, Wang X, et al. Battery SOH estimation method based on gradual decreasing current, double correlation analysis and GRU. Green Energy Intell Transp 2023;2(5):100108.
[12] Sylvestrin GR, Maciel JN, Amorim MLM, Carmo JP, Afonso JA, Lopes SF, et al. State of the art in electric batteries' state of health estimation with machine learning: A review. Energies 2025;18(3):746.
[13] Su L, Xu Y, Dong Z. State-of-health estimation of lithium-ion batteries: A comprehensive literature review from cell to pack levels. Energy Convers Econ 2024;5(4):224–42.
[14] Lanubile A, Bosoni P, Pozzato G, Allam A, Acquarone M, Onori S. Domain knowledge-guided machine learning framework for state of health estimation in lithium-ion batteries. Commun Eng 2024;3(1):168.
[15] Saha B, Goebel K. Battery data set, NASA prognostics data repository. NASA Ames Research Center, Moffett Field, CA; 2007.
[16] University of Oxford. Battery intelligence lab: data and code. 2022, https://howey.eng.ox.ac.uk/data-and-code/ (visited on 19/12/2024).
[17] Roman D, Saxena S, Robu V, Pecht M, Flynn D. Machine learning pipeline for battery state-of-health estimation. Nat Mach Intell 2021;3(5):447–56.
[18] University of Maryland. CALCE battery data. 2020, https://calce.umd.edu/battery-data#Storage (visited on 19/12/2024).
[19] Severson, et al. Data-driven prediction of battery cycle life before capacity degradation. Nature Energy 2019;4:383–91.
[20] MIT/Stanford University. Mit/stanford battery data. 2023, https://data.matr.io/1/.
[21] Zhao Z, Alzubaidi L, Zhang J, Duan Y, Gu Y. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. Expert Syst Appl 2024;242:122807.
[22] Hofmann T, Hamar J, Mager B, Erhard S, Schmidt JP. Transfer learning from synthetic data for open-circuit voltage curve reconstruction and state of health estimation of lithium-ion batteries from partial charging segments. Energy AI 2024;100382.
[23] Guo N, Chen S, Tao J, Liu Y, Wan J, Li X. Semi-supervised learning for explainable few-shot battery lifetime prediction. Joule 2024.
[24] Shen S, Sadoughi M, Li M, Wang Z, Hu C. Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries. Appl Energy 2020;260:114296.
[25] Fan G, Li J, Sun Z, Liu Y, Zhang X. Battery capacity estimation based on a co-learning framework with few-labeled and noisy data. J Power Sources 2024;600:234263.
[26] Qi Q, Liu W, Deng Z, Li J, Song Z, Hu X. Battery pack capacity estimation for electric vehicles based on enhanced machine learning and field data. J Energy Chem 2024;92:605–18.
[27] Li J, Chen W, Khalatbarisoltani A, Liu H, Lin X, Hu X. Tackling limited labeled field data challenges for state of health estimation of lithium-ion batteries by advanced semi-supervised regression (No. 2024-01-2200). 2024, SAE Technical Paper.
[28] Xiang Y, Fan W, Zhu J, Wei X, Dai H. Semi-supervised deep learning for lithium-ion battery state-of-health estimation using dynamic discharge profiles. Cell Rep Phys Sci 2024;5(1).
[29] Zhu J, Wang Y, Huang Y, Bhushan Gopaluni R, Cao Y, Heere M, et al. Data-driven capacity estimation of commercial lithium-ion batteries from voltage relaxation. Nat Commun 2022;13(1):2261.
[30] Lin C, Xu J, Mei X. Improving state-of-health estimation for lithium-ion batteries via unlabeled charging data. Energy Storage Mater 2023;54:85–97.
[31] Van Engelen JE, Hoos HH. A survey on semi-supervised learning. Mach Learn 2020;109(2):373–440.
[32] Belkin M, Partha N, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res 2006;7.
[33] Burkov A. In: The hundred-page machine learning book, vol. 1, Quebec City, QC, Canada: Andriy Burkov; 2019, p. 32.
[34] Chapelle O, Scholkopf B, Zien A. Semi-supervised learning. IEEE Trans Neural Netw 2009;20(3). 542-542.
[35] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on computational learning theory. 1998, p. 92–100.
[36] Zhu X. Semi-supervised learning with graphs. Carnegie Mellon University; 2005.
[37] Yang X, Song Z, King I, Xu Z. A survey on deep semi-supervised learning. IEEE Trans Knowl Data Eng 2022;35(9):8934–54.
[38] Zhang Y, Wen J, Wang X, Jiang Z. Semi-supervised learning combining co-training with active learning. Expert Syst Appl 2014;41(5):2372–8.
[39] Song Z, Yang X, Xu Z, King I. Graph-based semi-supervised learning: A comprehensive review. IEEE Trans Neural Netw Learn Syst 2022;34(11):8174–94.
[40] Sajun AR, Zualkernan I. Survey on implementations of GANs for semi-supervised learning. Appl Sci 2022;12(3):1718.
[41] Fujino A, Ueda N, Saito K. A hybrid generative/discriminative approach to semi-supervised classifier design. In: Proceedings of the National conference on artificial intelligence, vol. 20, (2). Menlo Park, CA; Cambridge, MA; London: AAAI Press; MIT Press; 2005, 1999, 2005.

[42] Triguero I, García S, Herrera F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowl Inf Syst 2015;42:245–84.
[43] Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Adv Neural Inf Process Syst 2020;33:596–608.
[44] ARule by the Environmental Protection Agency. 40 cfr 86.1815-27. 2024, https://www.ecfr.gov/current/title-40/part-86/section-86.1815-27.
[45] United Nations Global Technical Regulations (22) In-Vehicle Battery Durability for Electrified Vehicles. ECE/TRANS/180/Add.22. 2023, https://unece.org/transport/documents/2022/04/standards/un-gtr-no22-vehicle-battery-durability-electrified-vehicles.
[46] United Nations Economic Commission for Europe. Consolidated resolution on the construction of vehicles (ECE/TRANS/180/Add.22). 2023, https://unece.org/sites/default/files/2025-01/ECE-TRANS-180-Add.22-Amend.1-Appendix.1e.pdf.
[47] Office of energy efficiency and renewable energy: the official u.s. government source for fuel economy information. 2016, https://www.fueleconomy.gov/ (visited on 01/03/2024).
[48] Hofmann T, Hamar J, Rogge M, Zoerr C, Erhard S, Schmidt JP. Physics-informed neural networks for state of health estimation in lithium-ion batteries. J Electrochem Soc 2023;170(9):090524.
[49] Foroni D, Lissandrini M, Velegrakis Y. Estimating the extent of the effects of data quality through observations. In: Proceedings of the international conference on data engineering. ICDE, 2021, p. 1913–8.
[50] Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, Patzlaff H, et al. The effects of data quality on machine learning performance. 2022, arXiv:2207.14529.
[51] Chollet F. Deep learning with python. Simon and Schuster; 2021.
[52] Hu X, Xu L, Lin X, Pecht M. Battery lifetime prognostics. Joule 2020;4(2):310–46.
[53] Tian H, Qin P, Li K, Zhao Z. A review of the state of health for lithium-ion batteries: Research status and suggestions. J Clean Prod 2020;261:120813.
[54] Roy PK, Shahjalal M, Shams T, Fly A, Stoyanov S, Ahsan M, et al. A critical review on battery aging and state estimation technologies of lithium-ion batteries: prospects and issues. Electronics 2023;12(19):4105.
[55] Vasta E, Scimone T, Nobile G, Eberhardt O, Dugo D, De Benedetti, et al. Models for battery health assessment: a comparative evaluation. Energies 2023;16(2):632.
[56] Koehrsen W. A conceptual explanation of bayesian hyperparameter optimization for machine learning. 2018, Towards Data Science. https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f(visited on 08/03/2024).
[57] Scikit Learn. Compare the effect of different scalers on data with outliers. 2023, https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#plot-all-scaling-power-transformer-section (visited on 08/03/2024).
[58] Chen M, Du Y, Zhang Y, Qian S, Wang C. Semi-supervised learning with multi-head co-training. In: Proceedings of the AAAI conference on artificial intelligence, vol. 36, (6):2022, p. 6278–86.