Data-Driven Methods for Battery SOH Estimation:
Survey and a Critical Analysis

TSUYOSHI OJI¹, YANGLIN ZHOU², (Student Member, IEEE),
SONG CI², (Senior Member, IEEE), FEIYU KANG³, XI CHEN⁴,
AND XIULAN LIU⁴, (Member, IEEE)
¹Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 519071, China
²Department of Electrical Engineering, Tsinghua University, Beijing 100084, China
³Shenzhen Key Laboratory for Graphene-Based Materials, Engineering Laboratory for Functionalized Carbon Materials, Tsinghua Shenzhen International
Graduate School, Tsinghua University, Shenzhen 518055, China
⁴Beijing Electric Power Research Institute, Beijing 100075, China
Corresponding author: Song Ci (sci@tsinghua.edu.cn)

This work was supported in part by the Regional Key Research and Development Program of Inner Mongolia Autonomous Region in China under Grant 2020ZD00018, and in part by the Science and Technology Foundation of State Grid Corporation of Beijing under Grant 52022319005Q.

ABSTRACT State-of-health (SOH) estimation is a critical factor in ensuring the efficiency, reliability, and safety of lithium-ion batteries (LIBs) in electric vehicles (EVs). However, due to the complexity of electrochemical processes in batteries and the dynamics of working conditions, it is challenging to estimate SOH accurately, especially in real-world EV application scenarios. Thus, various data-driven methods with robust and adaptive features for SOH estimation have been widely proposed in the current literature. However, there is a lack of a comprehensive investigation and performance comparison of those methods, which makes them hard to be adopted in practice. Hence, in this paper, we have studied current major data-driven methods with real-world EV battery data to evaluate the performance. Besides, we summarize each method’s advantages and limitations with the consideration of the critical features required to achieve accurate SOH estimation in real-world applications. Hopefully, this paper provides a practical insight into the related fields.

INDEX TERMS Lithium-ion battery, state of health, data-driven methods, state estimation.

I. INTRODUCTION
With the merit of a long lifetime, high energy density, and fast response, lithium-ion batteries (LIBs) are widely used in electric vehicles (EVs) and energy storage scenarios [1], [2]. However, LIB performance declines over time (calendar aging) and use (cycle aging), which can lead to degraded performance, operational impairment, or even catastrophic consequences [3]. Since the complex internal electrochemical properties and uncertain external working environments, LIB degradation is an extremely complex process (as shown in Fig. 1), including physical mechanisms (e.g., thermal stress and mechanical stress) and chemical mechanisms (e.g., side reactions) [4], [5]. Generally, state of health (SOH) estimation is a critical metric in a battery management system (BMS) to quantify the extent of degradation. The most frequently used SOH indicator is battery capacity, which is defined as the ratio of current maximum available capacity \( C_i \) over the nominal value \( C_0 \) as \( \text{SOH} = \frac{C_i}{C_0} \times 100\% \).

With the gradually aging, when SOH reaches 70% – 80%, LIBs are more prone to thermal runaway and cause a safety risk [6]. An accurate and robust SOH estimation method ensures the safety, reliability, and cost-efficiency of a battery during operation [7].

In general, the methods for SOH estimation methods can be categorized into direct estimation and data-driven methods. Direct estimation usually includes Coulomb Counting and Impedance Spectroscopy [8]. In Coulomb counting, the integral of current is calculated with respect to time and divide by their difference of state of charge (SOC). It is a simple and widely used method, but its results are rather inaccurate, and the errors would accumulate [9]. Impedance spectroscopy applies a wide frequency spectrum to determine SOH [10]. However, the impedance spectroscopy method needs to carry out numerous experiments and requires adequate intermediate time to rest before the cell reaches its balanced potential.

Even though data-driven models have been applied in numerous SOH estimation processes, few available
investigations were focusing on the comprehension performance of each method. Andre et al. only compared the estimation performances of Neural Network (NN) and Extended Kalman Filter (EKF) for their accuracy and training speed [11]. Saha et al. employed Autoregressive Integrated Moving Average (ARIMA), EKF, Particle Filter (PF), and Relevance Vector Machine (RVM) to estimate the remaining useful life (RUL) for LIBs [12]. Each method’s algorithm is specifically introduced, but the estimation performance results were only presented in figures without a quantitative comparison. In [8], [13]–[16], various data-driven methods were studied for their pros and cons, and the methods were classified into different categories but lacking an objective performance comparison.

To address the aforementioned challenges, this paper makes the following main contributions:

- Review the major data-driven SOH estimation approaches for LIBs reported in the recent literature.
- Compare the performance of three model-based models and four model-less models, namely EKF, PF, ARIMA, Extreme Learning Machine (ELM), Long Short-term Memory (LSTM), Support Vector Machine (SVM), and RVM, on a real-world EV dataset.
- Provide an overall discussion about the aforementioned models in terms of their accuracies, confidence intervals, abilities to deal with nonlinearity, robustness, computation complexities, capabilities to deal with data sparsity, and generalization.

The remainder of this paper is organized as follows: For different data-driven methods, Section II provides a short theoretical explanation, their challenges, and a literature review focused on SOH estimation application. In Section III, the real-world operation EV data are shown, and the comparison experiment of previously described methods is explained in detail. The comparison result is analyzed and discussed in Section IV. In Section V, the different models previously described are discussed for actual application requirements. Section VI presents the conclusion drawn from this paper. For the reader’s convenience, Tab1 lists all the acronyms used in this article in alphabetical order.

II. REVIEW OF DATA-DRIVEN ESTIMATION METHODS

The data-driven approach is a method that builds a rough model and then refines the model with numerous data to make the model consistent with the data. If the initial model is an existing battery model, it is classified as a model-based method; otherwise, it is a model-less method.

The existing battery model in model-based methods usually involves Equivalent Circuit Model (ECM) and electrochemical model. ECM model uses appropriate circuit components to constitute an equivalent circuit, and the parameters of the circuit model can only obtain under laboratory conditions and will change through battery aging. The electrochemical model is to study the electrochemical process through battery aging, and its strength resides on no laboratory measurement required. However, developing a detailed mathematical model including phase-changing typically requires cell disassembly [17].

On the other hand, model-less methods can avoid analyzing the complex electrochemical reaction and directly use machine learning approaches to estimate the aging process. Such methods do not need prior knowledge of battery type and working conditions. The accuracy of the estimation largely depends on the training data size. Nevertheless, in most machine learning methods, we need to select a set of external characteristics that can best represent SOH, which may bring subjective factors into the estimation process.
### TABLE 1. List of acronyms and abbreviations.

| Acronym | Description                                      | Acronym         | Description                     |
|---------|--------------------------------------------------|-----------------|---------------------------------|
| ACF     | Autocorrelation Function                         | MCMC            | Markov Chain Monte Carlo        |
| ANN     | Artificial Neural Network                        | MLP             | Multi-layer Perceptron          |
| AR      | Autoregressive                                   | MSE             | Mean Square Error               |
| ARD     | Automatic Correlation Determination              | NN              | Neural Network                  |
| ARIMA   | Autoregressive Integrated Moving Average         | BMS             | Battery Management System       |
| PACF    | Partial Autocorrelation Function                 | CALCE           | Center for Advanced Life Cycle |
| PCC     | Possibilistic Clustering Classification          | CNN             | Convolutional Neural Network    |
| PCoE    | Prognostics Center of Excellence                 | CPSO            | Continuous Particle Swarm Optimization |
| PDF     | probability density function                     | DCNN            | Deep Convolutional Neural Network |
| PF      | Particle Filter                                  | DPSO            | Discrete Particle Swarm Optimi |
| PSO     | Particle Swarm Optimization                      | EAPs            | Electrode Aging Parameters      |
| RBF     | Radial Basis Function                            | ECM             | Equivalent Circuit Model        |
| RMSE    | Root Mean Square Error                            | EKF             | Extended Kalman Filter          |
| RMSProp | Root Mean Square Prop                            | HI              | Health Indicator                |
| RNN     | Recurrent Neural Network                         | ELN             | Extreme Learning Machine        |
| RUL     | Remaining Useful Life                            | ETV             | Electric Vehicles               |
| RVM     | Relevance Vector Machine                         | FFNN            | Feed Forward Neural Network      |
| SampEn | Sample Entropy                                    | GBR             | Gradient Boosted Regression     |
| SLFN    | Single Hidden Layer Feed Forward Neural Networks | GPR             | Gaussian Process Regression     |
| SMC     | Sequential Monte Carlo                           | HI              | Health Indicator                |
| SOC     | State of Charge                                  | LIB             | Lithium-ion Battery            |
| SOH     | State of Health                                  | LSTM            | Long Short-term Memory          |
| SVM     | Support Vector Machine                           | MAE             | Mean Absolute Error             |
| SVR     | Support Vector Regression                        | MAPE            | Mean Absolute Percentage Error  |
| UKF     | Unscented Kalman Filtering                       | MC              | Monte Carlo                     |
| UPF     | Unscented Particle Filter                         |                 |                                 |

### TABLE 2. Different capacity degradation models.

| No. | Selected model                                                                 |
|-----|--------------------------------------------------------------------------------|
| 1   | \( C = a \times k^2 + b \times k + c \)                                      |
| 2   | \( C = a \times \exp(b \times k) + c \times k^2 + d \)                        |
| 3   | \( C = a \times \exp(b \times k) + c \times k^2 + d \times k + e \)          |
| 5   | \( C = a \times \exp(b \times k) + c \times \exp(d \times k) \)             |
| 6   | \( C = \frac{a \times C_1}{b + c} \)                                         |

where \( a, b, c, d, e \) are the parameters in those models; \( k \) is the number of cycles; \( C_1 \) in No.6 is a constant through empirical test.

### A. MODEL-BASED METHODS

As mentioned earlier, model-based methods usually contain an ECM or electrochemical model. ECM model is mainly used to confirm SOC and difficult to estimate the remaining capacity. Hence, the electrochemical model is adopted to estimate SOH. Table 2 shows different electrochemical models for SOH estimation [18].

The model parameters can be determined through adaptive filter methods. The main concept is to filter the measurement noise so it can update the model parameters with new measurements. Kalman Filter (KF) and PF are two common methods adopted for this purpose.

1) KALMAN FILTER (KF)

KF is a statistical-based filtering method proposed in 1960. Through repeated iterations of the previous estimate and the current measurement values, a relatively accurate value can be derived. The state equation in KF is used to describe the state process of the system based on the prior information, and the measurement value obtained by the external observation system is described by using the measurement equation [19].

However, the standard KF cannot solve the nonlinear degradation model, so there are some improved algorithms such as EKF and Unscented Kalman Filtering (UKF).

The basic principle of KF is to use the expansion of the Taylor series to linearize the nonlinear equation and then to solve the linearized equation using the KF framework. Therefore, it may be more suitable for battery state estimation. Plett et al. showed that although EKF was usually used in SOC estimation in the past decades, it may also be used to estimate power fade and can keep the SOC estimate accurate throughout the cell lifetime even though its dynamics changing as it ages [20]. Before using the differential quotient to calculate the state matrix and the measurement matrix, Zhou et al. combined Gaussian Process Regression (GPR) with EKF to approximate the state equation, the measurement equation, and the noise equation of EKF [21].

In addition, UKF [22] was proposed in estimation process at [23] and [24]. The advantage of UKF is that there is no specific form of the nonlinear equation, so there is no demand for the derivative and Jacobian matrix calculation.

2) PARTICLE FILTER (PF)

In PF, the particles are generated and recursively updated from a nonlinear process that involves a system under analysis, a measurement model, and a priori estimate of the state probability density function (PDF) [25]. That is to say, using Monte Carlo (MC) simulations, PF is a method for implementing a recursive Bayesian filter, and this is also known as the Sequential MC (SMC) method.

With regard to PF-based methods, Su et al. divided the model into three categories: polynomial model, exponential model, and Verhulst [26] model, to compare their
performance [18]. Miao et al. presented an improved PF algorithm, the unscented particle filter (UPF), which combined the idea of PF and UKF to improve the RUL prediction accuracy [27]. UPF can be divided into two steps: firstly, the UKF method was used to get the proposal distribution; secondly, the standard PF method was applied to get the final results. Zhang et al. proposed an improved UPF based on the Markov chain Monte Carlo (MCMC) method, in which, after resampling in UPF, MCMC was adopted to approximate the estimated state [28]. Therefore, it can maintain the particle’s diversity and suppress particle degradation to a certain extent. Some other work also tried to solve the importance function chosen and degradation of diversity in sampling particles problems, using the linear optimization approach to produce new particles from chosen particles and abandoned particles [29].

3) AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA)

ARIMA requires only the historical time series data. This model is fitted to time series data to forecast future points in the series. Therefore, ARIMA is more suitable for single-step estimation instead of multi-step prediction.

In an ARIMA model, the choice of its parameters is usually subjective. In this regard, Long et al. used a high-order Autoregressive (AR) model to replace the ARIMA model and transformed the problem of seeking nonlinear parameters of ARIMA into seeking linear parameters in the AR model [30]. Furthermore, he proposed the Particle Swarm Optimization (PSO) algorithm to avoid the uncertainty of human subjective order determination. The data metabolic technology is also used to make the AR model change adaptively.

In order to improve the accuracy in the long-term prediction process, Liu et al. introduced an influence factor to characterize ‘accelerated’ degradation and combined this influencing factor with the AR model [31]. In his following work [32], he introduced a regularized particle filter to further improve the prediction accuracy.

B. MODEL-LESS METHODS

Instead of considering the electrochemical reaction and the failure mechanism inside batteries, the model-less methods, which require no explicit battery models and regard the battery system as a black box, and then infer battery SOH or lifespan directly from extracted features. In literature, many statistical, computational, and artificial intelligence algorithms and models, such as Artificial Neural Network (ANN) [33]–[38], SVM [39], [40], RVM [41], [42], GPR [43], Gradient Boosted Regression (GBR) [44], [45], have been adopted for battery state estimation in various applications. However, data-driven techniques are usually unstable as they may show different performances with different datasets [46]. Among all these model-less methods, ANNs and SVMs are regarded as the two most representative ones for nonlinear modeling [47].

1) ARTIFICIAL NEURAL NETWORK (ANN)

ANN is intended to imitate the human brain’s behavior, with artificial neurons arranged at the input layer, hidden layer(s), and output layer, respectively. The ANN topology is illustrated in Fig. 2. The input layer gathers the preprocessed information and acts as a conduit to the hidden layer(s). Each neuron can be represented by a weighted linear combination and contains a mathematical model based on its input for determining its output in the hidden layers [48]. The ANN-based method has long been used for modeling as they provide automated knowledge extraction and high inference accuracy if a sufficient amount of operation data is used for model training [49]. There are many kinds of network selection methods. A typical type having been successfully applied for SOH estimation is Feed Forward Neural Network (FFNN), also known as Multi-layer Perceptron (MLP), which is usually trained by the back-propagation algorithm.

So far, it is generally believed that the internal resistance is the most representative feature of the remaining useful capacity of the battery. Xia et al. used FFNN to prove the relationship between the complex impedance zero-phase crossing frequency of a battery and its SOH [33]. However, the measurement speed of internal resistance is usually slow, which is almost impossible to realize for online applications. Zhang et al. proposed an online method for SOH and RUL monitoring based on the fusion of partial incremental capacity and FFNN [34]. After smoothed the initial partial incremental curve and carried out the Spearman correlation analysis, two strongly correlated features were extracted from the partial incremental curve as input, and then two FFNN models for simultaneous estimation of SOH and RUL were established, leading to a simple model structure and the satisfactory accuracy and generalization performance.

However, ANNs usually suffer from slow training speed and high computational requirements. Therefore, a kind of fast learning model called ELM has been proposed for onboard estimation.
2) EXTREME LEARNING MACHINE (ELM)
Different from other ANNs, the connection weight between the input layer and hidden layer and the threshold of the hidden layer can be set randomly, with no adjustment required after setting. Moreover, the connection weights between the hidden layer and the output layer can not be adjusted iteratively but determined by solving the generalized inverse matrix [50]. Thus, compared with the traditional neural networks, especially with the Single Hidden Layer Feed Forward Neural Network (SLFN), it delivers faster performance than traditional learning algorithms in the premise of ensuring learning accuracy.

Pan et al. used the Thevenin equivalent model to calculate the ohmic internal resistance and polarization internal resistance of the battery with easily measured terminal voltage, load current, and ambient temperature [35]. The increment of the two resistance values was taken as the health factor, and the ELM method was used to estimate the battery life online. Compared with the traditional FFNN, the results showed that the estimation error is significantly reduced with a faster training speed.

3) DEEP LEARNING NETWORK
With multi-layer perceptron and hidden layers, the concept of deep learning is originated from the ANN and rising in recent years. Shen et al. first applied the deep learning method to the online capacity estimation of Li-ion batteries [36]. He utilized a deep convolutional neural network (DCNN) for the battery capacity estimation based on the voltage, current, and charge capacity measurements during a partial charge cycle. The proposed structure successfully avoids the manual feature extraction process, which has the risk of dropping useful information. Tian et al. also construct a convolutional neural network (CNN) to estimate electrode capacities and initial SOCs, termed electrode aging parameters (EAPs) [51].

Recurrent Neural Network (RNN) is another type of the deep learning methods that have certain advantages in learning nonlinear features of sequences. With the assumption that the attenuation of battery capacity is continuous in time, Eddahech et al. demonstrated an RNN model to predict the remaining capacity and internal resistance through the collected SOC difference, pulse current, temperature, and three latest-predicted internal resistance values [37].

However, RNN suffers from learning long-term dependencies. If RNN stores information over a period of time, the network gradient tends to vanish, meaning that the network is unable to learn anymore. Zhang et al. synthesized a data-driven battery RUL predictor by using Long Short-term Memory Recurrent Neural Network (LSTM-RNN) [38]. The Root Mean Square Prop (RMSProp) method for small batch training data samples was used to train the constructed neural network [52], and a rejection technique was proposed to solve the over-fitting problem [53]. The results showed that with similar accuracy, compared with other methods [18], [54]–[57], the number of offline training data samples can be reduced by 20% - 50%.

4) SUPPORT VECTOR MACHINE (SVM)
The main concept of SVM is to find a small set of support vectors out of a large number of data samples, which can still describe the system. SVM has been successful in a wide range of applications, especially for nonlinear problems with small samples, and can effectively prevent local minimization. In theory, there is a global optimum and can avoid the defect of the local extremum. Nevertheless, it is sometimes troublesome to determine the optimal kernel function and hyperparameters for nonlinear modeling [47].

In [39], PSO was employed to obtain the Support Vector Regression (SVR) kernel parameter. By adopting a fresh validation method, the fusion PSO–SVR model can well grasp the global degradation trend of SOH with little interference from local regenerations and fluctuations. Tao et al. imbedded the PF method into the SVR paradigm to optimize the hyperparameters due to its ability to update the parameters dynamically, providing the PDF of the optimal parameters [40]. Possibilistic Clustering Classification (PCC) was also induced to cluster different operational states to clusters and then estimate through their belonging model.

Another drawback of SVM is lacking the ability to output the confidence intervals, so the variation model, RVM, is widely adopted.

5) RELEVANCE VECTOR MACHINE (RVM)
RVM is also a supervised learning method similar to SVM. It is based on the Bayesian framework theory, which can eliminate the irrelevant points through the Automatic Correlation Determination (ARD) and then derive the sparse model. Compared with SVM, RVM can construct any kernel function without the restriction of the Macy’s theorem. In terms of parameters, SVM needs to be initialized manually, and different values may have a great influence on the results, while RVM can be operated automatically.

However, RVM involves a combination of kernel functions, and determining the weight of each function is vital to the performance. In [58], Yang et al. proposed a fusion method by using the Discrete Particle Swarm Optimization (DPSO) [59] algorithm for selecting kernels adaptively, and Continuous Particle Swarm Optimization (CPSO) [60] to adjust kernel combinations as well as kernel parameters adaptively.

In order to find an alternative online health indicator (HI) to quantify the battery degradation, Zhou et al. chose the mean voltage falloff as a HI in each charging and discharging cycle [42]. After the extraction, Box-Cox transformation was adopted to enhance the degree of linear correlation between the HI and the capacity. Then the author compared the performance of the simple statistical regression model and the RVM model. The result indicated that the proposed RVM model was more accurate than the simple statistical regression model and with the confidence interval.

To verify the SVM and RVM performance, Widodo et al. used the Sample Entropy (SampEn) [61] as a HI to estimate
TABLE 3. A synopsis of the reviewing results for the data-driven estimation methods.

| Method | Reference | Description of experiment dataset | Estimation precision | Novelty |
|--------|-----------|-----------------------------------|----------------------|---------|
| KF     | [20]      | 7.5 Ah & 5 Ah LiMn$_2$O$_4$/Graphite | the estimation converges to the correct values | Using dual and joint EKF to estimate power fade and capacity fade |
|        | [21]      | Battery pack on electric bus      | Error <8%             | Proposed an iterated GPR-EKF algorithm for online in-use EV SOH estimation |
|        | [24]      | 18650-type lithium battery        | Error <0.1%           | Designed a Double Unscented KF to calculate SOC and SOH at the same time |
| PF     | [18]      | A1, A2, A3, and A4 from CALCE$^*$ | accurate estimation  | Proposed an interacting multiple model particle filter for estimating the RUL with an improved battery capacity degradation model |
|        | [27]      | A1, A2, A3, and A4 from CALCE     | Error <5%             | Introducing the unscented particle filter into the battery RUL prediction |
|        | [28]      | A1, A2, A3, and A4 from CALCE     | Error 2.04%           | Proposed an improved unscented particle filter for LIB RUL prediction based on Markov chain Monte Carlo |
|        | [29]      | #5, #6, #7, and #18 from PCoE$^b$ | Error ≤ 4.6%          | Using linear optimizing combination resampling to improve the prediction accuracy of the unscented particle filter |
| ARIMA  | [30]      | A1, A2, A3, and A4 from CALCE     | Error <14.6%          | Using particle swarm optimization to improve the AR model |
|        | [31]      | #5, #6, #7, and #18 from PCoE and CS2-8, CS2-21, and CS2-33 from CALCE | MAE$^c$ <10% in most of the cases | Introducing an optimized nonlinear degradation AR time series model for RUL estimation |
| ANN    | [33]      | TenneT 3000X-0 LIB                 | N/A                   | Presenting a SOH estimation method based on the complex impedance zero-phase crossing frequency of a battery |
|        | [34]      | #5, #6, #7, #18 from PCoE         | MAE in range of 2.79-3.52 | Proposed a novel on-line synthesis method based on the fusion of partial incremental capacity and ANN to estimate SOH and RUL |
| ELM    | [35]      | ICR 1850-26P from Samsung, LiNiMnCo$_{1/3}$O$_2$/Graphite | Maximum Error 2.22%, MAB 1.72% | Introducing an ELM to capture the underlying correlation between the internal resistance and capacity degradation to improve the speed and accuracy for SOH estimation |
| CNN    | [36]      | 20 cells from PCoE and 10-years cycling tests on eight Li-ion cells | Maximum Error <9.47% | Utilizing DCNN for battery capacity estimation based on the voltage, current, and charge capacity measurements during a partial charge cycle |
|        | [51]      | eight Kokam pouch, LiNiCoAl$_x$O$_2$/Graphite | Error <1.6%           | Using CNN to estimate electrode capacities and initial SOCs |
| RNN    | [37]      | LiNi$_{0.5}$Co$_{0.5}$Al$_{0.5}$O$_2$/Graphite | MSE$^d$ 0.462         | Using RNN to monitor the SOH |
| LSTM   | [38]      | NCR 18650PF from Panasonic, LiNiCoAl$_x$O$_2$/Carbon | Error <11.6%          | Using LSTM and RNN to predict the RUL, and a dropout technique is adopted to address the overfitting problem |
| SVM    | [39]      | #5, #6, #7, and #45 from PCoE     | MAPE$^e$ 0.82%-2.28%  | Using particle swarm optimization to obtain the SVR kernel parameter for SOH estimation |
|        | [40]      | one Li-ion battery from PCoE      | accurate estimate     | Using the probabilistic clustering classification to determine the system, particle filtering to select SVR parameters according to the system state, and predicate remaining useful cycle through SVR |
| RVM    | [41]      | #18 and #28 from PCoE             | RMSE$^f$ 0.54         | Using SampEn as a HI to estimate SOH by using RVM |
|        | [42]      | #5 from PCoE                      | RMSE 0.0135           | Choosing the mean voltage falloff as a HI in each charging and discharging cycle, and estimating the capacity through RVM model |

$^a$ Batteries data from NASA Ames Prognostics Center of Excellence. $^b$ Batteries data from the Center for Advanced Life Cycle Engineering at the University of Maryland. $^c$ Mean Absolute Error. $^d$ Mean Square Error. $^e$ Mean Absolute Percentage Error. $^f$ Root Mean Square Error.

SOH through both methods [41]. The results showed that RVM outperforms SVM-based battery health prognostics in the aspect of accuracy.

So far, the primary data-driven estimation methods in the literature have been reviewed, and a detailed synopsis of the estimation results is provided in Table 3. As can be seen from the table, the dataset is vital to the estimation performance. Most of these studies used the publicly available data from NASA Ames Prognostics Center of Excellence (PCoE) [62], or the Center for Advanced Life Cycle Engineering (CALCE) at the University of Maryland [56]. However, even with the same dataset, some articles proposed abnormally highly accurate estimation results, which may bring confusion to actual implementation. Therefore, it is necessary to evaluate the performance of major data-driven methods on a real-world dataset with the same evaluation indices.

III. A COMPREHENSIVE AND QUANTITATIVE STUDY WITH REAL-WORLD DATA

As mentioned earlier, different types of models exhibit controversial results since they had been tested with different data sets. Therefore, in this work, we have chosen the aforementioned major data-driven methods for battery SOH estimation
A. REAL-WORLD BATTERY OPERATION DATA

The real-world data adopted in this work was obtained from six electric buses running in Yingshang County, Fuyang City, Anhui Province in China. An electric bus is shown in Fig. 3. All buses are loaded with lithium iron phosphate battery packs produced by Hefei Gotion High-tech Power Energy CO., Ltd.

BMS units are used for logging all battery operate data, including voltage, current, temperature, and 20 other parameter values. The data logging interval is 30 seconds, and the time frame of the recorded data is from January 2017 to June 2019. The specifications of the electric bus and the battery pack data are summarized in Table 4.

Due to the dynamics of discharging process, which makes it difficult to calculate the remaining useful capacity during the discharging process. However, the charging current remains at a relatively stabilized value. Hence, the charge capacity of a battery pack can be expressed as:

$$Q_i = \int_{t_1}^{t_2} I dt$$  \hspace{1cm} (1)

where $Q_i$ stands the total capacity charged in $i$ cycle. $t_1$ and $t_2$ represents the initial charging time and ending time, respectively. Accordingly, the available capacity of the battery pack was:

$$C_i = \frac{Q_i}{(SOC_{end} - SOC_{star})} \times 100.$$  \hspace{1cm} (2)

As shown in (2), the denominator will approach zero when the difference of SOC is small. Hence, we only consider the deep charging cycle (SOC difference $\geq$ 30%).

As we mentioned in Section I, since SOH quantifies the degradation degree of LIBs, it will provide guidance for battery replacement [63]. Most current research works use characterization parameters of battery aging (e.g., capacity, internal resistance, and power) to define SOH [64]. Among those parameters, internal resistance can only be acquired in the laboratory condition, and power measurement makes the computation more complex. Hence, for convenience in real-world applications and maximum efficiency, SOH in this paper is defined as the ratio of current maximum available capacity $C_i$ over the nominal value $C_0$, as:

$$SOH = \frac{C_i}{C_0} \times 100\%$$  \hspace{1cm} (3)

In this study, four hundred deep charging cycles were selected for each bus. After ranking the absolute percentage error between each data point and the average value in the corresponding cycle, the data points with the highest five percentage errors were regarded as outliers. Their curve diagrams and the statistical results are shown in Fig. 4 and Table 5, respectively. It can be seen that some onboard measurement data points show large deviations, and the calculated remaining useful capacity dropped about 8% at around the hundredth cycle for all six datasets. Therefore, an appropriate data-driven estimation model must be able to reject those outliers.
B. EXPERIMENTAL SETTINGS

For the data-driven methods mentioned above, some data-driven models can not deal with the time series directly, so it's necessary to use other features to train them. In addition, the optimization algorithms can not automatically optimize some parameters called hyperparameters, which need to be set manually. Therefore, the following is a description of the selected features and hyperparameters of the models used in this paper. These hyperparameters are determined through cross-validation, but there is no guarantee that all parameters are the globally optimal values.

1) EXTENDED KALMAN FILTER

By using the No.6 capacity degradation model listed in the Table 1 as the measurement equation and the four parameters adding noise as the estimation equation, EKF iterates through the Kalman gain and updates four parameters of the model when new measurement data are available. The four parameters and covariance matrix are initialized to 
\[[0.5, 0, 0.5, 0]\] and 
\[[1; 0.005; 1; 0.005]\], respectively. We have also adopted the one-step estimation method, i.e., estimating the next cycle SOH from the existing model, because the radical change SOH in real-world applications may not be estimated by a fixed model and then need to be updated in a real-time fashion.

2) PARTICLE FILTER

The PF model is developed with the same measurement equation and initial parameters as used in EKF. One-step estimation is adopted as same as used in EKF.

3) AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

The ARIMA model is possible to handle the non-stationary series of data if the series of data can achieve stationary by differentiating it to a sufficient degree. Therefore, the parameters of the ARIMA model, including the differentiating times, 'd', autoregressive terms, 'p', and moving average terms, 'q', can be determined through the autocorrelation function (ACF) and the partial autocorrelation function (PACF). In our case, the model derives the best estimation performance when (p, d, q) is set to 
\'4,1,2\'.

4) EXTREME LEARNING MACHINE

Different from the model-based algorithms, model-less methods estimate SOH through observing a few external feature values. In this paper, five values - starting SOC, ending SOC, starting voltage, ending voltage, and average charging current - extracted from the battery charging data sets are used to train the ELM model. In our case, we set the activation function as the hyperbolic tangent activation function and a single hidden layer with 30 nodes.

5) SUPPORT VECTOR MACHINE

For the SVM model, the five values mentioned above in ELM are used to train this model, the penalty coefficient is 100, and the kernel function is radial basis function (RBF) with a variance of 16.67.

6) RELEVANCE VECTOR MACHINE

For the RVM model, the aforementioned five values in ELM are also used to train the RVM model, and the kernel function is RBF.

7) LONG SHORT-TERM MEMORY

LSTM cannot directly process the time series, so the time series obtained by a sliding window is used to train the LSTM model. As shown in Fig. 5, previous \(N-1\) cycles are used to form the observation matrix, with the SOH of the subsequent cycles as the target value [65]. In our case, the size of the sliding window is ten, and the neuron in the single hidden layer is four.

The experiments are conducted on a computing platform with Intel Core i7-6700K processor at 4.0 GHz using 32 GB of RAM, running Win10 Pro version.

IV. PERFORMANCE EVALUATION AND ANALYSIS

To compare the performance of the aforementioned data-driven models, 10% of one dataset preprocessed as described in Section III-A are used for training, and the remaining 90% are used for testing. The corresponding estimation results are shown in Fig. 6(a). It can be seen that the SOH estimation results for EKF, PF, and LSTM models are closer to the real SOH than ARIMA, ELM, SVM, and RVM model results.
The fitting curve for ARIMA even shows an upward trend in the rest of the life cycle, indicating a large estimation error contained within. The error mainly comes from the outliers from the training data because ARIMA is not equipped with any mechanism to reject those outliers. After training, the intrinsic structure was fixed, and the estimation results were only related to the cycle index.

The estimation results for all the models mentioned above by using 40% data samples for training and 60% data samples for testing are shown in Fig. 6(b). It can be found that the results of EKF, PF, and LSTM methods are still closer to the real SOH than others. Specifically, the fitting curve for ELM shows a large deviation in the beginning, which is similar to what ARIMA presents by using 10% data samples for training. However, ELM failed to estimate SOH because it could not deal with data sparsity. The single hidden layer and the non-iterative training process in ELM would lead to underfitting with small training samples.

The estimation results for all aforementioned data-driven models by using 70% data samples for training and 30% data samples for testing are shown in Fig. 6(c). As is observed from the figure, the estimation results are all close to the real SOH, which validates the effectiveness of all the aforementioned models with the substantial number of training data samples. In this sense, confidence intervals for the outputs are more significant than their accuracy. The estimation results of RVM and SVM by using 70% data samples for training are compared with the real values and presented in Fig. 7. RVM is advantageous because it can output confidence intervals, and most of the real values reside in the 95% confidence interval of estimation results.

To further discuss the impacts of different size of training set on SOH estimation performance, we apply three metrics to evaluate the aforementioned models. The first metric is the mean absolute percentage error (MAPE) to evaluate the general accuracy, which are defined as follows:

$$MAPE = \frac{1}{m} \sum_{i=1}^{m} \left| \frac{y_i - \hat{y}_i}{\hat{y}_i} \right| \times 100\%$$ (4)
where $m$ is the total number of data samples; $y_i$ and $\hat{y}_i$ are the estimated and real values of cycles $i$, respectively. The second metric is the time to train each model, and the third metric is the computational time to complete SOH estimation. The comparison of three metrics under different training set for each model by using one dataset are presented in Fig. 8(a), Fig. 8(b), and Fig. 8(c), respectively.

In Fig. 8(a), the MAPE of PF and EKF are lower than other methods across all sizes of training samples, while LSTM presented a similar accuracy after 50% of the data samples are used for training. The MAPE of ELM and ARIMA are over 7.5% before adopting 30% of data samples for training, indicating that both methods need substantial training samples to avoid underfitting. In addition, SVM outperforms RVM after using 60% data samples for training.

Fig. 8(b) and Fig. 8(c) show the average training and computation time of different models with varied training samples, respectively. The training and computation steps were interlaced in EKF and PF models so that the time results are not shown in Fig. 8(b) and Fig. 8(c). Among other data-driven models, the training speed of ELM is the fastest. For the rest of the models, the training speed of ARIMA is almost equal to that of SVM or RVM, whereas the LSTM model is more than ten times slower than others. However, although the computation times between LSTM and other methods are still vastly different by ten times, the absolute difference is only within 0.005s, which is tolerable in actual practice.

To find out a more generalized conclusion, the average MAPE, training time, and computation time results by using all the six datasets are shown in Fig. 8(d), Fig. 8(e), and Fig. 8(f), respectively. Each dataset is tested ten times and records to find the average result so as to reduce randomness to the utmost extent. The results are graphically depicted in Table 6. It can be seen that PF shows the best average accuracy, followed by EKF, LSTM, SVM, RVM, ARIMA, and ELM. However, ELM is the fastest model for training on average, followed successively by ARIMA, SVM, RVM, and LSTM. Additionally, the average computation speed of SVM is the fastest, followed by ARIMA, RVM, ELM, and LSTM. These rankings are similar as reported.

The above comparison results indicate that EKF and PF exhibit a relatively high accuracy (MAPE <2.5%) by using any proportion of data samples for training. In another word, as long as an electrochemical model and its initial parameters are correctly chosen, these adaptive filters could accurately estimate SOH without extensive operating data. However,
this conclusion can only stand if high-precision data measurement devices are applied. Another model-based method, ARIMA, is somewhat on the opposite. Unlike EKF and PF, it needs a substantial set of data to fit with a time series model, which makes it more suitable for short-term prediction.

On the other hand, without requirements on prior knowledge, the ‘black box’ framework adopted in model-less methods would produce results that are entirely dependent on the training samples. Among these models, ELM is the fastest in training speed because no iteration was involved in its training. However, it suffers from low accuracy and the incapability to reject the outliers. On the opposite, LSTM demonstrates the best accuracy with the longest training time. SVM is superior to RVM in terms of accuracy, as well as shorter training and computation time. Despite that, RVM can output confidence intervals, which makes it one of the most prospective models in the future.

V. FURTHER DISCUSSION

Application categories and resource limits are determined in which cases a data-driven method can be applied. Thus, different elements should be viewed when choosing the appropriate method for a certain context [66]. In this section, some aspects that need considerations in future applications of LIB aging estimation will be discussed.

A. ACCURACY

The accuracy of a data-driven method is one way to indicate how successfully it fulfills its goal by a fair metric. For instance, EKF proposed in section IV was evaluated by MAPE, which was shown to be less than 3%. In this case, we can say it is a highly accurate method.

B. CONFIDENCE INTERVAL

When applying data-driven methods, the bias-variance trade-off needs to be considered [67]. Those methods that specialize in error minimization are possible to end up with the overfitting issue. Therefore, instead of getting rid of errors, a better result should give a confidence interval within which the true value is located. In this regard, an ideal estimation model should be probabilistic so that it can provide a range of values to represent the estimation results with a specified confidence level.

Among all the methods reviewed, RVM employs a probabilistic Bayesian framework, while PF is based on Bayesian filtering and Monte Carlo simulation. Thus, both methods are capable of producing confidence intervals as their output.

C. ABILITY TO DEAL WITH NONLINEARITY

SOH recession is a strongly nonlinear process, so the capability to model nonlinear relations is crucial for data-driven methods. Both EKF and PF can take nonlinear equations for measurement and transition. Besides, ANN and SVM frameworks also allow nonlinear regressions. However, ARIMA is one of the linear regression models. Consequently, limitations could be seen in practical applications.

D. ROBUSTNESS

A severe limitation for current data-driven methods is that their development has been limited to aspects about measurement accuracy. In the literature, such measurement data were considered as known and accurate [68]. Nevertheless, when estimations are performed in practice, such data typically contain a certain amount of noise. This inaccurate measurement can lead to a huge estimation error, resulting in misleading conclusions. Hence, it is of practical importance to build a robust model against noises.

SVM and RVM involve robust mechanisms to deal with small data fluctuation and aberration. In fact, they are engaged with an inherent sparse mechanism that allows them to neglect small data variation, and RVM can even discard irrelevant data [68]. Besides, EKF is also competent in terms of noise exclusion. On the other hand, ARIMA and ELM do not include any comparable mechanism to reinforce estimation robustness. Furthermore, in PF, the outliers will also cause filter divergence, leading to unwanted estimation performance [25].

E. COMPUTATION COMPLEXITY

The computation complexity is evaluated as the resources required by a data-driven method to run. In particular, it focuses on their time (amount of time it takes to run an algorithm) and memory (amount of memory space required to solve an instance) requirements. The determination of a model’s computational complexity is useful because by this way, we can (i) decide whether a part of the assignments should be carried out online or offline, (ii) distribute storage...
space in a more effective manner, and (iii) suggest modifications that would improve the computation results [66].

Since the computation complexity is generally difficult to quantify, one common representation of it is the asymptotic behavior expressed by big-$O$ notations, which characterize functions according to the correlation between run time or space requirements and the input size in a data-driven method.

Among the aforementioned models, the complexity of PF is independent of the state dimension and increases by the function of the particle number ($N_p$). The corresponding time complexity is $O(N_p)$ [68]. The computational complexity of ARIMA was found to depend on its order. In this sense, its big-$O$ complexity cannot be determined. The computational complexity of the ANN framework model is varied based on the number of training samples, input dimensions, hidden units, and outputs [69].

| F. CAPABILITY TO DEAL WITH DATA SPARSITY | Models | Model-based | Model-less |
|------------------------------------------|--------|-------------|------------|
| accuracy                                 | EKF    | high        | low        |
|                                          | PF     | high        | low        |
|                                          | ARIMA  | low         | medium     |
|                                          | ELM    | low         | medium     |
|                                          | SVM    | medium      | medium     |
|                                          | RVM    | medium      | high       |
|                                          | LSTM   | high        | medium     |
| confidence interval                      | ×      | ✓           | ×          |
|                                        | ✓      | ×           | ✓          |
| ability to deal with nonlinearites      | ✓      | ✓           | ✓          |
|                                        | ×      | ✓           | ✓          |
| robustness                              | ✓      | ✓           | ✓          |
|                                        | ✓      | ✓           | ✓          |
| capability to deal with data sparsity   | high   | high        | high       |
|                                        | low    | low         | high       |
|                                        | low    | medium      | medium     |
|                                        | low    | medium      | high       |
|                                        | medium | high        | medium     |

*Depending on the number of particles. *Big-O* complexity cannot be determined. *Depending on the input dimension, the number of training samples, the number of hidden units, and the number of outputs.

The characteristics of the aforementioned SOH estimation methods are summarized in Table 7.

### VI. CONCLUSION

Recently, data-driven approaches have been widely adopted to develop methods for accurate SOH estimation to ensure the efficiency, reliability, and safety of LIBs in EVs. Although data-driven methods have been applied in numerous SOH estimation processes, few comprehensive studies have compared the performance of these methods. Therefore, in this study, several different data-driven methods, namely EKF, PF, ARIMA, ELM, SVM, RVM, and LSTM, were investigated and evaluated. To our best knowledge, this is the first work to compare their performance with the real-world EV operation data.

The comparison showed that PF yielded the highest performance in terms of the average accuracy, while ELM was the model with the fastest training and SVM was the model with the fastest computation. Hence, none of the aforementioned methods can be considered an absolutely superior method, and a trade-off among the desired accuracy, the output confidence interval, the ability to deal with nonlinearity, robustness, computation costs, the ability to deal with data sparsity, and generalization should be considered for each particular situation. Table 7 gives a summary of the aforementioned methods.

Finally, this investigation is limited because only the different methods were compared, but the electrochemical model and HI might also affect the estimation performance. In the future, more explorations are needed by comparing the estimation inputs to provide insights into designs.

### REFERENCES

[1] N. Nitta, F. Wu, J. T. Lee, and G. Yushin, “Li-ion battery materials: Present and future,” Mater. Today, vol. 18, no. 5, pp. 252–264, 2015.

[2] R. Thomas, G. Despesse, S. Bacquet, E. Fernandez, Y. Lopez, P. Ramalheira-Andry, and L. Cassarino, “A high frequency self-reconfigurable battery for arbitrary waveform generation,” World Electr. Vehicle J., vol. 12, no. 1, p. 8, Jan. 2021.

[3] B. Gou, Y. Xu, and X. Feng, “State-of-health estimation and remaining-useful-life prediction for lithium-ion battery using a hybrid data-driven method,” IEEE Trans. Veh. Technol., vol. 69, no. 10, pp. 10854–10867, Oct. 2020.

[4] X. Hu, L. Xu, X. Lin, and M. Pecht, “Battery lifetime prognostics,” Joule, vol. 4, no. 2, pp. 310–346, Feb. 2020.

[5] C. R. Birkl, M. R. Roberts, E. McTurk, P. G. Bruce, and D. A. Howey, “Degradation diagnostics for lithium ion cells,” J. Power Sources, vol. 341, pp. 373–386, Feb. 2017.
