Lithium-ion Battery State-of-Health Estimation via Histogram Data, Principal Component Analysis, and Machine Learning

Junran Chen*, Phillip Kollmeyer*, Fei Chiang†, Ali Emadi*
*McMaster Automotive Resource Center (MARC), McMaster University, Hamilton, ON, Canada
†Department of Computing and Software, McMaster University, Hamilton, ON, Canada

2023 IEEE Transportation Electrification Conference & Expo (ITEC) | 979-8-3503-9742-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITEC55900.2023.10187012

***Abstract*—Lithium-ion batteries are widely used in electric vehicle powertrain systems. As batteries age, their state of health (SOH), indicated by their usable capacity and power capability, decreases. For reliable battery operation, accurate estimation and prediction of SOH are essential. This paper proposes an algorithm for estimating battery capacity SOH from an open-source fast charging dataset with many different charge profile types. Histogram data is created from the measured time domain data and fed into a feedforward neural network (FNN). To capture the impact of different charge profiles on aging, current and state of charge (SOC) are multiplied together to create an additional synthetic input to the estimator. To reduce the number of inputs to the FNN to only those that contain valuable information, we use principal component analysis to reduce the total number of inputs by 80%. An SOH algorithm is proposed that can estimate capacity throughout the battery's life with a 1.03% root mean square percentage error (RMSPE) and 0.68% mean absolute percentage error (MAPE).**

I. Introduction

In 2019, the transportation sector was Canada's second-largest source of greenhouse gas (GHG) emissions, increasing by 54% since 1990 [1]. Meanwhile, the study in [2] states that an additional 32 gigatons of global annual carbon dioxide equivalent emissions must be reduced by 2030 to prevent global warming from exceeding 1.5°C. Therefore, electric vehicles are revolutionizing the industry and are expected to replace fossil fuel vehicles globally. Among the components of electric vehicle powertrains, the most expensive is the lithium-ion battery [3]. Due to degradation, the battery's usable capacity and power capability fade over time. Battery state-of-health (SOH) information is necessary to determine battery replacement time and remaining useful life [4], making it critical to accurately estimate battery SOH for safe and reliable operation.

Battery SOH is typically defined as capacity degradation and/or resistance increase. Only capacity degradation SOH is investigated in this work, and resistance increase will be investigated in future work. Capacity SOH can be directly measured by fully charging and then fully discharging the battery in an experimental environment [5]. Unfortunately, it is typically infeasible to perform a charge/discharge test to measure an electric vehicle's battery capacity because a full cycle would take many hours during which the vehicle could not be operated. Instead, experimental testing is typically used to generate an aging data set with significant SOH degradation and an SOH estimation or prediction algorithm is created using this data. However, predicting lithium-ion battery SOH during operation is challenging since battery degradation mechanisms are influenced by various factors, including calendar time, power magnitude, and ambient temperature [6].

Prior research has proposed many techniques to estimate battery SOH [5], [7]. Model-based methods are one of the most popular among them. These methods first build an equivalent circuit [8] or electrochemical battery model [9] and include battery SOH as a model parameter. Then, an adaptive filter, such as an extended Kalman filter [10], will adjust the battery model's parameters, including SOH, during operation. However, since battery degradation is affected by multiple factors, a complicated battery model, ideally including electrochemical aging mechanisms, is needed for this method. By contrast, data-driven strategies extract the factors driving aging directly from experimental data. Fig. 1 illustrates how the battery charging voltage profile changes significantly as the battery ages, allowing SOH to be estimated based on such charging data. In one application of this method, a cloud battery SOH estimation platform is created by inputting battery time-series charging profiles to recurrent neural networks (RNNs) with long short-term memory (LSTM) cells [11].

While battery aging can be directly observed or predicted from time series operation data, including current, voltage, temperature, etc., it can be difficult for algorithms to handle such a large amount of raw data. Several works including [12] propose to convert battery time-series data into accumulated time histogram form, significantly reducing the amount of data saved and making real-time device storage and machine learning estimation feasible. For example, time series voltage data

could be reduced to accumulated time at different voltage ranges such as 2.5 V to 2.6 V, 2.6 V to 2.7 V, etc. The histogram data, which may be a total of a few hundred values, each referred to as a feature, can then be further reduced in size only to include data that impacts aging. In [13] and [12] Spearman and Pearson correlation analysis is used to determine each feature's correlation with SOH, and those features with higher scores are selected for use as inputs to the machine learning model. By contrast, the feature selection in [4] is achieved by the K-means clustering algorithm, which aims to group raw features into several clusters. The centroid of each cluster is then used as a new feature. The work in [12] compares the performance of different machine learning models, including support vector regression, Gaussian process regression, random forest regression, and artificial neural networks (ANN), and finds ANNs to perform the best.

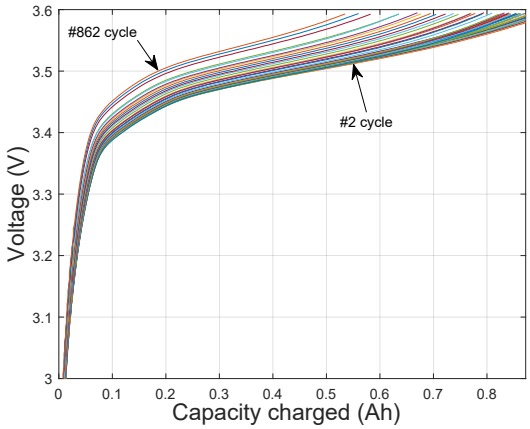


Fig. 1. 4.8 C constant current charge voltage versus charged capacity and cycle number for an A123 LFP cell [14]

In this work, we propose a battery SOH estimation algorithm. First, using a battery charging and discharging time series dataset, we construct an accumulated time histogram for an extensive battery dataset [14]. The time-series current and SOC data are multiplied together and used to create a new histogram feature which captures the relation between SOC, current, and aging. A principal component analysis (PCA) algorithm is applied to the histogram data to reduce the total number of histogram features and refine the battery aging information. A feedforward neural network (FNN) is then used to model the relationship between SOH and the histogram features created using PCA. The accuracy improvements achieved with the new synthetic feature and PCA are discussed. Furthermore, the SOH estimation accuracy of the proposed method is compared to another method in [12], which uses the same data sets used in this study.

II. Experimental Aging Data

The battery data used in this work is from [14] and includes data for 124 cylindrical 1.1 Ah lithium iron phosphate (LFP) cells. These cells are each charged with a unique fast charging profile with current ranging from 2 C-rate to 8 C-rate and are discharged at a fixed 4 C-rate. Therefore, these cells are each aged differently, and the SOH metric used throughout this work is each cycle's 4 C-rate discharge capacity. Fig. 2 shows the capacity (SOH) versus the number of cycles for all 124 cells, of which data for 80 cells (63,749 cycle capacity records) were used as the training set, 24 cells (18,623 cycle capacity records) as the validation set, and 20 cells (16,868 cycle capacity records) as the testing set.

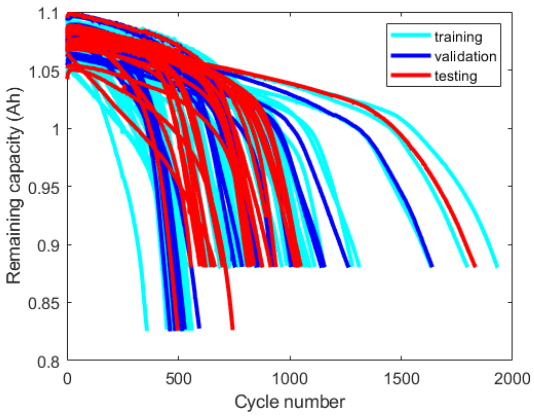


Fig. 2. Measured capacity versus number of cycles for 124 LFP cells each charged at a different rate and distribution of data into training, validation, and testing sets.

III. SOH Estimation Method

Fig. 3 shows the pipeline for transforming raw measured time-series data into to histogram data, for constructing features and reducing the number of features using PCA and for estimating SOH using an FNN.

A. Histogram features construction

Histogram transformation is performed by counting the accumulated time the battery operates in different current, voltage, or temperature ranges. The current range is -10 A to 10 A with 0.1 A interval histogram bins, the voltage range is 1.5 V to 4 V with 0.1 V intervals, and the temperature range is 20° C to 40° C with 1° C intervals. A set of histogram data is created for each cycle, with data accumulated over the battery's life and each set of histogram data having a unique SOH value calculated from the cycle's discharge capacity.

Fig. 4(a) shows though that two different charge current profiles, one with a higher current at low SOC and another with a higher current at high SOC, result in the same current histogram data. To capture this

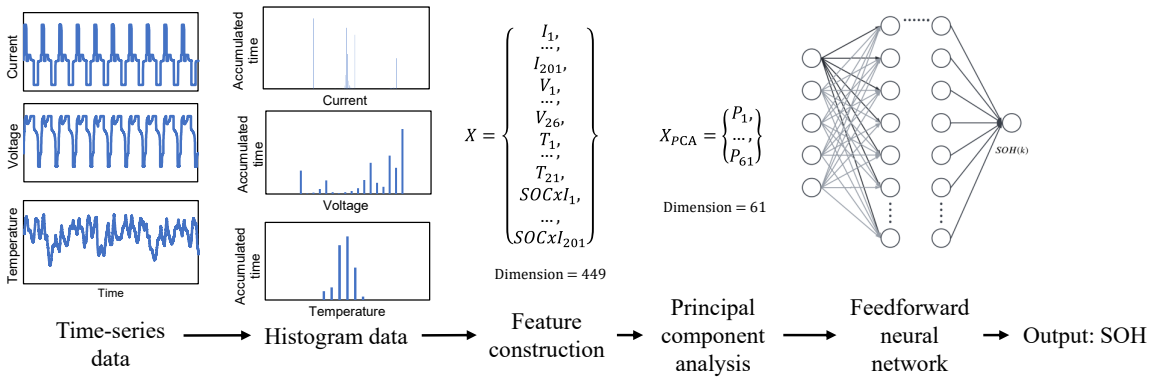
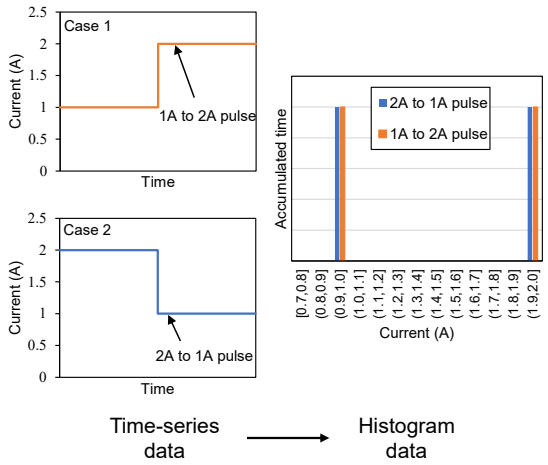
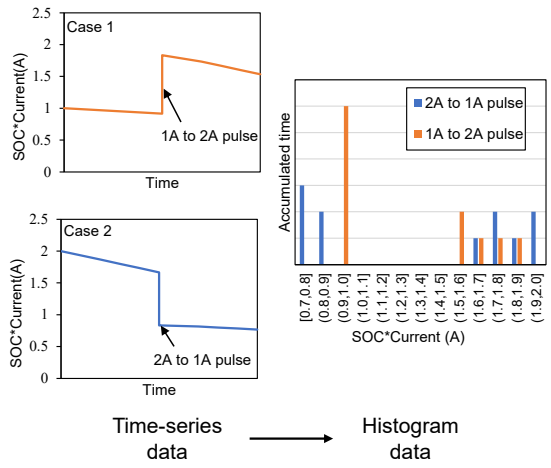


Fig. 3. Workflow of the proposed battery SOH estimation algorithm.



(a) Current data histogram transformation.



(b) SOC times current data histogram transformation.

Fig. 4. Schematic of showing how added feature $SOCxI$ extracts battery current operation sequence information.

relation of SOC and current, a new feature is added in this work, $SOCxI$, which is current multiplied by the corresponding battery state-of-charge (SoC). As shown in Fig. 4(b), the histogram plot of $SOCxI$ distinguishes between battery charge profiles where current is high at high SOC versus current is high at low SOC. The range of $SOCxI$ is -10A to 10A with 0.1 A intervals.

B. Principal component analysis

The time domain data is transformed to a total of 449 current, voltage, temperature, and $SOCxI$ histogram data values. These features X are sparse because the battery will rarely or never operate at some current, voltage, and temperature ranges, resulting in some histogram bins being at or near zero. Some features (i.e. histogram bins) also may not contribute to aging. To ensure good fitting of the FNN to the training set and to reduce training time, a PCA algorithm is implemented to reduce the total number of inputs to the FNN to less than 449. PCA is an unsupervised machine learning method for dimensionality reduction, feature extraction and lossy data compression. It does this by looking at the data and finding a way to represent it with fewer details, while still keeping the most important information [15]. $d(= 449)$ is the dimension of the original feature X and d' is the dimension of the new PCA created features X_{PCA}. First, PCA calculates the covariance matrix $cov(X)$ of original features, and sorts its vectors $(w_1, w_2, ..., w_d)$ by their eigenvalues λ_i. Then, d' is determined by how many features are needed to achieve the threshold reconstruction level t as shown in (1).

$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geqslant t \tag{1}$$

For t = 99%, as used in this work, $d' = 61$. Knowing d', the solution matrix $W^* = (w_1, w_2, ..., w_{d'})$ is part of $cov(X)$. Finally, new features X_{PCA} can be calculated as $X_{\text{PCA}} = W^* X$, which is a linear combination of original features.

C. Feedforward neural network

An FNN, as shown in Fig. 5, is used to model the relationship between battery SOH and features X_{PCA}. The FNN includes two hidden layers with 512 neurons each and two activation functions, including Tanh (hyperbolic tangent) and leaky ReLU (rectified linear unit). TABLE I shows the tuned hyperparameters for training the FNN. The FNN model parameters are determined by repeating the training process with random initial parameters ten times and selecting the trained network with the minimum average validation error.

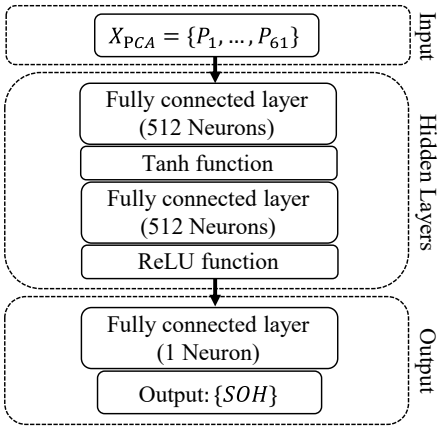


Fig. 5. Structure of feedforward neural network used for SOH estimation.

TABLE I
HYPERPARAMETERS USED FOR FNN TRAINING PROCESS

Optimizer	Adam
Learning rate drop period	1000
Learning rate drop factor	0.1
Validation patience	200
Initial learning rate	0.01
Mini-batch size	full-batch
L2 Regularization factor	1

IV. SOH ESTIMATOR PERFORMANCE WITH AND WITHOUT PCA AND SOCxI FEATURE

To highlight the benefits of PCA and the SOCxI feature, the SOH estimator is trained with neither (case #1), one or the other (case #2 and #3), or both (case #4) of these aspects as shown in TABLE II. The model is trained separately for each configuration and the best trained model is selected using validation error per the method described in the prior section. The error for each case for the twenty cell testing dataset is summarized in TABLE II and plotted in Fig. 6 to Fig. 9.

A. Accelerated training time due to PCA

PCA reduces the number of dimensions of input data to the FNN, making it easier for the FNN to learn the relation between the inputs and battery SOH. Cases #3 and #4 in Table II, both of which include PCA, have fewer total learnable parameters and require about one quarter to one third the training time of configurations #1 and #2 which have PCA. PCA also has potential to reduce model over fitting by reducing the input dimensions. However, it is important to note that the PCA data compression method does result in some loss of information. As a result, for the no PCA versus PCA only cases (case #1 versus case #3), the error is found to increase, going from 1.35% root mean square percentage error (RMSPE) to 2.59% RMSPE. Therefore, while PCA does reduce model size and training time, these improvements may come at the expense of worse estimation accuracy.

B. Improved estimation accuracy due to $SOCxI$

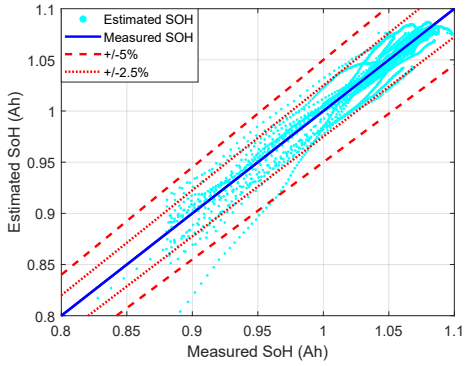
As demonstrated in Fig. 4, the relation between SOC and current during aging is lost due to the accumulated time histogram conversion. The proposed synthetic feature SOCxI captures this relation in the histogram data. Cases #2 and #4 both include the SOCxI feature. Table II shows that this feature reduces RMSPE from 1.35% to 1.19% without PCA respectively from 2.59% to 1.03% with PCA.

C. Improved overall performance due to PCA and $SOCxI$

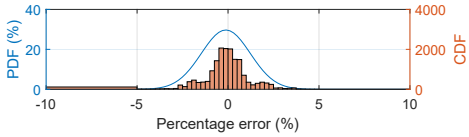
Fig. 10 compares SOH estimated for the four cases versus cycle number for three different cells from the testing dataset. The cells include '4.8C(80%)-4.8C', '8C(20%)-3.6C', and '6C(40%)-3.6C' where '8C(20%)-3.6C' means that the cell is charged to 20% SOC at 8C current and charged to 100% SOC at 3.6C current. The figure shows that all of the models estimate SOH quite accurately even once SOH has started to rapidly decline, and that case #4 with PCA and SOCxI is generally the most accurate.

Overall, case #4 achieves an RMSPE and MAPE of 1.03% and 0.68%, respectively. The probability density function (PDF) and cumulative distribution function (CDF) of the SOH estimation error for case #4 are presented in Fig. 9(b). The maximum error is around -6%, and most of the estimation points are distributed within the error range of -2% and +2%. In contrast, the work in [12], which creates an SOH algorithm with the same data set but with different features and uses correlation analysis to perform feature extraction instead of PCA, has a considerably higher RMSPE and MAPE of 1.92% and 1.13%, respectively, demonstrating the utility and improved performance of the approach proposed in

this paper. Moreover, when adding PCA and SOCxI the number of learnable model parameters is reduced from 0.39 million to 0.3 million, reducing training time as well as the computational burden of the algorithm, which is important for electric vehicle and other mobile applications.

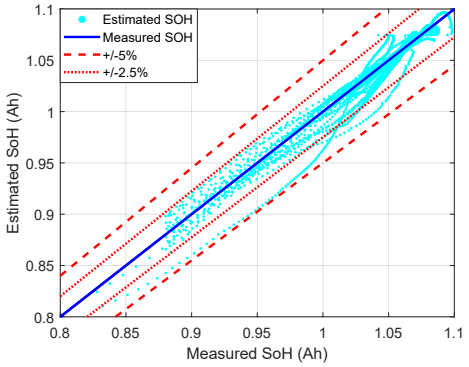


(a) Estimated SOH vs. measured SOH for 20 cell testing set.

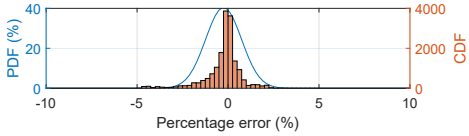


(b) Probability distribution function (PDF) and cumalative distribution function (CDF) of SOH estimation percentage error for the testing set.

Fig. 6. Configuration #1: Case without SOCxI and PCA.

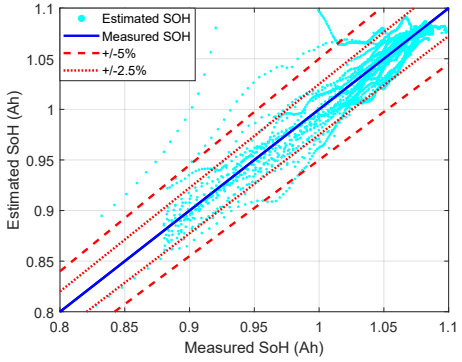


(a) Estimated SOH vs. measured SOH for 20 cell testing set.

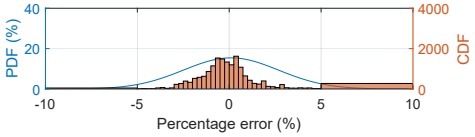


(b) Probability distribution function (PDF) and cumalative distribution function (CDF) of SOH estimation percentage error for the testing set.

Fig. 7. Configuration #2: Case with SOCxI and without PCA.

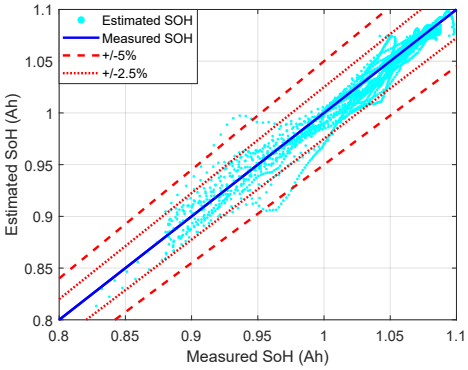


(a) Estimated SOH vs. measured SOH for 20 cell testing set.

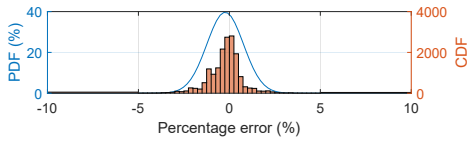


(b) Probability distribution function (PDF) and cumalative distribution function (CDF) of SOH estimation percentage error for the testing set.

Fig. 8. Configuration #3: Case without SOCxI and with PCA.



(a) Estimated SOH vs. measured SOH for 20 cell testing set.



(b) Probability distribution function (PDF) and cumalative distribution function (CDF) of SOH estimation percentage error for the testing set.

Fig. 9. Configuration #4: Case with SOCxI and PCA.

TABLE II
COMPARISON OF FNN HISTOGRAM SOH ESTIMATOR PERFORMANCE WITH AND WITHOUT PCA AND SOCxI FEATURE

	PCA	SOCxI	RMSPE	MAPE	MaxE	Input dimension	Learnable parameters	Training time
#1			1.35%	0.88%	11.1%	248	0.39 million	1653 seconds
#2		✓	1.19%	0.69%	6.2%	449	0.49 million	2286 seconds
#3	✓		2.59%	1.36%	22.2%	59	0.29 million	586 seconds
#4	✓	✓	1.03%	0.68%	6.8%	61	0.30 million	663 seconds

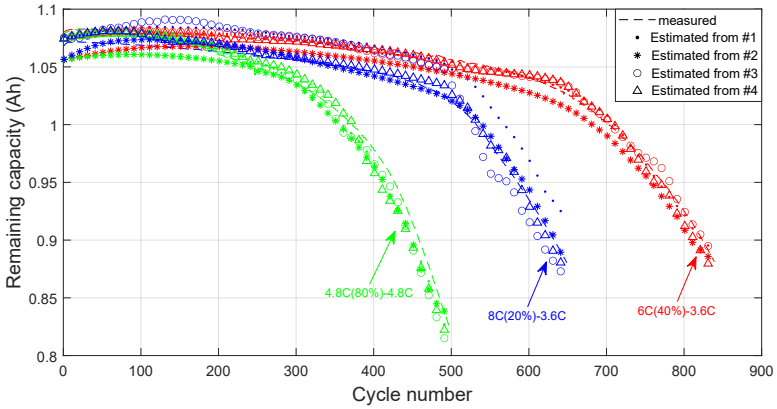


Fig. 10. Estimated vs. measured SOH for '4.8C(80%) to 4.8C', '8C(20%) to 3.6C', and '6C(40%) to 3.6C' cell .

V. CONCLUSION

This work proposes a battery SOH estimation algorithm whose inputs are accumulated time histogram data created from battery time-series operation data. The proposed model achieves very accurate SOH estimation with 1.03% RMSPE and 0.68% MAPE over the twenty battery testing dataset. Overall, the following novel contributions are made: 1) A synthetic SOCxI feature is added to capture the impact of the relation between current magnitude and SOC on aging. The RMSPE and MAPE are reduced by around 25% after adding this feature and PCA. 2) PCA is used to reduce the dimension of the FNN SOH estimator inputs from 449 to 66, reducing model training time by about 1/3 while also reducing model error by 13%.

REFERENCES

[1] "Environment and climate change canada (2021) canadian environmental sustainability indicators: Greenhouse gas emissions." https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/greenhouse-gas-emissions.html, 2018.

[2] "Emissions gap report 2020," *UN environment programme*, 2020.

[3] A. Babin, N. Rizoug, T. Mesbahi, D. Boscher, Z. Hamdoun, and C. Larouci, "Total cost of ownership improvement of commercial electric vehicles using battery sizing and intelligent charge method," *IEEE Transactions on Industry Applications*, vol. 54, no. 2, pp. 1691–1700, 2017.

[4] G.-w. You, S. Park, and D. Oh, "Real-time state-of-health estimation for electric vehicle batteries: A data-driven approach," *Applied energy*, vol. 176, pp. 92–103, 2016.

[5] R. Xiong, L. Li, and J. Tian, "Towards a smarter battery management system: A critical review on battery state of health monitoring methods," *Journal of Power Sources*, vol. 405, pp. 18–29, 2018.

[6] C. R. Birkl, M. R. Roberts, E. McTurk, P. G. Bruce, and D. A. Howey, "Degradation diagnostics for lithium ion cells," *Journal of Power Sources*, vol. 341, pp. 373–386, 2017.

[7] C. Vidal, P. Malysz, P. Kollmeyer, and A. Emadi, "Machine learning applied to electrified vehicle battery state of charge and state of health estimation: State-of-the-art," *IEEE Access*, vol. 8, pp. 52796–52814, 2020.

[8] G. L. Plett, "Extended kalman filtering for battery management systems of lipb-based hev battery packs: Part 2. modeling and identification," *Journal of Power Sources*, vol. 134, no. 2, pp. 262–276, 2004.

[9] R. Ahmed, M. El Sayed, I. Arasaratnam, T. Jimi, and S. Habibi, "Reduced-order electrochemical model parameters identification and soc estimation for healthy and aged li-ion batteries part i: Parameterization model development for healthy batteries," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 2, no. 3, pp. 659–677, 2014.

[10] G. L. Plett, "Extended kalman filtering for battery management systems of lipb-based hev battery packs: Part 1. background," *Journal of Power Sources*, vol. 134, no. 2, pp. 252–261, 2004.

[11] W. Li, N. Sengupta, P. Dechent, D. Howey, A. Annaswamy, and D. U. Sauer, "Online capacity estimation of lithium-ion batteries with deep long short-term memory networks," *Journal of Power Sources*, vol. 482, p. 228863, 2021.

[12] Y. Zhang, T. Wik, J. Bergström, M. Pecht, and C. Zou, "A machine learning-based framework for online prediction of battery ageing trajectory and lifetime using histogram data," *Journal of Power Sources*, vol. 526, p. 231110, 2022.

[13] S. Greenbank and D. Howey, "Automated feature extraction and selection for data-driven models of rapid battery capacity fade and end of life," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 2965–2973, 2022.

[14] K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh, and R. D. Braatz, "Data-driven prediction of battery cycle life before capacity degradation," *Nature Energy*, vol. 4, no. 5, pp. 383–391, 2019.

[15] Z.-H. Zhou, *Machine learning*. Springer Nature, 2021.