Article

Research on Minimization of Data Set for State of Charge Prediction

Tun Liu, Jundong Zhao, Chaoqun Xiang and Shu Cheng *

School of Traffic & Transportation Engineering, Central South University, Changsha 410075, China; liutuncs@csu.edu.cn (T.L.); zhaojd@csu.edu.cn (J.Z.); xiangchq@csu.edu.cn (C.X.)
* Correspondence: chengshu@csu.edu.cn

Abstract: The quick estimation and prediction of lithium-ion batteries’ (LIBs) state of charge (SoC) are attracting growing attention, since the LIB has become one of the most essential power sources for daily consumer electronics. Most deep learning methods require plenty of data and more than two LIB parameters to train the model for predicting SoC. In this paper, a single-parameter SoC prediction based on deep learning is realized by cleaning the data for lithium-ion battery parameters and constructing the feature matrix based on the cleaned data. Then, by analyzing the feature matrix’s periodicity and principal component to obtain two kinds of the original eigenmatrix’s substitution matrices, the two substitutions are fused to obtain an excellent prediction effect. In the end, the minimization method is verified with newly measured lithium battery data, and the results show that the MAPE of the SoC prediction reaches 0.96%, the input data are reduced by 93.33%, and the training time is reduced by 96.68%. Fast and accurate prediction of the SoC is achieved by using only a minimum amount of voltage data.

Keywords: long short-term memory (LSTM); state of charge; principal components analysis (PCA); data fusion

1. Introduction

The lithium-ion battery (LIB) has become one of the essential mobile power sources, from electric vehicles (EVs) to cell phones and from microgrids to laptops, owing to its mature technology, low cost, high energy density, and long service time [1]. It is more and more critical to quickly and accurately obtain an LIB’s state of charge (SoC). However, most of the SoC estimation methods proposed in the literature are based on studies of large equipment that are conducted during the service process, such as with vehicles. This paper investigates LIB’s initial SoC in daily consumer electronics to provide the convenience of future edge computing and the Internet of Things. The initial SoC is helpful to estimate the SoC in use accurately, and is an essential indicator of the state of health (SoH). This paper aims to quickly and accurately predict LIB’s SoC with minimum data.

The existing SoC estimation and prediction methods can be divided into two categories: direct methods and indirect methods. The direct methods include electrochemical impedance spectroscopy (EIS), current integral [2], and open-circuit voltage (OCV) methods, which are dependent on the battery’s measurable variables [3]. The indirect methods usually exploit data-driven methods and model-based methods. The data-driven approaches can be classified into the observer method [4], variants of the Kalman filter (KF) [5–8], the particular filter (PF) [9], and neural network-based deep learning methods [10]. Some data-driven methods also depend on the battery models, which are classified into electrochemical and circuit models.

EIS needs a specific measuring instrument to obtain ohmic resistance, charge resistance, and noise sensitivity, which cannot be adopted for online estimation [11–13].
Current integral/coulomb-counting estimation needs to know the SoC, and relies significantly on precise current measure sensors. Additionally, the EIS is easily affected by temperature variations, the C-rate, and accumulated errors [14,15]. OCV must have a predefined table describing the monotonic relationship between OCV and SoC [16], and needs rest time to obtain the precise value.

Various model-based SoC estimation approaches have been proposed. The electrochemical mechanism model [17] can describe the inherent battery mechanisms, but it is complicated and time-consuming. The equivalent circuit model (ECM) usually uses first-order or second-order models that utilize parallel RC elements (open-circuit voltage terms and resistances) to model the battery’s dynamic behavior [18], and is combined with a data-driven classification algorithm. The first-order models are widely adopted by unscented Kalman filter (UKF) [6], extended Kalman filter (EKF) [7], double particle filter (DPF) [8], sliding-mode observer (SMO) [19], discrete-time nonlinear observer (DNLO) [4], and dual H infinity filter [20] models. There are some second-order models applied in EKF [21] and support vector machine (SVM) [22] models. Other methods, such as mathematical models, use the improved Volterra equations [23] to dynamically describe the battery’s degradation. Though these methods achieve satisfactory performance, building a precise model is still challenging, and plenty of experimental data are necessary.

Currently, neural network-based deep learning methods are the state-of-the-art methods for various applications. Deep learning is model-free and usually only dependent on the battery’s measurable variables. The typical deep learning algorithms include long short-term memory (LSTM) [24,25], convolutional neural networks (CNN) [26–29], and complex CNN-LSTM [30,31]. The deep learning methods can save the tedious work of building the model, but they need a large amount of data from more than two parameters.

This paper proposes a novel approach combined with LSTM to realize promising SoC prediction with only voltage data. The principal component analysis and periodic analysis obtained the feature matrix’s substitution components, and then we fused the two parts. Combining these two kinds of data can enlarge the difference between the original matrices and achieve a better prediction effect. Simultaneously, this method reduces the amount of input data by more than 93% and reduces the training time by 96%.

2. Construction of Data Set

2.1. Data Background

The NASA lithium-ion battery data sets from the Prognostics Center of Excellence Data Repository [32] are chosen for the experiment that only uses voltage data for SoC prediction. The data were collected under the constant current–constant voltage (CC-CV) principle at 43 °C. Specifically, the charging was conducted at a constant current of 1.5 A until the battery reached the upper charging limit of 4.2 V, and then it continued in constant voltage (CV) mode until it reached the current cut-off at 20 mA. The discharging was conducted under a constant current of 4 A until the cell voltage dropped to 2.0 V, 2.2 V, 2.5 V, and 2.7 V (Table 1). The LIB voltage curves are shown in Figure 1a, and the capacity curves are shown in Figure 1b. This paper uses the #30, #31, and #32 batteries’ charging voltage data as training data and the #29 battery’s data as test data.

| Constant Current (A) | Upper Voltage Limit (V) | Cut-Off Current (mA) | Total Data Amount |
|----------------------|-------------------------|----------------------|------------------|
| #29                  | 1.5                     | 4.2                  | 104,136           |
| #30                  | 1.5                     | 4.2                  | 104,136           |
| #31                  | 1.5                     | 4.2                  | 104,136           |
| #32                  | 1.5                     | 4.2                  | 104,136           |
Figure 1. (a) CC-CV charging curve; (b) discharging capacity.

2.2. Characteristic Parameter Cleaning

Selecting appropriate test parameters can reduce the required parameter types and the amount of test data. The measurable data of the battery are mainly pressure, temperature, internal resistance, voltage, and current. The characteristics of these parameters will be analyzed next.

1. Pressure

A.J. Louli [33] found that the irreversible capacity loss of batteries was correlated with the growth of SEI film by studying the negative electrode of different batteries, including the NCA of a ternary lithium battery, the LiCoO2 of a cobaltate lithium battery, and the NMC of a nickel manganese cobalt battery, and discovered that the thickening of the SEI film could be measured by the pressure. During charging, the overall pressure of the battery increases because of the lithium embedded in the negative electrode, which increases the volume of the battery. Conversely, the pressure drops during discharge. The tension generated on the SEI film is weak, and high-precision pressure sensors are often required. In A.J. Louli’s study, the sensor used for measurement was the ultra-small industrial compression weighing sensor Omega-LCKD, and the sensor data processor is OMEGA dp25B-S. However, the sensor and data processor are very high precision and expensive, suitable only for the laboratory.

2. Temperature

Battery heating is mainly concentrated in the battery interior, and even the internal heating is uneven due to the influence of the manufacturing process. It can be seen from the curve of NASA’s single temperature measurement in Figure 2 that the temperature varies over a wide range and takes a long time to stabilize, which also indicates that it is difficult to accurately obtain the internal temperature of the battery through external measurement. It is not an appropriate single parameter to characterize battery degradation. Simultaneously, the external environment easily affects the temperature and pressure parameters.
Figure 2. Battery temperature change curve.

3. Internal resistance

The internal resistance of lithium-ion batteries is usually measured using electrochemical impedance spectroscopy (EIS). An equivalent circuit usually represents the internal resistance of each part of a battery. The resistance values of each part of the equivalent circuit can be obtained by combining the equivalent model with the analysis of the Nyquist diagram. However, the process is complicated by placing the battery in a particular instrument to test it. The battery’s impedance obtained from the NASA data can be seen in Figure 3a. It is not easy to intuitively obtain the impedance, which requires particular analysis software or an AC impedance spectrum tester to analyze. Meanwhile, it can be seen from Figure 3b that the total resistance change of the whole cycle test does not reflect capacity degradation.

Figure 3. Internal resistance: (a) impedance Nyquist diagram; (b) resistance.

4. Voltage and current

Generally, battery management systems provide voltage and current detection functions that do not need additional sensors. However, the current parameter is affected by environmental noise, the load, and the control of the charging instrument. Voltage is the external reflection of the migration of elements inside the battery, and it is not affected by
the external environment, making it an ideal parameter for characterizing the degradation of energy storage devices [34,35]. When charged at a certain current, the voltage can reflect the physical and chemical characteristics of the battery better than the current.

2.3. Data Set Construction

The LSTM network can store long-term and short-term states in two cells, \( C_t \) and \( h_t \), respectively (Figure 4). \( C_t \) denotes the memory cell state, which can store long-term data features. \( h_t \) stores short-term states. The fundamental of LSTM is that the state of the memory cell \( C_t \) can be well propagated and store data features from a long time ago [24]. Therefore, The LSTM deep learning algorithm has the advantage of processing sequence data, and the voltage data are precisely this kind of data. The definitions of the forget gate \( f_t \), input gate \( i_t \), candidate state \( \tilde{C}_t \), new cell state \( C_t \), output gate \( o_t \), and hidden state \( h_t \) are calculated as follows (1)–(6):

\[
f_t = \sigma(w_f \cdot [x_t, h_{t-1}] + b_f) \\
i_t = \sigma(w_i \cdot [x_t, h_{t-1}] + b_i) \\
\tilde{C}_t = \tanh(w_c [x_t, h_{t-1}] + b_c) \\
C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\
o_t = \sigma(w_o \cdot [x_t, h_{t-1}] + b_o) \\
h_t = o_t \odot \tanh (C_t)
\]

Figure 4. LSTM cell structure.

The LSTM algorithm uses a column vector as an input feature (Figure 5). The voltage data of each cycle are converted into a 40 × 40 matrix so that the deep learning models can process it. According to the characteristics of the LSTM, every 40-interval data point of a column is taken as two adjacent factors to make sure that the matrix has the most significant standard deviation in the column.
Then, the voltage data and voltage derivative are combined column by column after being normalized by the z-score method. The normalizations of xmin-xmax, atan, and log limit the output value in the ranges of 0 to 1, −1 to 1, and 0 to 1, respectively, which cannot increase the difference between the data. Therefore, the data are normalized with the z-score method, which places no restrictions on the mapping range of the data, as shown in Equations (7)–(9):

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$  \hspace{1cm} (7)  \\
$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2}$$  \hspace{1cm} (8)  \\
$$\tilde{x}_i = \frac{x_i - \mu}{s} \hspace{1cm} i = 1, 2, \ldots, n$$  \hspace{1cm} (9)  

where $\mu$ is the mean value, $s$ is the standard deviation, and $\tilde{x}_i$ is the normalized value.

The voltage derivative is imported as input data to enlarge the feature of the data set, then arranged with voltage data according to the characteristics of the LSTM algorithm, the arrangement method of which is illustrated in Figure 6. Figure 7 shows the visualization of the arrayed matrix.
3. Analysis of Critical Features of Data Set

The constructed matrix uses up to 3200 pieces of data, and a total of 512,000 pieces of data are used throughout the test cycle of the four batteries. If the battery has more cycles, the amount of data used will be larger, which will increase the calculation burden and training time, so the data need to be minimized.

3.1. Data Periodicity Analysis

One way to reduce the dimensionality of the data is to replace the original data with its minimum period data if the data are periodic. From the visualization of NASA’s battery data matrix in the previous section, the amplitude varies roughly periodically. In order to verify its periodic change, the correlation index between the even and odd columns of the matrix is analyzed by using the correlation coefficient function in Equation (4). For even columns, take the first two columns 1–2, 1–4, 1–6, and 1–8 with the subsequent even sequence, such as (i + 1)–(i + 2) column i = 2, 4, 6, ..., k/2; (i + 1)–(i + 4) column i = 4, 6, ..., k/4; (i + 1)–(i + 6) column i = 6, ..., k/6; (i + 1)–(i + 8) column i = 8, ..., k/8, and carry out the correlation coefficient analysis. The correlation coefficient formula is as follows:

\[ Corr = \frac{\sum_{m} \sum_{n} (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\sum_{m} \sum_{n} (A_{mn} - \bar{A})^2} \sqrt{\sum_{m} \sum_{n} (B_{mn} - \bar{B})^2}} \] (10)

Figure 8a shows that the correlation coefficient between even columns presents an approximate linear downward trend, but the correlation coefficients remain above 0.98. The high correlation coefficients indicate a high similarity and linear relationship between even columns. The correlation coefficients of odd columns in Figure 8b show significant fluctuation and poor similarity, so odd columns cannot be used to replace the remaining data.
Figure 8. The correlation coefficients of the matrix of (a) even columns and (b) odd columns.

We changed the discrete sampling signal into the time domain signal with the sampling period of 1 s steps, then analyzed its periodicity through the Fourier transform equation in Equation (11):

$$ x_k = \sum_{n=0}^{N-1} x_n e^{-j \frac{2\pi k}{N} n} \quad (k = 0, 1, 2 \ldots N - 1) \quad (11) $$

The Euler expansion of the above formula can be obtained in (12):

$$ x_k = \sum_{n=0}^{N-1} x_n \left( \cos \left(2\pi k \frac{n}{N} \right) - j \sin \left(2\pi k \frac{n}{N} \right) \right) \quad (12) $$

After Fourier series expansion, the spectrum is converted to the ordinate and listed as columns. It can be seen that the amplitude of the discrete time domain signal also changes periodically every four columns, as shown in Figure 9.

Figure 9. Fourier analysis diagram.
3.2. Data Difference Analysis

The global standard deviation (Figure 10a) shows that although there are periodic changes, the standard deviation difference between an odd column and even column is too large. The difference in each column is too small to distinguish the data that can provide more different features.

![Figure 10](image1.png)

**Figure 10.** The global diagram for (a) standard deviation and (b) ACV.

The absolute coefficient of variation (ACV) (Equation (13)) is introduced in this section to further analyze the degree of data dispersion. The absolute value of the ratio of the standard deviation to the mean is the absolute coefficient of variation, which can be used to measure the relative discrete relationships between data. The smaller the coefficient of variation is, the smaller the dispersion degree of the data is; the larger the coefficient of variation is, the greater the dispersion degree of the data is.

\[
ACV = \left| \frac{s}{\mu} \right| 
\]  

(13)

![Figure 11](image2.png)

**Figure 11.** Box plot of a single cycle.
The global diagram of ACV is generated with the all circulating data of NASA in line with the column direction, as shown in Figure 11. In the graph, the abscissa is the number of columns of the matrix, and the ordinate is the number of cycles. To better analyze the ACV distribution, the maximum ACV is limited to 10. Figure 10b shows that the column numbers with a large coefficient of variation are evenly distributed.

The global analysis results cannot provide data differences within a single cycle. Therefore, the box plot in statistics is introduced to analyze data variability within a single cycle. In the box plot of the NASA data (Figure 11), it can be seen that in the even columns, the median of even columns is basically the same, and the quartile distance is not significantly different, while the median of the odd number columns fluctuates slightly. The variation decreased with the increase in the column number. Thus, we consider taking the even columns in front of the matrix as a substitute for the whole matrix.

3.3. Principal Component Analysis

Principal component analysis (PCA) is another widely used data feature substitution method (1–3) adopted to find the matrix variation’s principal components. Finding the principal components to replace the original matrix through PCA can reduce the size of the original matrix and the amount of data.

The steps of PCA are as follows:

1. Normalization of the matrix

Establish the input matrix $X = [x_1, x_2, \cdots, x_m]^T$ (14), where $x_i = (\bar{x}_{i1}, \bar{x}_{i2}, \cdots, \bar{x}_{in})$. $\bar{x}_{ij}$ is computed by Equation (15):

$$
\bar{x}_{ij} = \frac{x_{ij} - \mu}{s} \quad (i=1,2,\cdots, m; j=1,2,\cdots, n) \tag{15}
$$

where $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $s = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2$

2. Correlation coefficient calculation

The normalized data are used to calculate the correlation coefficient matrix $C = (c_{ij})_{m \times n}$ (16), where $c_{ii} = 1$ and $c_{ij} = c_{ji}$. $c_{ij}$ denotes the correlation coefficient between the $i$th column and the $j$th column (17).

$$
C = \begin{bmatrix}
    r_{11} & \cdots & r_{1n} \\
    \vdots & \ddots & \vdots \\
    r_{m1} & \cdots & r_{mn}
\end{bmatrix} \tag{16}
$$

$$
c_{ij} = \frac{\sum_{k=1}^{m} \bar{a}_{ki} \bar{a}_{kj}}{m-1} \quad (i, j=1,2,\cdots, m; k<i, j) \tag{17}
$$

where $\bar{a}_{ki} = (x_{ki} - \bar{x}_{ki})$ and $\bar{a}_{kj} = (x_{kj} - \bar{x}_{kj})$.

3. Eigenvalue and principal matrix

$$
|\lambda E - C| = 0 \tag{18}
$$

Through Equation (18), the eigenvalue of $R$ can be obtained, where $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_m)$ and eigenvector $u = (u_{11}, u_{21}, \cdots, u_{mn})^T$. $y_i$ denotes the $i$th principal component of the matrix. The principal matrix $Y = [y_1, y_2, \cdots, y_m]$ is shown in (19):

$$
[Y_1, Y_2, \cdots, Y_m] = \begin{bmatrix}
    x_{11} & \cdots & x_{1n} \\
    \vdots & \ddots & \vdots \\
    x_{m1} & \cdots & x_{mn}
\end{bmatrix} \begin{bmatrix}
    u_{11} & \cdots & u_{1n} \\
    \vdots & \ddots & \vdots \\
    u_{m1} & \cdots & u_{mn}
\end{bmatrix} \tag{19}
$$

4. Contribution ratio
The contribution \( b_j \) of \( y_j \) is calculated through Equation (20), and \( a_p \) in Equation (21) is the accumulative contribution rate:

\[
b_j = \frac{\lambda_j}{\sum_{k=1}^{m} \lambda_k} \tag{20}
\]

\[
a_p = \frac{\sum_{k=1}^{p} \frac{1}{\lambda_k}}{\sum_{k=1}^{m} \lambda_k} \tag{21}
\]

when \( a_p \) is close to 1 (>85%) and the \( p \) PC matrix is usually chosen to replace the original matrix. Figure 12 shows that the contribution rate of the first several columns in the principal component matrix is the largest, and the contribution rate of the subsequent columns decreases sharply. The accumulative contributions of PC8 and PC20, respectively, are over 88% and 98%.

![Figure 12. The contribution rate of PC matrices.](image)

### 4. Prediction and Minimization

#### 4.1. Prediction Based on Periodic and Principal Components Data

In this section, the even-numbered columns of the original matrix and the principal component matrix are respectively tested by the LSTM. The testing platform is equipped with an Intel Core (TM) i7-9750H CPU, 16 GB RAM, and NVIDIA GeForce GTX 1650 Max-Q GPU. The LSTM used has five layers: the first layer is the sequence layer, which is set to 30 according to the matrix characteristics parameter, and the second layer is the LSTM functional layer; the third layer is the full connection layer; the fourth layer is the normalized function layer; the fifth layer is the classification layer. Training success is defined by a training accuracy of over 98%. The SoC is defined in Equation (10), where \( Q_{\text{rated}} \) is the battery’s rated energy and \( Q_{\text{remain}} \) is the battery’s remaining energy. When the \( Q_{\text{remain}} \) is the next maximum releasable energy after the battery is fully charged, the initial SoC can be obtained (22). The mean squared error (MSE) (23) and the mean absolute percent error (MAPE) (24) are used as error evaluation indicators.

\[
\text{SoC} = \frac{Q_{\text{remain}}}{Q_{\text{rated}}} \tag{22}
\]

\[
\text{MSE} = \frac{1}{k} \sum_{k=1}^{k} (\hat{C}_k - C_k)^2 \tag{23}
\]

\[
\text{MAPE} = \frac{1}{k} \sum_{k=1}^{k} \left| \frac{\hat{C}_k - C_k}{C_k} \right| \times 100\% \tag{24}
\]

The prediction results are shown in Figure 13 and Table 2. Table 2 illustrates that when the complete normalized matrices are used, the MSE and MAPE are \( 2.56 \times 10^{-4} \) and 1.42%, respectively. Though a good prediction was obtained in the test, the model was trained by using a large amount of input data in one minute and thirty-eight seconds. A large amount of input data increases the computational load and the data sampling and
storage costs. Additionally, it delays the results output. This is challenging for online real-time prediction. When using different even columns for prediction, the training time and input data are significantly reduced, but the prediction effect is slightly worse than the original matrices.

![Graph](image)

**Figure 13.** The prediction result of the (a) original matrix with different columns and (b) PC matrices.

**Table 2.** Error Evaluation Results.

| Errors Data | MSE       | MAPE   | Used Data Rate | Training Time |
|-------------|-----------|--------|----------------|---------------|
| 40 × 80     | 2.41 × 10^{-4} | 1.28%  | 100.00%        | 1 min 38 s    |
| 40 × 2      | 5.10 × 10^{-4} | 1.87%  | 2.50%          | 6 s           |
| 40 × 4      | 2.46 × 10^{-4} | 1.36%  | 5.00%          | 4 s           |
| 40 × 6      | 2.41 × 10^{-4} | 1.33%  | 7.50%          | 7 s           |
| 40 × 8      | 2.50 × 10^{-4} | 1.34%  | 10.00%         | 24 s          |
| 40 × 20     | 3.20 × 10^{-4} | 1.54%  | 10.00%         | 5 s           |
| Fusion data | 5.02 × 10^{-4} | 2.01%  | 25.00%         | 60 s          |
| 40 × 4      | 2.39 × 10^{-4} | 0.89%  | 5.00%          | 4 s           |

Overall, the prediction results of using 1–4 original columns are slightly worse than those using the complete original matrix. In contrast, using 1–4 original columns can reduce the training time and the data use ratio by 96% and 95%, respectively. Nevertheless, the prediction effect of the principal component matrices is much worse than that of the original matrices.

4.2. Minimization

The prediction results of periodic data substitution and principal component substitution are worse than those of the original matrix, but significantly reduced the training time. Since both data substitutions represent the main features of the original matrix, further fusion is considered for testing the prediction effect. To combine the advantages of both substitution matrices, we created a matrix by combining the principal components with the periodic data, as shown in (25):

\[ X = [x_1, x_2, x_3, x_4, PC_1, PC_2, PC_3, \ldots, PC_8]^T \]  

(25)

In the first four columns of the matrix, \(x_{1-4}\) is the first four columns of the original data, and the next eight columns are the eight principal components of \(PC_{1-8}\). After standardization to calculate the correlation coefficient \(C_{ij}\), the eigenvalue \(\lambda\) and eigenvectors...
U are then obtained by solving the eigenequation using the Jacobian matrix method. A new fusion matrix is obtained by multiplying the first four eigenvectors of the composite matrix X and the corresponding matrix U. The cumulative contribution rate of the new fusion data’s first four principal components is more than 92%. Figure 13b displays the prediction curve.

Table 2 shows that the MSE and MAPE of the fusion data are reduced by 0.46% and 30.47%, respectively, and the training time is also reduced by 96%. We conclude that the fusion data, combining the advantages of the two alternative data, has the best prediction effect.

Verification

We tested four LIBs (LIR2032) under the CC-CV charging process at room temperature to verify the previous analysis results. The charging was conducted under a constant current (CC) of 20 mA until it reached the upper voltage limit of 4.2 V. Then, it was continued in constant voltage (CV) mode until the charge current dropped to 2 mA. The discharging was conducted under a constant current of 20 mA until the voltage dropped to 3 V. In the discharging process, we obtain the available discharging capacities. Table 3 lists the work conditions.

| Constant Current (mA) | Upper Voltage Limit | Cut-off Current (mA) | Total Data Amount |
|-----------------------|---------------------|----------------------|------------------|
| Battery #1            | 20                  | 4.2                  | 1,238,394        |
| Battery #2            | 20                  | 4.2                  | 1,094,865        |
| Battery #3            | 20                  | 4.2                  | 1,230,828        |
| Battery #4            | 20                  | 4.2                  | 1,071,722        |

The rated capacity of the new LIB was 35 mAh, which is lower than that of the NASA data, resulting in fewer sample numbers, but more new LIB test cycles. First, the testing data are normalized and changed to 30 × 60 matrices with the voltage data and voltage derivative. Then, we calculated the principal components of the four batteries’ charging data through PCA. Next, the predictions of SoC based on the matrix that takes the same columns from the original matrix via the LSTM model were compared. The #1–3 LIB charging voltage data are used as training data, and the #4 LIB charging voltage data are used as testing data. The amount of data for the test is 624,600, accounting for 58.28% of the total testing data of 1,071,722.

We found that a few PCs still account for a large proportion of the contribution, and the accumulative contributions are over 88% and 98% when taking four PCs and 14 PCs, respectively. The new batteries’ matrices display periodic changes in every two columns, and the even columns have a correlation coefficient of over 0.96 with the subsequent even columns. The average standard deviation of even columns is much smaller than the change in PC, and fluctuates around 0.5. Figure 14a shows the standard deviation of the original matrix, which shows periodic changes in the even-numbered columns; in Figure 14b, it rapidly decreases as the number of columns increases. Besides, Figure 14c illustrates that the high ACV values appear in the first few columns; Figure 14d indicates that the ACV of the PC matrices are distributed evenly among the columns and cycles. Then, the periodic data and principal component data of the measured data are fused, and the three kinds of data are tested separately.
Figure 14. The (a) standard deviations of 694 cycles and (b) PC matrices; the ACV of (c) original matrices and (d) PC matrices.

Figure 15a and Table 4 show that the MSE and MAPE of using the original full matrices are $2.14 \times 10^{-4}$ and 1.37%, respectively. Using two columns from the original matrix, the MSE and MAPE increased by 105.6% and 45.8%, respectively; using 1–4 columns of the original matrix, the MSE and MAPE respectively increased by 10.3% and 6.7%, and the training time and data used ratio respectively decreased by 92.9% and 95%. When using PC4 matrices, the MSE and MAPE increased by 69.6% and 22.5%, respectively; for PC14 matrices, the MSE and MAPE increased by 143% and 56.7%, respectively. The MSE and MAPE of 1–4 columns are 132.7% and 50% lower than PC14. This verifies that the prediction based on directly using the even columns of the matrices composed of voltage and voltage derivatives is better than that of the PC matrices. However, the training speed of using the principal component is much faster than using the even columns. The results listed in Table 4 demonstrate that the MSE of the first four columns and MAPE respectively increase only by 10.28% and 6.67% compared with the complete matrix and the PC4, with an accumulating contribution rate over 88%. Therefore, the study considers combining the even columns and PC4 to achieve their prediction accuracy and speed advantage.
Figure 15. The prediction results of (a) a variety of PC matrices and the original matrix with different columns and (b) fusion data matrix.

Table 4. Error evaluation results.

| Errors Data Type | MSE       | MAPE   | Used Data Rate | Training Time |
|------------------|-----------|--------|----------------|---------------|
| Original 30 × 60 | 2.14 × 10⁻⁴ | 1.20%  | 100.00%        | 2891 s        |
| First columns   | 4.40 × 10⁻⁴ | 1.75%  | 3.30%          | 295 s         |
|                  | 2.36 × 10⁻⁴ | 1.28%  | 6.70%          | 203 s         |
| PC               | 3.39 × 10⁻⁴ | 1.47%  | 6.70%          | 98 s          |
|                  | 5.35 × 10⁻⁴ | 1.69%  | 22.10%         | 68 s          |
| Fusion data 30 × 4 | 2.09 × 10⁻⁴ | 0.96%  | 6.70%          | 96 s          |

Figure 15b shows that the prediction curve of the principal components combined with the first four columns exhibits promising results. Table 4 illustrates that the MSE and MAPE reach the lowest values of 2.09 × 10⁻⁴ and 0.96% in all predictions, which are 9.16% and 20.00% lower than the complete matrix, respectively. The input data were reduced by 93.33%, from the original 1800 to 120, accomplishing the goal of using the minimum amount of input data to achieve an accurate prediction of an LIB’s SoC. Meanwhile, the training time was reduced by 67.46% and 96.68%, compared with the original first four columns and the original matrix, respectively.

5. Conclusions

The normalized voltage data and voltage derivative data are combined to form a new matrix with a high similarity (the correlation coefficient value is over 0.96) between even-numbered columns, realizing SoC prediction based on a single parameter. We compared the predictions based on the new matrices and the PCA matrices, which consisted of PCs with a cumulative contribution rate of more than 85%, and found that the former predictions are better than those based on the PC matrices. The prediction effect of both alternate data is worse than that of using the original matrix, but the training time is much shorter.

Then, the prediction effect was tested by fusing data. The fusion matrix of PCs and even columns contribute to better prediction results than the complete matrix, with an MSE of 2.09 × 10⁻⁴ and a MAPE of 0.96%. By minimizing the data set, we achieved high-precision SoC prediction with less than 10% of the input data based on a single parameter. Simultaneously, the novel approach reduced the training time by 96.68%.
Author Contributions: Investigation, methodology, writing—review and editing, T.L.; validation, C.X.; visualization, J.Z.; supervision, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52072414.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The NASA’s batteries data can be found at https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/. The rest data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

\begin{align*}
  b_f, b_i, b_c, b_o & \quad \text{The bias of different processing units in LSTM} \\
  w_f, w_i, w_c, w_o & \quad \text{The weights of different processing units in LSTM} \\
  \sigma & \quad \text{Sigma function} \\
  \mu & \quad \text{The mean value of input data} \\
  \bar{x}_i & \quad \text{The normalized value of input data} \\
  v', \Delta v' & \quad \text{The normalized voltage value and voltage derivative value} \\
  b_j & \quad \text{The principal component contribution rate} \\
  a_p & \quad \text{Principal components accumulative contribution rate} \\
  PC_i & \quad \text{Principal components matrix}
\end{align*}

References

1. Wu, J.; Zhang, C.; Chen, Z. An online method for lithium-ion battery remaining useful life estimation using importance sampling and neural networks. *Appl. Energy* **2016**, 173, 134–140. https://doi.org/10.1016/j.apenergy.2016.04.057.

2. Liu, X.; Wu, J.; Zhang, C.; Chen, Z. A method for state of energy estimation of lithium-ion batteries at dynamic currents and temperatures. *J. Power Sources* **2014**, 270, 151–157. https://doi.org/10.1016/j.jpowsour.2014.07.107.

3. Song, X.; Yang, F.; Wang, D.; Tsui, K.-L. Combined CNN-LSTM Network for State-of-Charge Estimation of Lithium-Ion Batteries. *IEEE Access* **2019**, 7, 88894–88902. https://doi.org/10.1109/access.2019.2926517.

4. Li, W.; Liang, L.; Liu, W.; Wu, X. State of Charge Estimation of Lithium-Ion Batteries Using a Discrete-Time Nonlinear Observer. *IEEE Trans. Ind. Electron.*** **2017**, 64, 8557–8565. https://doi.org/10.1109/TIE.2017.2703685.

5. Paschero, M.; Storti, G.L.; Rizzi, A.; Mascioli, F.M.F.; Rizzoni, G. A Novel Mechanical Analog-Based Battery Model for SoC Estimation Using a Multicell EKF. *IEEE Trans. Sustain. Energy* **2016**, 7, 1695–1702. https://doi.org/10.1109/tste.2016.2574755.

6. Wang, W.; Wang, X.; Xiang, C.; Wei, C.; Zhao, Y. Unscented Kalman Filter-Based Battery SOC Estimation and Peak Power Prediction Method for Power Distribution of Hybrid Electric Vehicles. *IEEE Access* **2018**, 6, 35957–35965. https://doi.org/10.1109/access.2018.2850743.

7. Chen, Z.; Fu, Y.; Mi, C. State of Charge Estimation of Lithium-Ion Batteries in Electric Drive Vehicles Using Extended Kalman Filtering. *IEEE Trans. Veh. Technol.* **2013**, 62, 1020–1030. https://doi.org/10.1109/tvt.2012.2235474.

8. Xiong, R.; Zhang, Y.; He, H.; Zhou, X.; Pecht, M.G. A Double-Scale, Particle-Filtering, Energy State Prediction Algorithm for Lithium-Ion Batteries. *IEEE Trans. Ind. Electron.* **2017**, 65, 1526–1538. https://doi.org/10.1109/tie.2017.2733475.

9. Zhou, D.; Zhang, K.; Ravey, A.; Gao, F.; Miraoui, A. Online Estimation of Lithium Polymer Batteries State-of-Charge Using Particle Filter-Based Data Fusion With Multimodels Approach. *IEEE Trans. Ind. Appl.* **2016**, 52, 2582–2595.

10. How, D.N.T.; Hannan, M.A.; Lipu, M.S.H.; Sahari, K.S.M.; Ker, P.J.; Muttaqi, K.M. State-of-Charge Estimation of Li-Ion Battery in Electric Vehicles: A Deep Neural Network Approach. *IEEE Trans. Ind. Appl.* **2020**, 56, 5565–5574.

11. Xiao, D.; Fang, G.; Liu, S.; Yuan, S.; Ahmed, R.; Habibi, S.; Emadi, A. Reduced-Coupling Coestimation of SOC and SOH for Lithium-Ion Batteries Based on Convex Optimization. *IEEE Trans. Power Electron.* **2020**, 35, 12332–12346. https://doi.org/10.1109/tpele.2020.2984248.

12. Wang, Q.-K.; He, Y.-J.; Shen, J.-N.; Hu, X.-S.; Ma, Z.-F. State of Charge-Dependent Polynomial Equivalent Circuit Modeling for Electrochemical Impedance Spectroscopy of Lithium-Ion Batteries. *IEEE Trans. Power Electron.* **2017**, 33, 8449–8460. https://doi.org/10.1109/tpele.2017.2780184.

13. Rodrigues, S.; Munichandraiah, N.; Shukla, A. A review of state-of-charge indication of batteries by means of a.c. impedance measurements. *J. Power Sources* **2000**, 87, 12–20. https://doi.org/10.1016/s0378-7753(99)00351-1.

14. Zhang, Y.; Song, W.; Lin, S.; Feng, Z. A novel model of the initial state of charge estimation for LiFePO4 batteries. *J. Power Sources* **2014**, 248, 1028–1033. https://doi.org/10.1016/j.jpowsour.2013.09.135.
15. Ng, K.S.; Moo, C.-S.; Chen, Y.-P.; Hsieh, Y.-C. Enhanced coulomb counting method for estimating state-of-charge and state-of-health of lithium-ion batteries. *Appl. Energy* **2009**, *86*, 1506–1511. Available online: https://ideas.repec.org/a/eee/apene/v86y20099pp1506-1511.html (accessed on 16 August 2020).
16. Lu, L.; Han, X.; Li, J.; Hua, J.; Ouyang, M. A review on the key issues for lithium-ion battery management in electric vehicles. *J. Power Sources* **2013**, *226*, 272–288. https://doi.org/10.1016/j.jpowsour.2012.10.060.
17. Hu, X.; Sun, F.; Zou, Y. Comparison between two model-based algorithms for Li-ion battery SOC estimation in electric vehicles. *Simul. Model. Pr. Theory* **2013**, *34*, 1–11. https://doi.org/10.1016/j.simpat.2013.01.001.
18. Cambron, D.C.; Cramer, A.M. A Lithium-Ion Battery Current Estimation Technique Using an Unknown Input Observer. *IEEE Trans. Veh. Technol.* **2017**, *66*, 6707–6714. https://doi.org/10.1109/tvt.2017.2657520.
19. Il-Song, K. Nonlinear State of Charge Estimator for Hybrid Electric Vehicle Battery. *IEEE Trans. Power Electron.* **2008**, *23*, 2027–2034.
20. Xu, W.; Xu, J.; Lang, J.; Yan, X. A Multi-Timescale Estimator for Lithium-Ion Battery State of Charge and State of Energy Estimation Using Dual H Infinity Filter. *IEEE Access* **2019**, *7*, 181229–181241. https://doi.org/10.1109/access.2019.2959396.
21. Bhangu, B.S.; Bentley, P.; Stone, D.; Bingham, C. Nonlinear Observers for Predicting State-of-Charge and State-of-Health of Lead-Acid Batteries for Hybrid-Electric Vehicles. *IEEE Trans. Veh. Technol.* **2005**, *54*, 783–794. https://doi.org/10.1109/tvt.2004.842461.
22. Li, R.; Xu, S.; Li, S.; Zhou, Y.; Zhou, K.; Liu, X.; Yao, J. State of Charge Prediction Algorithm of Lithium-Ion Battery Based on PSO-SVR Cross Validation. *IEEE Access* **2020**, *8*, 10234–10242. https://doi.org/10.1109/access.2020.2964852.
23. Sidorov, D.; Tynda, A.; Mufahov, I.; Dreglea, A.; Liu, F. Nonlinear Systems of Volterra Equations with Piecewise Smooth Kernels: Numerical Solution and Application for Power Systems Operation. *Mathematics* **2020**, *8*, 1257. https://doi.org/10.3390/math8081257.
24. Greff, K.; Srivastava, R.K.; Koutnik, J.; Steunebrink, B.R.; Schmidthuber, J. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2222–2232. https://doi.org/10.1109/tnnls.2016.2582924.
25. Park, K.; Choi, Y.; Choi, W.J.; Ryu, H.-Y.; Kim, H. LSTM-Based Battery Remaining Useful Life Prediction With Multi-Channel Charging Profiles. *IEEE Access* **2020**, *8*, 20786–20798. https://doi.org/10.1109/access.2020.2968939.
26. Shomron, G.; Weiser, U. Spatial Correlation and Value Prediction in Convolutional Neural Networks. *IEEE Comput. Arch. Lett.* **2019**, *18*, 10–13. https://doi.org/10.1109/lca.2018.2890236.
27. Sun, Y.; Xue, B.; Zhang, M.; Yen, G.G.; Lv, J. Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification. *IEEE Trans. Cybern.* **2020**, *50*, 3840–3854. https://doi.org/10.1109/tcyb.2020.2983860.
28. Liu, C.; Tan, J.; Shi, H.; Wang, X. Lithium-Ion Cell Screening With Convolutional Neural Networks Based on Two-Step Time-Series Clustering and Hybrid Resampling for Imbalanced Data. *IEEE Access* **2018**, *6*, 59001–59014. https://doi.org/10.1109/access.2018.2875514.
29. Li, W.; Liu, H.; Wang, Y.; Li, Z.; Jia, Y.; Gui, G. Deep Learning-Based Classification Methods for Remote Sensing Images in Urban Built-Up Areas. *IEEE Access* **2019**, *7*, 36274–36284. https://doi.org/10.1109/access.2019.2903127.
30. Zheng, Q.; Fang, J.; Hu, Z.; Zhang, H. Aero-Engine On-Board Model Based on Batch Normalize Deep Neural Network. *IEEE Access* **2019**, *7*, 54855–54862. https://doi.org/10.1109/access.2018.2885199.
31. Alhussein, M.; Aurangzeb, K.; Haider, S.I. Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting. *IEEE Access* **2020**, *8*, 180544–180557. https://doi.org/10.1109/access.2020.3028281.
32. B. Saha and K. Goebel. Battery Data Set, NASA Ames Prognostics Data Repository (http://ti.arc.nasa.gov/project/prognostic-data-repository) 2007, NASA Ames Research Center, Moffett Field, CA.
33. Louli, A.; Ellis, L.; Dahn, J. Operando Pressure Measurements Reveal Solid Electrolyte Interphase Growth to Rank Li-Ion Cell Performance. *Joule* **2019**, *3*, 745–761. https://doi.org/10.1016/j.joule.2018.12.009.
34. Mairé, P.; Kaiser, H.; Schiefele, W.; Novák, P. Colorimetric determination of lithium-ion mobility in graphite composite electrodes. *J. Electroanal. Chem.* **2010**, *644*, 127–131. https://doi.org/10.1016/j.jelechem.2009.09.011.
35. DuBeshter, T.; Jorne, J. Pulse Polarization for Li-Ion Battery under Constant State of Charge: Part I. Pulse Discharge Experiments. *J. Electrochem. Soc.* **2017**, *164*, E3539–E3546. https://doi.org/10.1149/2.055171jes.