Article

Data Cleaning and Restoring Method for Vehicle Battery Big Data Platform

Shuangqi Li[1, 2], Hongwen He[1, *], Pengfei Zhao[3], Shuang Cheng[2]

[1] National Engineering Laboratory for Electric Vehicles, School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China

[2] Department of Electronic and Electrical Engineering, University of Bath, Bath, UK

[3] The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[*] Corresponding author: hwhebit@bit.edu.cn (Hongwen He)

Abstract: Battery is one of the most important and costly devices in electric vehicles (EVs). Developing an efficient battery management method is of great significance to enhancing vehicle safety and economy. Recently developed big-data and cloud platform computing technologies bring a bright perspective for efficient utilization and protection of vehicle batteries. However, a reliable data transmission network and a high-quality cloud battery dataset are indispensable to enable this benefit.

This paper makes the first effort to systematically solve data quality problems in cloud-based vehicle battery monitoring and management by developing a novel integrated battery data cleaning framework. In the first stage, the outlier samples are detected by analyzing the temporal features in the battery data time series. The outlier data in the dataset can be accurately detected to avoid their impacts on battery monitoring and management. Then, the abnormal samples, including the noise polluted data and missing value, are restored by a novel future fusion data restoring model. The real electric bus operation data collected by a cloud-based battery monitoring and management platform are used to verify the performance of the developed data cleaning method. More than 93.3% of outlier samples can be detected, and the data restoring error can be limited to 2.11%, which validates the effectiveness of the developed methods. The proposed data cleaning method provides an effective data quality assessment tool in cloud-based vehicle battery management, which can further boost the practical application of the vehicle big data platform and Internet of vehicle.

Keywords: Big data, Internet of vehicle, electric vehicles, data cleaning, battery management system, battery state estimation.

1. Introduction

Along with the increasingly severe energy depletion and environmental pollution problems, the demand for transportation electrification has grown rapidly within the past few decades [1, 2]. According to a strategy carried out by the United Kingdom government in 2018, the sales of traditional diesel and hybrid vehicles will be fully banned after 2040 [3]. China also attaches great importance to the promotion of electric vehicles (EVs), and by 2020, more than 60% of the public electric buses have been successfully electrified [4]. However, the concerns about the security and cost of the battery pack still make the consumers worried [5, 6].

In recent years, the development of big data and data transmission technologies bring a bright perspective for efficient utilization and protection of vehicle batteries. By uploading their operation data to a cloud platform or data center, the EV batteries can be better monitored and managed by using advanced algorithms [7-9]. A cloud-based battery management framework is established in [10] based on end-edge cloud technology. With the developed cloud computing platform, the performance of battery state-of-X estimation and thermal management systems can be significantly improved. In [11], the Internet of Things technology is employed to upload the measured battery operation data to the data center. A cloud-based digital twin management system is established to estimate the state of charge (SoC) and state of health of lithium-ion and lead-acid batteries. The data-driven machine learning and deep learning algorithm are further employed in [12] and [13] to estimate battery SoC values of EVs. Experimental results highlight the effectiveness of artificial intelligence algorithms in improving the accuracy, adaptability, and robustness of battery state estimation. Big-data platforms and artificial intelligence algorithms provide a new solution for efficient vehicle battery monitoring and management. However, data quality greatly impacts the performance of cloud battery management and monitoring systems [14, 15]. Different from conventional data collection and transmission systems, bad data frequently appears in cloud-based battery management platforms due to the mobility of EVs and the harsh working condition of onboard sensors. Therefore, a reliable data transmission network and a high-quality cloud battery dataset are indispensable to enable this benefit [16]. To the best of the author's knowledge, no work has been carried out to systematically solve data quality problems in cloud-based vehicle battery monitoring and management by developing a novel integrated battery data cleaning framework.

This paper aims to bridge the aforementioned research gap and proposes a novel data cleaning method for improving the quality of the vehicle battery data in cloud-based battery management systems. A novel integrated battery data cleaning framework is designed, which is able to comprehensively assess the quality of the battery data and restore the bad samples. In the first stage, the outlier samples are detected by analyzing the temporal features in the battery data time series. Then, the abnormal samples, including the noise polluted data and missing value, are restored by a novel future fusion data restoring model. The real electric bus operation data collected by a cloud-

based battery monitoring and management platform are used to verify the performance of the developed data cleaning method. Experimental results revealed that the established data quality assessment and restoring models are able to detect and reconstruct the dirty data accurately.

1.1. Literature review

Time-series analysis and regression analysis are the most commonly used data cleaning methods in engineering applications, and their characteristics are summarized in Table I. Data cleaning is realized by analyzing temporal dependence in the dataset in time-series analysis methods [17-19]. The autoregressive moving average algorithm is used in [20] to analyze and correct the error and noise in high-frequency velocity data in measuring devices, and experimental results on several industrial datasets showed that the developed moving average method could effectively assess the quality and fix the errors in the collected data. In a further study, a time series analysis-based data cleaning and repairing framework is carried out in [21] for improving the quality of probe vehicle data. The exponential smoothing method is used in their work to detect and restore the errors in the collected vehicle speed dataset, and simulation results revealed that the data quality could be significantly improved for meeting the traffic-state measure requirement. Time-series analysis method has been proved effective for repairing mistakes in single-property data by utilizing the temporal dependence information in it. However, unlike conventional data cleaning issues, the battery operation dataset consists of four time series: terminal voltage sequence, current sequence, SoC sequence, and temperature sequence [22, 23]. Each of them can be regarded as an independent time series but with low autocorrelation, and thus conventional time-series analysis method can hardly capture the temporal dependence relationship [24].

Table I. Summarization of data cleaning methods in the existing literature.

Data cleaning method	Literature	Model dependence	Temporal dependence	Deep features
Time-series analysis	[17], [18], [19], [20], [21]	×	✓	×
Regression analysis	[25], [26], [27], [28], [29]	✓	×	×
Feature-fusion method		✓	✓	✓

Although autocorrelation information in battery operation data can hardly be utilized in data cleaning issues directly, the model dependence information provided by the battery mathematical model provides a new solution for data cleaning [30-32]. The regression analysis has been recognized as one of the most effective ways to analyze the model dependence relationship between different variables in data cleaning. Paper [25] developed a data cleaning method for improving the quality of power equipment condition monitoring dataset based on the random forest regression algorithm. The missing data restoration is modeled as a multiple regression problem, and simulation results indicate that it can correctly identify the abnormal data and accurately fill the

missing data. In [26] and [27], the backpropagation algorithm and support vector machine are further used to improve the quality of wind power data and power grid monitoring data, where the data cleaning task are both resolved by the regression analysis method.

Regression analysis methods can detect and fix the bad data by utilizing the model dependence information [28, 29]. However, compared to mechanical and electrical systems, the battery is an integrated electrochemical system with complex external characteristics [33, 34]. Conventional regression analysis methods cannot deeply excavate the model features in the battery dataset, and thus the data restoring accuracy and stability is usually unsatisfactory. Recent developed artificial intelligence brings a bright perspective to the battery data cleaning issue [35, 36]. In [37], the deep neural network is used to excavate battery model features in the dataset to enhance battery management systems' performance. Further, Deep-long-short-term-memory (Deep-LSTM) algorithm [38], which is specially designed for excavating the temporal features in time series; and Denoising Autoencoder (DAE) algorithm [39], which specializes in deeply excavating the model dependence relationship between different series, have been proved effective in complex system state estimation and prediction issues. However, to the best of the author's knowledge, no published works have studied the use of deep learning methods in vehicle battery data cleaning.

1.2. Contribution and innovation

This paper aims to get around the above difficulties and proposes a novel data cleaning method for improving the quality of the collected vehicle battery operation data in cloud-based battery management systems. The main contribution of this paper can be summarized as follows:

1) To the best of the authors' knowledge, this paper is the first effort to systematically analyze and solve data quality problems in cloud-based vehicle battery monitoring and management
2) A novel integrated battery data cleaning framework is designed, which is able to comprehensively assess the quality of the battery data and restore the bad samples. With the developed framework, data quality in cloud-based vehicle battery management can be significantly ensured and improved. Compared to conventional data cleaning methods, deep features in the dataset can be better utilized to improve the sensitivity and accuracy of the established model.
3) A novel data quality assessment model is established by analyzing the temporal features in the battery dataset. Compared to conventional time series analysis methods, not only the autocorrelation but also the cross-correlation in the battery dataset can be utilized to boost model sensitivity. With the developed method, the outlier data in the dataset can be accurately detected to avoid their impacts on battery monitoring and management.

4) It further develops a novel data restoring model for improving the integrity of the collected battery dataset. By using both the temporal and model dependence features, the abnormal data, including the noise polluted and missing data, can be accurately reconstructed.

Furthermore, the theoretical and practical significance of the developed methodology can be summarized as follows:

1) The proposed data cleaning method provides an effective data quality assessment tool in cloud-based vehicle battery management, which can further perfect the design theory and boost the practical application of the vehicle big data platform and Internet of vehicle technology.
2) The established data restoring model brings a bright perspective for improving the accuracy and stability of the cloud-based battery model and further promotes the efficient utilization and protection of vehicle batteries.

1.3. Organization of the paper

The rest of the paper is organized as follows: The developed integrated battery data cleaning framework is described in Section 2. Section 3 and 4 present the developed future-oriented data quality assessment and data restoring models, respectively. The performance of the developed battery data cleaning method is illustrated in Section 5, followed by concluding remarks in Section 6.

2. Integrated battery data cleaning framework

The data transmission process in a cloud-based vehicle battery monitoring and management platform is shown in Fig. 1. Firstly, the battery operation state of road EVs is real-time estimated by onboard battery management systems (BMS). Then the collected battery operation data, including terminal voltage, current, SoC, and temperature, are uploaded to the cloud platform through a data transmission network for further analysis. However, the collected data may be polluted by the errors and noises in the following sectors:

1) Data collection sector: On the one hand, the collected data may be impacted by the abnormal operation state of the battery pack. The data collected under the fault operation state of the electrochemical system negatively influences battery modeling. On the other hand, error and noise occur in BMS data collection and state estimation processes also greatly impact the quality of the collected dataset.
2) Data transmission sector: The link between the cloud and vehicle is performed by the communication network: EVs' real-time battery operation data is uploaded to the cloud through the Internet of vehicle and cellular network technologies. The data transmission error, noise, and missing also poison the quality of collected battery data.

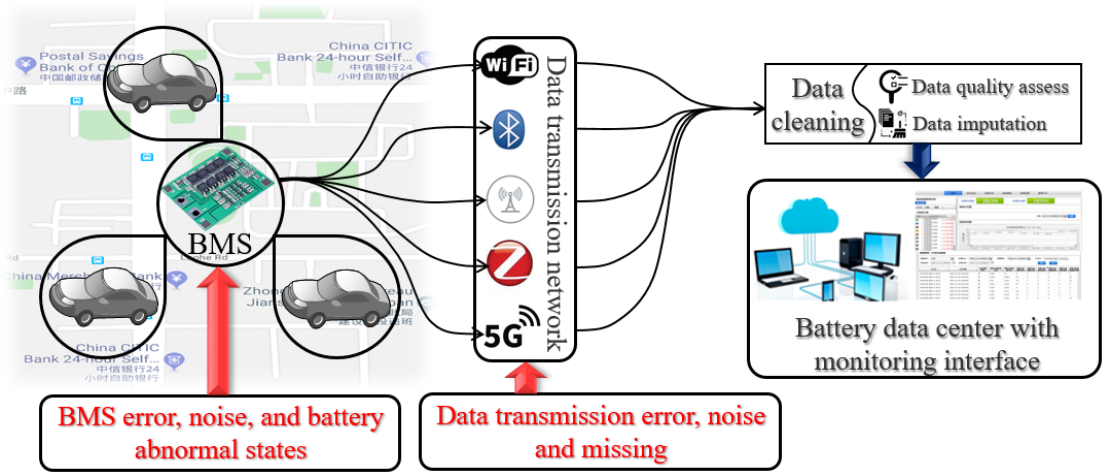

Fig. 1. Data collection and transmission in cloud-based vehicle battery monitoring and management platform.

Therefore, it is necessary to detect and clean the bad data from the battery operation dataset before data mining. Data cleaning in battery data mining will achieve two objectives. Firstly, when bad data occurs, the data cleaning scheme should detect outlier samples to avoid their influence on the data mining process. Meanwhile, it is also necessary to restore the bad data as much as possible to guarantee the integrity of the dataset. This section proposes an integrated data cleaning framework for vehicle battery big data platforms. As shown in Fig. 1, the developed battery data cleaning framework consists of two stages: data quality assessment and data restoring.

In the first stage, a data quality assessment model is established to detect the bad data in the database by analyzing temporal features in battery data time series. As shown in Fig. 1, the temporal dependence information in battery current, SoC, and temperature time series are analyzed and extracted by the LSTM unit for further analysis. Compared to original data in the time domain, the extracted deep features can better reflect the time dependence information in battery operation data time series. Then, the residual analysis method is used to assess the quality of the battery data. The battery terminal voltage series owns more stable characteristics compared to the current series, which can improve the stability of the data quality assessment model; while compared to battery SoC, the terminal voltage sequence better reflect battery external characteristics, which can boost the sensitivity of the model. Therefore, the terminal voltage is selected as the observation (output) variable, and a sequence-to-sequence regression model is established based on the extracted temporal features from battery current, SoC, and temperature time series. The outlier data is detected by analyzing the residual error of the built sequence-to-sequence regression model.

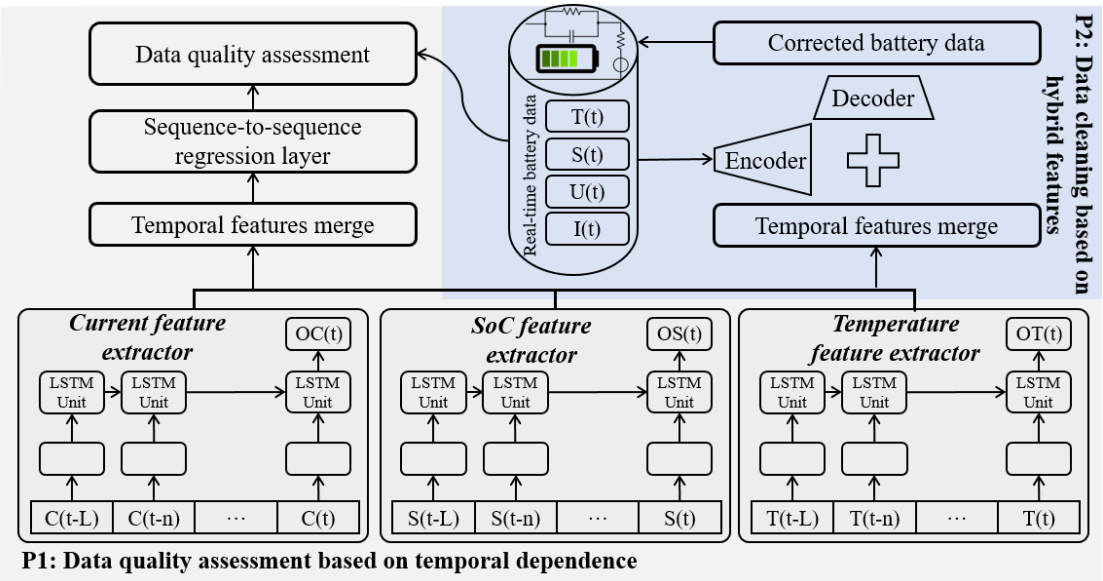

Fig. 2. Integrated battery data quality assessment and restoring framework.

In the second stage, the outlier samples are restored by comprehensively analyzing both temporal and model features in the battery database. As shown in Fig. 1, an encoder is designed to process the real-time battery data, with which the model features that reflect battery external characteristics can be extracted for data restoring. Furthermore, the temporal dependence characteristics extracted in the first stage are also used to fix the bad data in the developed data restoring model. Battery external characteristics change in various working conditions, such as under different temperatures and discharging current scenarios. The introduction of temporal features provides the learning model with additional battery historical state information, which can further boost the adaptability of the established data restoring model. With the extracted temporal and model features from the battery real-time and historical operation dataset, the data restoration is modeled as a sample generation and reconstruction process by a decoder network.

With above data cleaning framework, the bad data can be detected and restored timely, and thus the quality of the collected battery data can be significantly improved. In the rest part of the paper, the detailed mathematical principle when establishing the data quality assessment model and data restoring model will be introduced.

3. Data quality assessment by analyzing temporal features

In our work, the deep recurrent neural network (Deep-RNN) is used to assess the quality of the collected battery data by analyzing the temporal features in time series. The three most important battery external characteristic variables, including the temperature, current, and SoC, are extracted from the dataset and used as the input vectors of the Deep-RNN model, and the training input matrix $\mathcal{FT}_{t,L}$ can be presented as:

$$\mathcal{FT}_{t,L}=\begin{bmatrix} I(t-L) & \cdots & I(n) & \cdots & I(t) \\ S(t-L) & \cdots & S(n) & \cdots & S(t) \\ T(t-L) & \cdots & T(n) & \cdots & T(t) \end{bmatrix} \tag{1}$$

Where: I, S, and T are battery current, SoC, and temperature state sequence; L is the length of the input variable. $\mathcal{FT}_{t,L}$ is used as the input of the Deep-RNN network shown in Fig. 2.

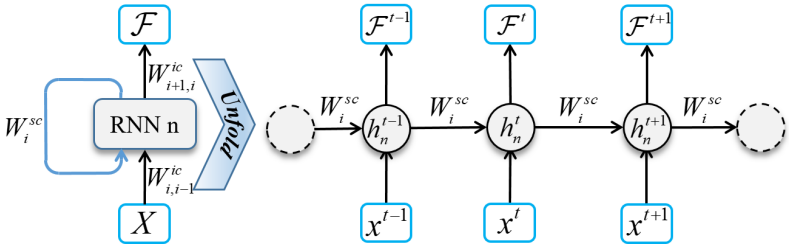

Fig. 3. Structure and topology of the deep recurrent neural network.

The temporal features in the above three time series are captured and mapped to the feature domain by the following equations:

$$H_i^t = b_i + W_i^{sc} \bullet H_i^{t-1} + W_{i,i-1}^{c} \bullet H_{i-1}^{t} \tag{2}$$

$$\mathcal{F}_i^t = f_{act}(H_i^t) \tag{3}$$

Where: W_i^{sc} is self-connection weight, which is used to reflect the temporal dependence within the time-series; $W_{i,i-1}^{c}$ the connection weight between the neurons in different layers, which is used to transmit the extracted temporal features in the deep network. H_i^t is the extracted features from the input data, with which the multi-time step temporal dependence information in time-series can be reflected by numerical results. $\mathcal{F}_i^t$ is the standardized output after processed by the neuron activation function f_{act}. The long short-term memory (LSTM) unit [40] is further employed in this study to boost the performance of the established temporal feature extraction model. Three different gates are further deployed to control the information flow in the RNN network, as shown in Fig. 4.

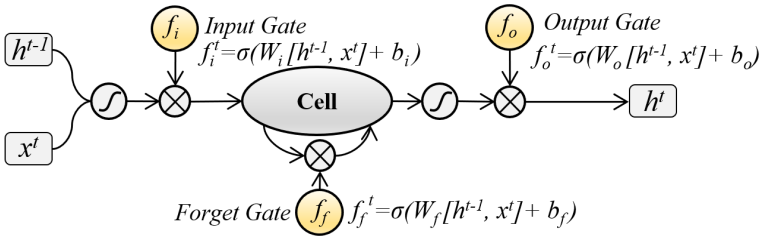

Fig. 4. Information flow in the long short-term memory unit.

To further improve the performance of the data cleaning model, we set several hidden layers in Deep-LSTM network. With the feature extraction process between multi-network layers, the temporal dependence in the battery operation data sequence $\mathcal{FT}_{t,L}$

can be transferred and presented more clearly in the feature domain. The extracted battery current, SoC, and temperature features on the top layer are labeled as OC, OS, and OT, respectively. To simplify the subsequent processing, the above three vectors are merged together to generate a feature matrix $\mathcal{F}_h^t$, which can be represented as:

$$\mathcal{F}_h^t = \left[OC_1^t \quad \cdots \quad OC_L^t \quad OS_1^t \quad \cdots \quad OS_L^t \quad OT_1^t \quad \cdots \quad OT_L^t \right] \tag{4}$$

The battery terminal voltage sequence is selected as the output of the network to predict correct battery voltage states based on the extracted temporal information. A fully connected regression layer is set on the top of the Deep-LSTM network, and the whole network is trained as a sequence-to-sequence regression model. The model training target is defined to better regress the battery terminal voltage time-series $\mathcal{TAR}_{t,L}$:

$$\mathcal{TAR}_{t,L} = \left[U(t-L) \quad \cdots \quad U(n) \quad \cdots \quad U(t) \right] \tag{5}$$

Where: U is the observed battery terminal voltage value at t.

After the Deep-LSTM model is fully trained, its sequence-to-sequence regression error is used to assess the battery data quality in real-time:

$$\hat{e}_{t+1} = \frac{\left| U(t+1) - \hat{U}(t+1) \right|}{\max\{\mathcal{TAR}_{t,L}\} - \min\{\mathcal{TAR}_{t,L}\}} \tag{6}$$

L samples in the historical battery operation data sequence are used as the model input to estimate battery terminal voltage $\hat{U}(t+1)$, and its difference with the observed value $U(t+1)$ is used to assess data quality. If the calculated relative estimation error is large, the corresponding sample is judged as bad data; while if it is within the threshold, the corresponding sample is recognized as normal data.

4. Feature-fusion based data restoring model

The established Deep-LSTM model can extract and analyze the temporal information in battery data time series, but it is not enough for data restoring. This section further establishes a novel battery data restoring model based on the feature fusion method. As shown in Fig. 3, both the temporal and model features in the battery data are utilized in the developed data restoring model, and the fixing of the battery data is carried out by a data reconstruction process.

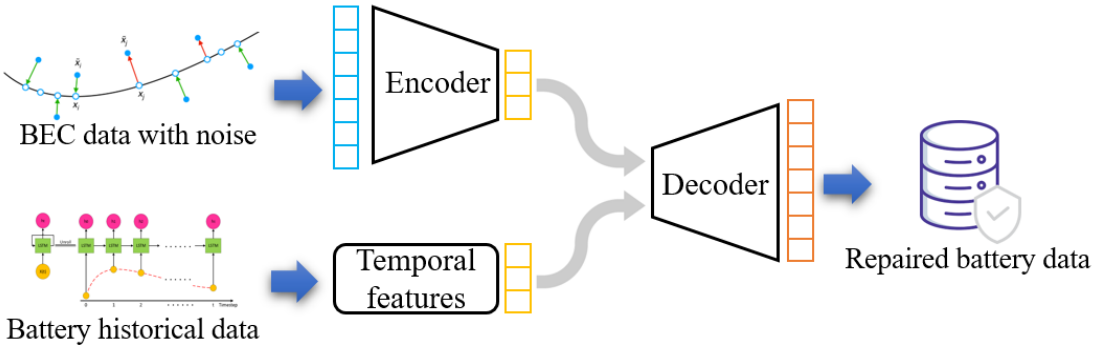

Fig. 5. Battery data restoring model based on the feature-fusion method.

The model-dependent relationship between battery real-time current, SoC, temperature, and terminal voltage states should be extracted and utilized to restore the damaged samples. DAE algorithm is one of the most commonly used methods for extracting the hidden features in the dataset and restoring the corrupted data from noise and fault. In this study, the DAE algorithm is employed to fix the bad data in the battery database. As shown in Fig. 3, the battery model features are firstly extracted from battery external characteristic (BEC) data by an encoder, and the training model input X and output Y can be depicted by the following equation:

$$X = \left[I(t) \quad U(t) \quad S(t) \quad T(t) \right] + w_k \tag{7}$$

$$Y = \left[I(t) \quad U(t) \quad S(t) \quad T(t) \right] \tag{8}$$

$$w_k \sim (0, Q_k) \tag{9}$$

Where: w_k is the white gaussian noise (WGN) added in the model input, which is used to enhance the data reconstruction capability of DAE model. The intensity of WGN is assigned as Q_k. A three-layer DAE is employed in this paper to extract the battery model features from the dataset, including an input layer, a hidden layer, and an output layer. The training target of DAE is to reconstruct the input data while filtering the noise information, which can be described by the following equations:

$$\hat{Y} = f_\theta(X) = s(W_e X + b_e) \tag{10}$$

$$KL(Y \,||\, \hat{Y}) = Y \log \frac{Y}{\hat{Y}} + (1-Y) \log \frac{1-Y}{1-\hat{Y}} \tag{11}$$

Equation (15) gives the forward propagation process of the network, W_e and b_e are the DAE parameters. The training target is to minimize the difference between the model reconstruction result $\hat{Y}$ and the original sample Y. As described in equation (16), the relative entropy method is used as the loss function of the established DAE. The front part of the network, including all the weights and biases between the input layer and hidden layer, is further separated from the trained DAE to generate an encoder

to extract the battery model features from the battery dataset. The output of the encoder MF can be represented as:

$$\mathcal{MF} = \begin{bmatrix} \mathcal{F}_m^1 & \mathcal{F}_m^2 & \cdots & \mathcal{F}_m^n \end{bmatrix} \tag{12}$$

Where: $F_m(n)$ is the output of the neuron n. With the established DAE, the corresponding battery model feature can be extracted. However, battery external characteristics also relate to its previous discharging behaviors. For example, battery terminal voltage drops transiently after experiencing a high current discharging scenario, making the data restoring process difficult. Therefore, in our work, the temporal features are also used in the established data restoring model to improve its accuracy and stability further. The trained Deep-LSTM based data quality assessment model is directly used as the temporal feature extractor. Combining with model-based features in (17) and temporal features in (4), the training input $\mathcal{F}_{TRAIN}$ of data reconstruction model can be depicted as:

$$\mathcal{F}_{TRAIN} = \begin{bmatrix} \mathcal{F}_h^1 & \mathcal{F}_h^2 & \cdots & \mathcal{F}_h^n \\ \mathcal{F}_m^1 & \mathcal{F}_m^2 & \cdots & \mathcal{F}_m^n \end{bmatrix} \tag{13}$$

A fully connected layer is used as the decoder to reconstruct and fix the bad samples, and the normal battery data without noise in (13) is used as the output to train the data restoring model.

5. Results and discussion

The cloud-based vehicle battery monitoring and management platform established in our previous work [41] is used in this study to collect battery operation data. As shown in Fig. 5, the operation data of battery packs in EVs are uploaded to the cloud platform to generate a cloud battery database and realize cloud-based battery management. In this study, we mainly focus on detecting and restoring the dirty data that is polluted by noise and missing, and battery operation data of a 10 m electric bus designed by Yutong bus Co., Ltd is used to verify the effectiveness of the proposed data cleaning method. The rated voltage and capacity of the studied Lithium iron phosphate battery pack are 480V and 199.4kWh. In this section, the performance of the established data quality assessment model will be firstly evaluated, then the data restoring experiments are carried out to fix the bad data.

Fig. 6. The cloud-based vehicle battery monitoring and management platform.

5.1. Data quality assessment results

Five different data pollution conditions are considered in our simulation, including voltage anomaly, current anomaly, SoC anomaly, temperature anomaly, and data missing. Two noise intensity levels: 3% and 7%, are added to normal samples to simulate the noise and interference situation in the data collection and transmission process.

The performance of the developed battery data quality assessment model is evaluated under data missing and noisy pollution scenarios. As shown in Table I, four different outlier data detection methods: cluster analysis [42], support vector machine (SVM) [43], RNN, and the developed feature-fusion method, are compared by their sensitivity under different preset noise intensities. The threshold of all detectors is set as 5% in 5 experiments; therefore, samples should be judged as normal data when noise intensity is lower than 5%, while as bad data when noise intensity is higher than 5%.

TABLE II. Performance comparison of different data quality assessment models

Fault types	Noise intensity	Cluster analysis	SVM	RNN	Feature-fusion
Voltage fault	3%	13.81%	9.47%	4.90%	2.58%
	7%	86.12%	90.70%	93.91%	97.34%
Current fault	3%	12.62%	7.24%	5.09%	3.75%
	7%	87.27%	88.85%	91.54%	96.69%
SoC fault	3%	14.95%	8.75%	5.46%	3.49%
	7%	84.96%	86.49%	91.76%	96.15%
Temperature fault	3%	19.51%	14.33%	9.64%	7.85%
	7%	61.44%	69.81%	76.25%	83.07%
Data missing		98.75%	100%	100%	100%

The cluster analysis model and SVM model are not able to clearly differentiate the bad data and normal data. As shown in Table I, when the added noise is only 3%, nearly 15% and 10% of samples are wrongly judged as bad data. The misdiagnosis

phenomenon in data cleaning process could result in serious information loss in the battery database because some valuable normal battery data are wrongly abandoned. Therefore, it is necessary to improve the accuracy of the quality assessment model to detect the bad data more clearly. Compared to cluster analysis and SVM methods, the bad data can be better filtered by RNN and feature-fusion methods; the reason is the temporal-dependent information in battery operation data series can be better utilized. With RNN and Deep-LSTM methods, the rate of detection mistakes can be limited to 6.3% and 4.4%, indicating that the established data quality assessment model can clearly differentiate the bad and normal data.

In bad data detection experiments, noise with higher intensity (7%) is added to normal samples. As shown in Table I, classification-based data cleaning methods, including cluster analysis and SVM methods, achieve a similar accuracy when detecting the bad data. However, only 79.9% and 83.9% of polluted data can be filtered from the dataset because of lucking temporal information. Compared with classification-based methods, the RNN method can better differentiate the bad data, and the detection accuracy is improved to 88.4% on average in voltage anomaly, current anomaly, SoC, and temperature anomaly scenarios. The developed feature-fusion method is further employed to further improve the sensitivity of the established data quality assessment model. Compared to the RNN method, the temporal information can be better excavated, and the bad data detection accuracy is further improved to 93.3% in four different data pollution experiments, which validates the effectiveness of the developed data cleaning method.

It should be figured out that the established model can better detect voltage anomaly compared to SoC and current. The reason is that the voltage is selected as the observation variable while the SoC and current are input variables in the built data quality assessment model. The noise in the observation variable is reflected in regression error directly, so the built model is more sensitive when detecting the noise in battery terminal voltage. Further, the established model shows an inferior performance when detecting temperature anomalies; the reason is that the temperature shows a limited and indirect influence on battery external characteristics. The detection accuracy only reaches 83.07%. Four methods show similar performance when detecting missing data in the battery database. In voltage missing, SoC missing, current missing, and temperature missing scenarios, nearly 100% of bad samples can be successfully filtered from battery operation data.

5.2. Performance evaluation of data restoring model

The performance of the developed data restoring model is evaluated under six different data cleaning cases: voltage missing (Case 1), current missing (Case 2), SoC missing (Case 3), voltage noise (Case 4), current noise (Case 5), and SoC noise (Case 6). The performance of the developed feature-fusion method is compared with conventional regression analysis and time-series analysis methods.

In data reconstruction experiments (Case 1 to 3), it is assumed that one of the battery external characteristic parameters is polluted by noise with 7% intensity. As shown in Fig. 7, the data reconstruction can be realized by both the conventional and developed feature-fusion methods. In the regression analysis method, data construction errors reach 3.32%, 4.71%, and 3.42% in cases 1 to 3, respectively. The reason is that the battery's complex external characteristics can hardly be accurately simulated through conventional regression analysis. Model accuracy can be improved by analyzing temporal features in the dataset with the time-series analysis method. Compared with the regression analysis method, data reconstruction error can be reduced by 19.7% on average. In noise polluted data reconstruction experiment, both the model and temporal features in the battery dataset can be deeply excavated by the developed feature-fusion method. As a result, the data imputation accuracy can be further improved. The voltage, current, and SoC data reconstruction error can be limited to 0.97%, 1.42%, and 0.62%, which validates the effectiveness of the developed method.

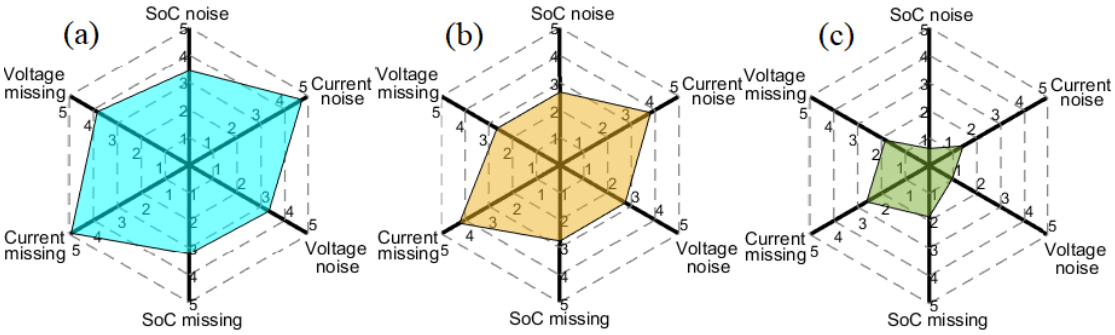

Fig. 6. Performance comparison between different data restoring methods in missing data imputation and noise polluted data reconstruction experiments. (a). Regression analysis method; (b) time-series analysis method; (c) the developed feature-fusion method.

In this study, data imputation experiments are carried out under SoC, current, and voltage missing scenarios, as shown in Cases 4 to 6 in Fig. 7. Similar to noise-polluted data reconstruction experiments, both the regression analysis and time-series analysis methods realize data imputation. However, the data imputation errors reach 3.95% and 3.22% on average in the above two methods. Compared to regression analysis and time-series analysis methods, data restoring accuracy can be significantly improved with the developed method by better-utilizing model and temporal features in the dataset. Data imputation errors are limited to 1.84%, 2.63%, and 1.86% in three cases, which validate the effectiveness of the developed method. With conventional regression analysis and time-series analysis method, model performance in Cases 1, 3, 4, and 6 is generally better than that of Cases 2 and 5. The reason is that battery dynamic characteristics reflected in the current sequence are much stronger than SoC and voltage sequences. The developed feature-fusion method can better excavate the deep features in the dataset. Therefore, the uneven performance phenomenon in different scenarios can be significantly avoided. Meanwhile, it should be figured out the developed feature-fusion method performs better in data reconstruction experiments than in data imputation

experiments. The reason is that more model dependence features are available in noise-polluted data compared with data missing situations.

Table III. Model accuracy and stability comparison of different data cleaning methods.

Methods	Data reconstruction		Data imputation	
	MAPE (%)	STD	MAPE (%)	STD
Regression analysis method	3.91	0.0536	3.89	0.0521
Time-series analysis method	3.17	0.0324	3.22	0.0355
Feature-fusion method	1.03	0.0131	2.11	0.0317

Performances of the developed battery data restoring method are further quantitatively compared with conventional regression analysis and time-series analysis methods in Table III. The regression analysis method achieves similar performance in data reconstruction and imputation scenarios. Model MAPE and standard deviation (STD) reach 3.9% and 0.053 on average, indicating the regression method's limited capability when dealing with vehicle battery data cleaning issues. Time-series analysis method shows better accuracy and stability compared with the regression analysis method. Model MAPE and STD are reduced by 18.1% and 35.9% on average by utilizing the temporal information in battery operation data. The performance of the developed feature-fusion method is further improved by better utilizing both the model and temporal features in the dataset. In data reconstruction scenarios, model accuracy is improved by 73.7% and 67.5% while stability is improved by 75.6% and 59.6% compared to regression analysis and time-series analysis methods. Further, in data imputation scenarios, model MAPE and STD can also be limited to 2.11% and 0.032, which validates the effectiveness of the developed feature-fusion method.

6. Conclusion

The deep learning algorithms and feature fusion method are employed in the paper to address the challenge of detecting and restoring dirty samples in the battery operation database. The real electric bus operation data collected by a cloud-based battery monitoring and management system is used to verify the performance of the developed data cleaning method. Through extensive simulations, the key findings are as follows:

(1) The established data quality assessment model can accurately detect the outlier samples by analyzing the temporal features in the battery data dataset. Compared to conventional cluster analysis, SVM, and RNN methods, the misdiagnosis phenomenon in the data cleaning process can be significantly avoided. Meanwhile, nearly 93.3% of noise-polluted samples and 100% of missing values can be successfully filtered from the database. With the developed method, the outlier data in the dataset can be accurately detected to avoid their impacts on battery monitoring and management.

(2) Both temporal and model features play an important role in restoring damaged battery operation data. With the developed feature fusion method, data

reconstruction accuracy is improved by 73.7% and 67.5% compared to regression analysis and time-series analysis methods. The average restoring error can be limited to 1.03% and 2.11% in noise-polluted data reconstruction and missing data imputation scenarios. With the developed integrated quality assessment and restoring framework, the quality of the collected battery operation data can be significantly improved to benefit cloud-based vehicle battery monitoring and management.

The proposed data cleaning method in this paper provides an effective data quality assessment tool in cloud-based vehicle battery management, which can further boost the practical application of the vehicle battery big data platform and Internet of vehicle technology.

Acknowledgments

This work is supported by the National Nature Science Foundation of China (No. U1864202).

Reference

[1] Y. Cao, R. C. Kroeze, and P. T. Krein, "Multi-timescale Parametric Electrical Battery Model for Use in Dynamic Electric Vehicle Simulations," *IEEE Transactions on Transportation Electrification,* vol. 2, no. 4, pp. 432-442, 2016.

[2] M. A. Hannan *et al.*, "Vehicle to grid connected technologies and charging strategies: Operation, control, issues and recommendations," *Journal of Cleaner Production,* vol. 339, p. 130587, 2022/03/10/ 2022.

[3] T. Chen *et al.*, "A review on electric vehicle charging infrastructure development in the uk," vol. 8, no. 2, pp. 193-205, 2020.

[4] W. Wen, S. Yang, P. Zhou, S. J. R. Gao, and S. E. Reviews, "Impacts of COVID-19 on the electric vehicle industry: Evidence from China," p. 111024, 2021.

[5] A. Chu, A. Allam, A. Cordoba Arenas, G. Rizzoni, and S. Onori, "Stochastic capacity loss and remaining useful life models for lithium-ion batteries in plug-in hybrid electric vehicles," *Journal of Power Sources,* vol. 478, p. 228991, 2020/12/01/ 2020.

[6] J. Bi, T. Zhang, H. Yu, and Y. Kang, "State-of-health estimation of lithium-ion battery packs in electric vehicles based on genetic resampling particle filter," *Applied Energy,* vol. 182, pp. 558-568, 2016/11/15/ 2016.

[7] X. Tang, K. Liu, X. Wang, F. Gao, J. Macro, and W. D. Widanage, "Model Migration Neural Network for Predicting Battery Aging Trajectories," *IEEE Transactions on Transportation Electrification,* vol. 6, no. 2, pp. 363-374, 2020.

[8] Y. Zhao, P. Liu, Z. Wang, L. Zhang, and J. Hong, "Fault and defect diagnosis of battery for electric vehicles based on big data analysis methods," *Applied Energy,* vol. 207, pp. 354-362, 2017/12/01/ 2017.

[9] L. Yang, Y. Cai, Y. Yang, and Z. Deng, "Supervisory long-term prediction of state of available power for lithium-ion batteries in electric vehicles," *Applied Energy,* vol. 257, p. 114006, 2020/01/01/ 2020.

[10] S. Yang *et al.*, "Implementation for a cloud battery management system based on the CHAIN framework," *Energy and AI,* vol. 5, p. 100088, 2021/09/01/ 2021.

[11] N. Dyantyi, A. Parsons, O. Barron, and S. Pasupathi, "State of health of proton exchange membrane fuel cell in aeronautic applications," *Journal of Power Sources,* vol. 451, p. 227779, 2020/03/01/ 2020.

[12] M. A. Hannan *et al.*, "Toward Enhanced State of Charge Estimation of Lithium-ion Batteries Using Optimized Machine Learning Techniques," *Scientific Reports,* vol. 10, no. 1, p. 4687, 2020/03/13 2020.

[13] M. A. Hannan *et al.*, "SOC Estimation of Li-ion Batteries With Learning Rate-Optimized Deep Fully Convolutional Network," *IEEE Transactions on Power Electronics,* vol. 36, no. 7, pp. 7349-7353, 2021.

[14] D. Markudova *et al.*, "Preventive maintenance for heterogeneous industrial vehicles with incomplete usage data," vol. 130, p. 103468, 2021.

[15] M. Dubarry and D. Beck, "Big data training data for artificial intelligence-based Li-ion diagnosis and prognosis," *Journal of Power Sources,* vol. 479, p. 228806, 2020/12/15/ 2020.

[16] N. Shehab, M. Badawy, and H. Arafat, "Big Data Analytics and Preprocessing," in *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*: Springer, 2021, pp. 25-43.

[17] G. SuryaNarayana, K. Kolli, M. D. Ansari, and V. K. Gunjan, "A Traditional Analysis for Efficient Data Mining with Integrated Association Mining into Regression Techniques," in *ICCCE 2020*: Springer, 2021, pp. 1393-1404.

[18] A. Zhang, S. Song, J. Wang, and P. S. J. P. o. t. V. E. Yu, "Time series data cleaning: From anomaly detection to anomaly repairing," vol. 10, no. 10, pp. 1046-1057, 2017.

[19] Q. Li, Y. Wang, Z. Pu, S. Wang, and W. J. T. R. R. Zhang, "Time series association state analysis method for attacks on the smart internet of electric vehicle charging network," vol. 2673, no. 4, pp. 217-228, 2019.

[20] S. Dilling and B. J. MacVicar, "Cleaning high-frequency velocity profile data with autoregressive moving average (ARMA) models," *Flow Measurement and Instrumentation,* vol. 54, pp. 68-81, 2017/04/01/ 2017.

[21] Z. Zhang, D. Yang, T. Zhang, Q. He, and X. Lian, "A Study on the Method for Cleaning and Repairing the Probe Vehicle Data," *IEEE Transactions on Intelligent Transportation Systems,* vol. 14, no. 1, pp. 419-427, 2013.

[22] Y. Gao, C. Zhu, X. Zhang, and B. J. E. Guo, "Implementation and evaluation of a practical electrochemical-thermal model of lithium-ion batteries for EV battery management system," vol. 221, p. 119688, 2021.

[23] X. Ding, D. Zhang, J. Cheng, B. Wang, and P. C. K. Luk, "An improved Thevenin model of lithium-ion battery with high accuracy for electric vehicles," *Applied Energy,* vol. 254, p. 113615, 2019/11/15/ 2019.

[24] X. Han, Z. Wang, and Z. Wei, "A novel approach for health management online-monitoring of lithium-ion batteries based on model-data fusion," *Applied Energy,* vol. 302, p. 117511, 2021/11/15/ 2021.

[25] S. Zhang, W. Yao, P. Sun, and Y. Zhang, "A Condition Monitoring Data Cleaning Method for Power Equipment Based on Correlation Analysis and Ensemble Learning," in *2020 IEEE International Conference on High Voltage Engineering and Application (ICHVE)*, 2020, pp. 1-4.

[26] Y. Mao and M. Jian, "Data completing of missing wind power data based on adaptive BP neural network," in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, 2016, pp. 1-6.

[27] W. Shi *et al.*, "Improving Power Grid Monitoring Data Quality: An Efficient Machine Learning Framework for Missing Data Prediction," in *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, 2015, pp. 417-422.

[28] A. Lew, M. Agrawal, D. Sontag, and V. Mansinghka, "PClean: Bayesian data cleaning at scale with domain-specific probabilistic programming," in *International Conference on Artificial Intelligence and Statistics*, 2021, pp. 1927-1935: PMLR.

[29] K. Rahul, R. K. J. I. J. o. I. T. Banyal, and D. Making, "Detection and Correction of Abnormal Data with Optimized Dirty Data: A New Data Cleaning Model," vol. 20, no. 02, pp. 809-841, 2021.

[30] J. Meng, D.-I. Stroe, M. Ricco, G. Luo, and R. J. I. T. o. I. E. Teodorescu, "A simplified model-based state-of-charge estimation approach for lithium-ion battery with dynamic linear model," vol. 66, no. 10, pp. 7717-7727, 2018.

[31] F. Feng *et al.*, "Co-estimation of lithium-ion battery state of charge and state of temperature based on a hybrid electrochemical-thermal-neural-network model," *Journal of Power Sources,* vol. 455, p. 227935, 2020/04/15/ 2020.

[32] R. Zhu, B. Duan, C. Zhang, and S. Gong, "Accurate lithium-ion battery modeling with inverse repeat binary sequence for electric vehicle applications," *Applied Energy,* vol. 251, p. 113339, 2019/10/01/ 2019.

[33] M. Jiao, D. Wang, and J. Qiu, "A GRU-RNN based momentum optimized algorithm for SOC estimation," *Journal of Power Sources,* vol. 459, p. 228051, 2020/05/31/ 2020.

[34] A. Farmann and D. U. Sauer, "Comparative study of reduced order equivalent circuit models for on-board state-of-available-power prediction of lithium-ion batteries in electric vehicles," *Applied Energy,* vol. 225, pp. 1102-1122, 2018/09/01/ 2018.

[35] B. Gou, Y. Xu, and X. Feng, "An Ensemble Learning-Based Data-Driven Method for Online State-of-Health Estimation of Lithium-Ion Batteries," *IEEE Transactions on Transportation Electrification,* vol. 7, no. 2, pp. 422-436, 2021.

[36] S. Li, P. Zhao, C. Gu, J. Li, S. Cheng, and M. Xu, "Online Battery Protective Energy Management for Energy-Transportation Nexus," *IEEE Transactions on Industrial Informatics,* pp. 1-1, 2022.

[37] D. N. T. How, M. A. Hannan, M. S. H. Lipu, K. S. M. Sahari, P. J. Ker, and K. M. Muttaqi, "State-of-Charge Estimation of Li-Ion Battery in Electric Vehicles: A Deep Neural Network Approach," *IEEE Transactions on Industry Applications,* vol. 56, no. 5, pp. 5565-5574, 2020.

[38] Y. Li, K. Li, X. Liu, Y. Wang, and L. Zhang, "Lithium-ion battery capacity estimation — A pruned convolutional neural network approach assisted with transfer learning," *Applied Energy,* vol. 285, p. 116410, 2021/03/01/ 2021.

[39] J. Dai, H. Song, G. Sheng, and X. Jiang, "Cleaning Method for Status Monitoring Data of Power Equipment Based on Stacked Denoising Autoencoders," *IEEE Access,* vol. 5, pp. 22863-22870, 2017.

[40] Y. Yu, X. Si, C. Hu, and J. J. N. c. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," vol. 31, no. 7, pp. 1235-1270, 2019.

[41] S. Li, H. He, and J. Li, "Big data driven lithium-ion battery modeling method based on SDAE-ELM algorithm and data pre-processing technology," *Applied Energy,* vol. 242, pp. 1259-1273, 2019/05/15/ 2019.

[42] X. Yang, G. Zhang, J. Lu, and J. J. I. T. o. F. S. Ma, "A kernel fuzzy c-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises," vol. 19, no. 1, pp. 105-115, 2010.

[43] T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, and A. J. I. T. o. I. I. Liu, "Big data cleaning based on mobile edge computing in industrial sensor-cloud," vol. 16, no. 2, pp. 1321-1329, 2019.