future internet

MDPI

Article

Analysis and Visualization of New Energy Vehicle Battery Data

Wenbo Ren [1,2,†], Xinran Bian [2,3,†], Jiayuan Gong [1,2,*], Anqing Chen [1,2], Ming Li [1,2], Zhuofei Xia [1,2] and Jingnan Wang [1,2]

[1] Institute of Automotive Engineers, Hubei University of Automotive Technology, Shiyan 442002, China; 202111205@huat.edu.cn (W.R.); 20180030@huat.edu.cn (A.C.); 202111203@huat.edu.cn (M.L.); 202011168@huat.edu.cn (Z.X.); 202111207@huat.edu.cn (J.W.)
[2] Shiyan Industry Technique Academy of Chinese Academy of Engineering, Shiyan 442002, China; xinran.bian@etu.uca.fr
[3] Information Systems and Decision Support, ISIMA, 63000 Clermont-Ferrand, France
* Correspondence: 20160013@huat.edu.cn
† These authors contributed equally to this work.

Abstract: In order to safely and efficiently use their power as well as to extend the life of Li-ion batteries, it is important to accurately analyze original battery data and quickly predict SOC. However, today, most of them are analyzed directly for SOC, and the analysis of the original battery data and how to obtain the factors affecting SOC are still lacking. Based on this, this paper uses the visualization method to preprocess, clean, and parse collected original battery data (hexadecimal), followed by visualization and analysis of the parsed data, and finally the K-Nearest Neighbor (KNN) algorithm is used to predict the SOC. Through experiments, the method can completely analyze the hexadecimal battery data based on the GB/T32960 standard, including three different types of messages: vehicle login, real-time information reporting, and vehicle logout. At the same time, the visualization method is used to intuitively and concisely analyze the factors affecting SOC. Additionally, the KNN algorithm is utilized to identify the K value and P value using dynamic parameters, and the resulting mean square error (MSE) and test score are 0.625 and 0.998, respectively. Through the overall experimental process, this method can well analyze the battery data from the source, visually analyze various factors and predict SOC.

Keywords: data visualization; KNN; SOC; vehicle battery; data analysis

Citation: Ren, W.; Bian, X.; Gong, J.; Chen, A.; Li, M.; Xia, Z.; Wang, J. Analysis and Visualization of New Energy Vehicle Battery Data. *Future Internet* **2022**, *14*, 225. https://doi.org/10.3390/fi14080225

Academic Editor: Lei Li

Received: 23 June 2022
Accepted: 21 July 2022
Published: 26 July 2022

1. Introduction

As the world is moving towards sustainable survival and development, the shortage of oil and increasingly prominent environmental pollution make research on new energy and renewable energy an inevitable trend for the development of all walks of life [1–6]. Among them, new energy vehicles have gradually become the main development object in the transportation industries of various countries, and the battery components necessary for new energy vehicles have become increasingly perfect with the continuous development of science and technology [7,8]. At present, lithium-ion batteries with low cost, small volume, and long service life have been put into production through continuous experiments and improvement, and their safety and reliability have been continuously improved [9–11]. Lithium-ion batteries have been widely used in new energy vehicles, electric bicycles, aerospace, the military, and other fields, especially in the field of electric vehicles [12,13]. However, the current lithium-ion battery has poor abuse resistance and is vulnerable to the external environment, resulting in safety-related accidents. In order to improve the utilization rate of the battery, prevent overcharge and overdischarge of the battery, prolong the service life of the battery, and monitor the state of the battery, major manufacturers have conducted in-depth research on battery technology, thus the battery management system came into being. The battery management system (BMS) will monitor the battery, including

real-time monitoring of battery physical parameters, battery state estimation and charging control. However, when relevant faults occur, the battery management system itself cannot analyze the original data generated by the battery. It can only artificially analyze the stored data and the messages in the CAN bus, and it can not find the root cause of the battery faults [14,15].

In recent years, with the continuous improvement and maturity of battery technology, the battery energy storage system (present battery maximum capacity at a certain condition is called the SOC of the battery) has been used as an important indicator to evaluate the battery state [16]. Since Li-ion batteries are renewable energy sources and intermittent in nature, the interpretation and analysis of SOC is important in the development of effective charging and discharging schemes [17], so the analysis and evaluation of battery energy storage is the top priority in the development of new energy vehicles. A previous paper [18] has conducted a detailed study on some data of new energy batteries, and introduced the cyclic neural network (RNN) to visualize and warn on battery data management; Ref. [19] proposed a method to analyze battery fault diagnosis of electric vehicles based on short-term and long-term memory networks. In reference [20], the author proposed a two-way coupled electrochemical thermal model to study and analyze the effects of water cooling liquid inlet and flow rate on the effectiveness of battery thermal management system. The original battery data and factors impacting SOC have not been explored in the aforementioned literature, despite the fact that a variety of approaches have been suggested to detect battery failure. However, the SOC of the battery is affected by many factors (vehicle state, voltage, temperature, etc.). The existing methods focus on the direct prediction of SOC but ignore the importance of analyzing and visualizing the original data. There is no practical method to analyze the factors affecting SOC.

In order to solve the shortage of existing parsing of original battery data, visual analysis, and analysis of factors affecting SOC, this paper is based on parsing the original battery data (hexadecimal) intuitively, visualizing and analyzing each index of the battery on the data, and finally using the indexes affecting SOC to realize the prediction of SOC by the KNN algorithm.

The organizational structure of this paper is as follows: Section 2 includes the relevant methods mainly used in data analysis, visual analysis, and SOC prediction. Section 3 describes the relevant data sets of the experiment and the various indicators of the data. Section 4 is divided into three parts. Part 1 describes how to visually analyze the obtained battery data; Part 2 makes a visual analysis of the analytical data obtained in the first part to find out the indicators that affect SOC; In the third part, the KNN algorithm is built for the analyzed indicators, and the SOC is predicted by comparing the selected parameters. Section 5 presents the results and analysis of the methods in Section 4. Finally, Section 6 summarizes the conclusions.

2. Related Work

Nowadays, there is little work carried out to analyze the original data of a battery, and it is very uncommon. The SOC that directly affects the battery is studied. In Ref. [21], Deng Ma proposed an adaptive tracking EKF (ATEKF) method to estimate the SOC of a battery. In Ref. [22], the authors compared machine learning methods with different characteristics to estimate the performance of battery SOC, showing that different methods of machine learning are useful for both measuring and predicting SOC. However, there is a lack of research on the original data generated by Li-ion batteries, because Lithium-ion batteries generate hexadecimal data, which are not intuitive, and the hidden voltage, current, temperature, and SOC are difficult to obtain directly.

Based on the observation results reported above, we introduced a scheme to realize the visual analysis of lithium battery data and SOC prediction from the source. The scheme helps to use abstract password-like lithium battery data to visualize the various metrics that affect battery performance and analyze them to predict SOC. Our work has two contributions: (I) through investigation and acquisition of a large number of lithium

battery data, we first make a rough analysis, and then conduct codification research to illustrate the intuitiveness and feasibility of the method; (II) the parsed data are cleaned and pre-processed, visualization studies are performed to filter out the valid data, the factors affecting SOC are analyzed in the visualization, and finally, SOC is predicted by the KNN algorithm.

3. Data

Data Analysis

- New Energy Vehicle Battery Dataset 1

The data provided include the message data obtained from the lithium battery, including protocol type, the server receiving time, message time, message type, and the original messages. We mainly extract and analyze the original messages, which include the current vehicle status, vehicle position, battery voltage, battery voltage, and engine status. However, the message data are all composed of hexadecimals, so it is difficult to directly obtain understandable data. Therefore, it is necessary to analyze and obtain intuitive related indicators to promote the next analysis. See Figure 1 for an explanation.

VIN	Internal c	license pl	agreemer	server rec	message	message	check	message	original message
			GB_T3296	2019-05-	2019-05-	vehicle lo	correct m	33	232304fe4c474a453133454138484d363132363738010008
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d
			GB_T3296	2019-05-	2019-05-	real-time	correct m	358	232302fe4c474a453133454138484d36313236373801014d

Figure 1. Power battery data format composition and original hexadecimal message.

- New Energy Vehicle Battery Dataset 2

The data set consists of one CSV file including 36 indicators of vehicle battery data (vehicle status; total voltage; cumulative mileage; total current; vehicle speed; SOC; operation mode; insulation resistance; DC–DC status; charging status; minimum voltage battery subsystem number; minimum voltage battery cell code; minimum temperature value; minimum temperature subsystem number, etc.).The collection interval is 10 s, which includes all the data of lithium battery operation, reducing the contingency of subsequent experiments. This data set is used for visual analysis and SOC prediction in Sections 4 and 5. See Figure 2 for an explanation.

time	vehiclesta	chargesta	runmodel	speed	summilea	sumvoltag	sumcurrer	soc	dcdcstatu	gearnum	insulation	maxvoltag	maxvoltag	maxbatter	minvoltag
2020/7/15 0:24	1	3	1	0	42886	371	2.1	68	1	13	8192	1	92	3.882	1
2020/7/15 0:24	1	3	1	0	42886	371.3	3.6	68	1	13	4896	1	92	3.882	1
2020/7/15 0:24	1	3	1	1.4	42886	371	4.9	68	1	13	4883	1	90	3.881	1
2020/7/15 0:24	1	3	1	2.7	42886	370.8	7.4	68	1	14	4984	1	90	3.881	1
2020/7/15 0:24	1	3	1	0	42886	371.3	2.8	68	1	14	5021	1	92	3.881	1
2020/7/15 0:25	2	3	1	0	42886	371.3	0	68	2	15	4933	1	92	3.881	1
2020/7/15 9:06	1	3	1	2	42886	371.3	3.1	68	1	13	4999	1	92	3.883	1
2020/7/15 9:07	1	3	1	3.2	42886	371	3.9	68	1	14	4999	1	92	3.883	1
2020/7/15 9:07	1	3	1	2.4	42886	371	4.4	68	1	13	4977	1	92	3.882	1
2020/7/15 9:07	1	3	1	1	42886	371	4.4	68	1	13	4948	1	92	3.882	1
2020/7/15 9:07	1	3	1	1.4	42886	371	3.6	68	1	13	4939	1	92	3.882	1
2020/7/15 9:07	1	3	1	7.9	42886	370.8	11.8	68	1	14	4946	1	92	3.882	1
2020/7/15 9:07	1	3	1	5.9	42886	371	2.6	68	1	14	4951	1	93	3.881	1
2020/7/15 9:08	1	3	1	0	42886	371.3	2.1	68	1	14	4965	1	93	3.881	1
2020/7/15 9:08	1	3	1	6.1	42886	370.8	10.8	68	1	14	4996	1	92	3.882	1
2020/7/15 9:08	1	2	1	25.7	42886	371.3	-8	68	1	14	4986	1	92	3.882	1
2020/7/15 9:08	1	3	1	13.6	42886	371	0.3	68	1	14	4987	1	92	3.883	1
2020/7/15 9:08	1	3	1	25.7	42887	370.3	18.5	68	1	14	4999	1	92	3.883	1
2020/7/15 9:08	1	3	1	20.9	42887	370.8	2.8	68	1	14	5014	1	92	3.883	1
2020/7/15 9:09	1	3	1	6	42887	371	2.3	68	1	14	5006	1	92	3.883	1

Figure 2. New energy vehicle battery dataset 2 structure.

4. Methods

Since the original data of lithium batteries are provided by new energy vehicles that all meet the production standards, all comply with the GB/T32960 standard that specifies the remote service and data format of electric vehicles. The hexadecimal messages generated by the battery are following its defined data format. In Section 4.1, the data set format, analysis method, and related algorithm structure defined in the GB/T32960 standard will be explained in detail. In Section 4.2, the new energy vehicle battery dataset 2 is used for visualization to find the factors with high SOC correlation. In the last subsection, how to design the KNN algorithm is explained.

4.1. GB/T32960 Standard Introduction and Data Format Analysis

4.1.1. Introduction to GB/T32960 Standard

GB/T32960, "technical specification for electric vehicle remote service and management system", is divided into three parts in terms of content, which are general, on-board terminal, communication protocol, and data format [23].

The general structure diagram of the electric vehicle remote monitoring system is given in GB/T 32960.1-2016, part I, general provisions. As can be seen from Figure 3, after the vehicle terminal obtains vehicle data, it uploads the data to the enterprise platform by means of CAN bus communication, and then the enterprise platform interacts with the public platform by means of CAN bus.

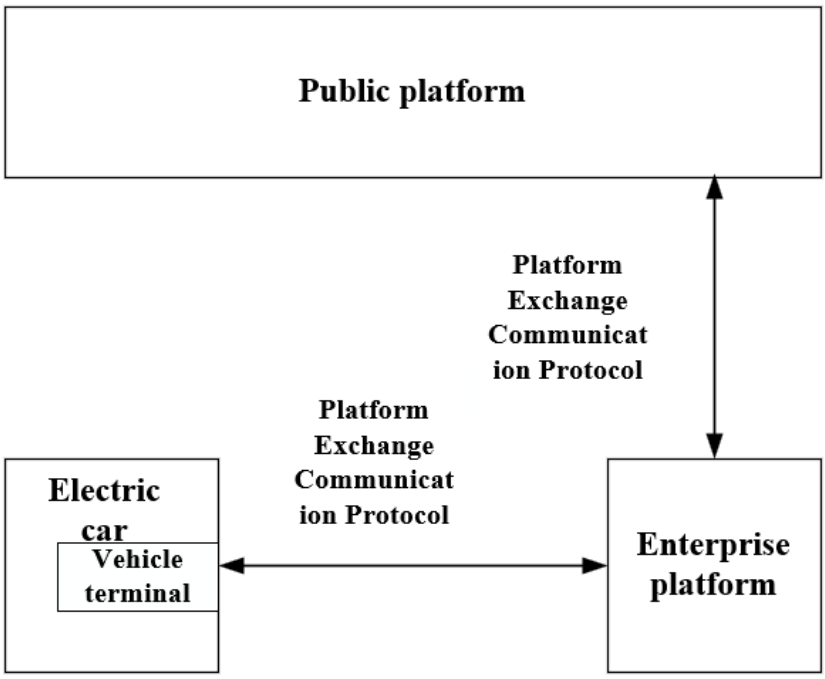

Figure 3. Structure of electric vehicle remote service and management system.

4.1.2. Data Format Analysis

GB/T 32960.3 specifies the protocol structure, communication connection, data packet structure and definition, data unit format, and definition in the electric vehicle remote service and management system. Before introducing the packet structure, first, the data types in the packet specified in the protocol are analyzed. The defined data types specify the composition of battery message information. The protocol has five data types: byte, word, dword, string, and byte[n]. It should be noted that the protocol uses the big end mode to transfer multi-byte data types.

A complete data message consists of a start, command cell, unique vehicle identifier, data encryption method, data unit length, data unit and check code. A battery packet sent

from the vehicle terminal to the server side always follows the general structure, as shown in Table 1.

Table 1. Packet structure definition.

Start Byte	Definition	Type of Data	Describe
0	Starter	STRING	Fixed to ASCII characters '#' i.e., 0 × 23, 0 × 23
2	Command ID	BYTE	See Table 2
3	Response ID	BYTE	0 × FE: command packet, received Party does not answer
4	VIN	STRING	Vehicle Unique VIN Car
21	Data encryption method	BYTE	0 × 01: Data is not encrypted 0 × 02: RSA encryption 0 × 03: AES128 encryption
22	Data unit length	WORD	Data Unit Length Total Words
24	Data unit	COMPOUND	Divided into information type flag and information body
Penultimate	Check code	BYTE	BCC XOR check

The header of the packet is first composed of two ASCII characters '#', representing the beginning of the packet. Then, the definition of response ID and command ID is shown in Table 2.

Table 2. Command identification.

Coding	Definition	Directions
0 × 01	Vehicle login	Go up
0 × 02	Real-time information reporting	Up
0 × 03	Real-time information reporting	Up
0 × 04	Vehicle logout	Go up
0 × 07	Heartbeat	Go up
0 × 08	Terminal time	Up
0 × 80	Query command	Down
0 × 81	Set command	Down

It can be seen that if it is a real-time information reporting frame, the third byte should be filled with 0 × 02. Next, you should fill in the unique vehicle identification number, namely VIN (vehicle identification number). Because the information length is not fixed, the data unit length represents the length of the next data information, so that the server can find the end of the frame when parsing. This paper mainly discusses the protocol content and packet structure involved in the most commonly used real-time information reporting as an example.

The real-time information reporting first includes the data collection time, which is represented by a 6-byte BCD code in the format of month, year, and day. Then is the information type; information type does not require the order and items can be freely combined. There are many types of information, such as vehicle data, drive motor data, fuel cell data, engine data, etc. See Table 3 for specific information type definitions. Finally, there is the message body, whose length and data type will vary depending on the type of message [23].

Table 3. Sign of information type.

Type Code	Illustrate
0 × 01	Vehicle data
0 × 02	Drive motor data
0 × 03	Fuel cell data
0 × 04	Engine data
0 × 05	Vehicle location
0 × 06	Extreme data
0 × 07	Alarm data
0 × 08	Rechargeable energy-storage device voltage Data
0 × 09	Rechargeable energy-storage device temperature data

Because there are many types of information, this paper uses the vehicle data format for example analysis. The vehicle data format, fuel cell data, drive motor data and other information types are detailed in the literature [23].

4.1.3. Analytical Thinking

From the above data format, the data packets generated by the lithium battery comply with the format shown in Table 1. Therefore, different states of different vehicles are expanded based on data in Table 1. The whole message data output is mainly divided into three types: Vehicle login, real-time information reporting, and vehicle logout. First, for the overall analysis of the idea, the detailed steps of the analysis are as follows.

1. The entire message is structured according to the structure and definition of the packet, and the message is divided into starters, command units, unique vehicle identifiers, data encryption methods, data unit lengths, data unit, and check code.
2. Judge the vehicle status (vehicle login, real-time information reporting, vehicle logout) contained in this message by the command ID in the command unit.
3. Further analyze the vehicle status in detail according to different modules defined in the data unit format. Details can be seen in Figure 4.

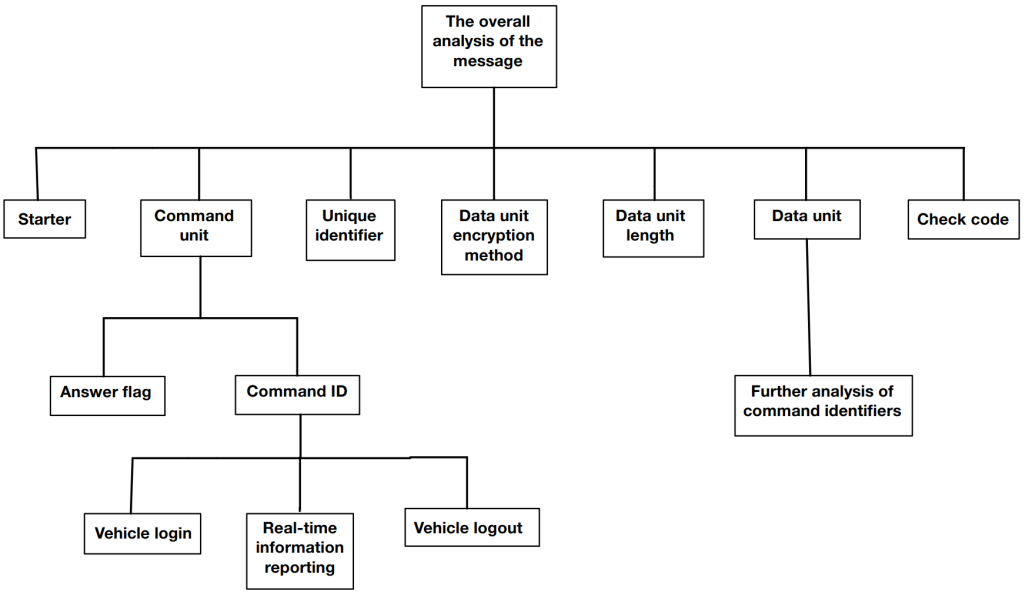

Figure 4. Global analysis.

Next, for the different types to parse, after the overall structure of the division of the hexadecimal message mainly for its command unit is further split, we look for the command ID to determine the type to which it belongs. Since the types obtained from the command unit are divided into three types, but the overall parsing structure is roughly the

same, the most complex real-time information is mainly reported. The analysis of real-time information-reporting type message is shown in Figure 5.

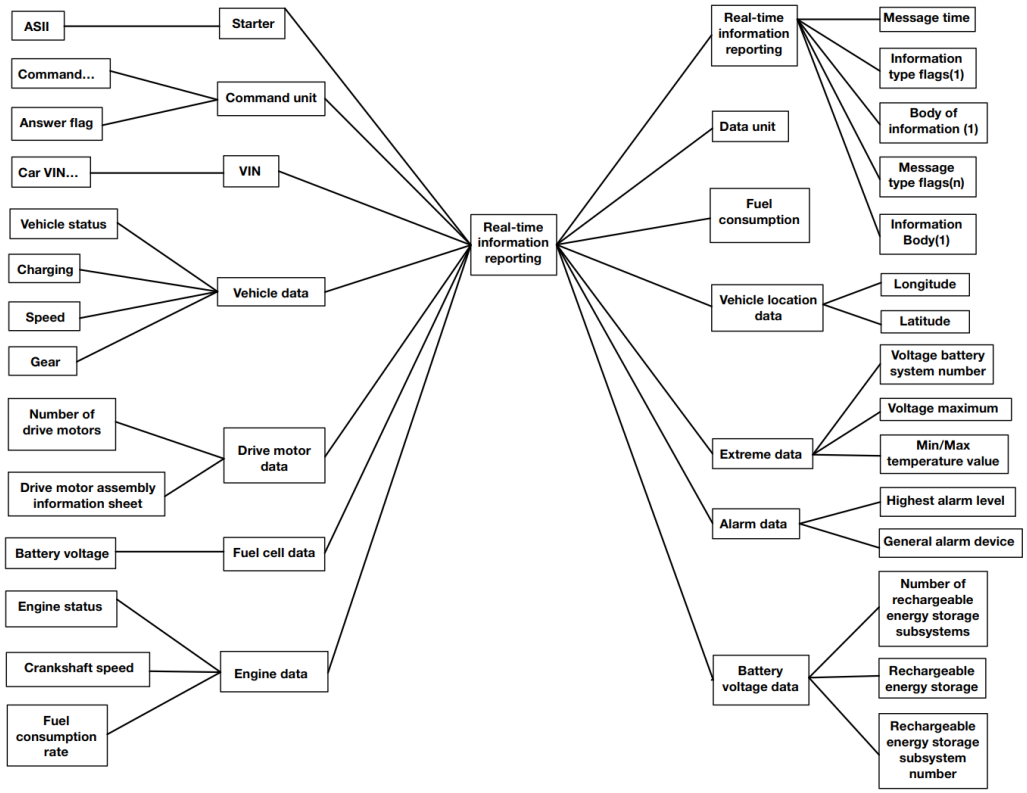

Figure 5. Analysis of real-time information-reporting type message.

The command unit in Table 1 finds 0 × 02, which indicates that this message is a real-time information report. After determining the type, the overall structure of the message is divided according to the format in which the information is reported, then finds the corresponding data type of the message according to Table 3 (data unit format definition), and finally fuses the parsing information. The real-time information-reporting message generally includes vehicle data, drive motor data, fuel cell data, position data, extreme value data, alarm data and battery voltage data. During parsing, each step is linked, and parsing is carried out based on the last two bits of the previous message (one byte corresponds to two messages).

4.1.4. Codification

After the division and analysis of Tables 1–3, the overall structure of the code-based parsing process is shown in Figure 6. The overall parsing structure is mainly divided into six modules to parse the command ID to start parsing, each parsing a module to find the first mark of the remaining messages, and grabs the command ID of the message to start the subsequent message parsing until all the parsing is complete. The whole process is similar to a workshop with six different workshops, and a car is sent to the workshop in turn to check it.

Figure 6. Analysis of the idea.

Because the message is hexadecimal, it is necessary to perform a binary conversion first. The main idea is to convert the hexadecimal to a decimal (multiply precision first,

then offset), and then the converted byte number corresponds to the description so as to achieve the purpose of parsing. "0 × FE" means exception and "0 × FF" means invalid two in consideration. A global var is defined to capture the global parsing type and pass it to the next parsing module. From the above reasoning, the algorithm structure is shown in Table 4.

Table 4. Algorithm structure.

Algorithm Structure	Parsing Algorithm
(1) Public fun	Base conversion
(2) Global Var	Define global variables
(3) Fun_01to06	Overall division analysis (Table 1)
(4) Self.nextMark	Expand with the first token of the remaining message
(5) Fun_07	Command bit parsing
(6) Self.ol	Display in columns
(7) Self. next	Identify remaining messages

Parsing process: (all the parsing is carried out according to the technical specification of electric vehicle remote service and management system part 3: communication protocol and data format). First, regardless of the vehicle status, the overall parsing (fun_01to06) is required: (start → command ID → response ID → unique identification code → data unit encryption method → data unit length). Idea: firstly, the number of bytes represented by each description is input, followed by binary conversion, using (self.ol) to display the original message in the form of columns, then (self.pj) to start parsing and display the parsing results in the form of columns (self.pl), and finally it continues to identify the remaining messages (self.next), find the first mark of the remaining messages (self.nextMark), and grab the command mark of the message to start the subsequent message parsing (self.mo). Next, the command mark bit is identified and parsed (fun_07), and the type of the whole message is parsed for subsequent parsing.

By constructing the main body, the message data to be detected are divided into the overall message structure and the type of the message (vehicle login, information reporting, vehicle logout, platform login, platform logout) is judged. Each module is detected and finally merged. See Figure 7 for details.

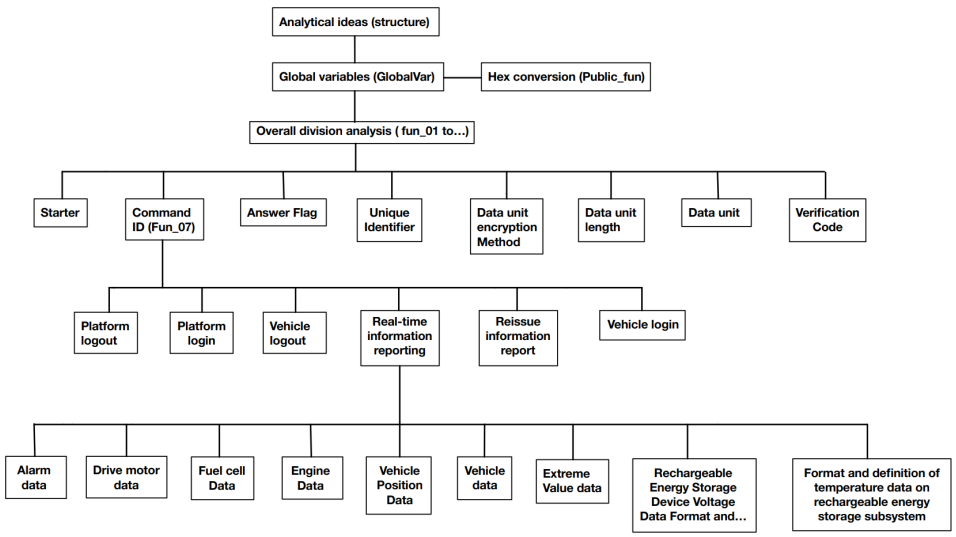

Figure 7. Analysis of code algorithm.

4.2. Visual Analysis

Based on the above Section 4.1, the abstract hexadecimal original message is parsed into intuitive data such as voltage, current, mileage, SOC, and temperature. In order to

further analyze the SOC-related factors that are crucial to the battery, we will visualize the parsed data to analyze the data of the lithium battery. Data visualization is scientific and technological research on the visual representation of data, which mainly uses graphical means to clearly and effectively convey and communicate information [24,25]. Here, the data visualization tool in Python is used to visualize the parsing [26].

Due to the rapid development of new energy vehicles, research on batteries is becoming more and more important. However, battery SOC is unable to be measured directly and can only be estimated by the parameters of the battery voltage, current and temperature, which are also affected by various uncertainties such as battery aging, environmental temperature changes and vehicle driving status, so accurate SOC estimation has become an urgent problem in the development of electric vehicles.

Data Preprocessing

Since the SOC predictions are to be made in Section 5, it is important to pre-process these metrics. The main characteristics included in the dataset are battery voltage, current, mileage, maximum temperature, and minimum temperature, which are organized as shown in Table 5. The collected data contain a total of 36 pointa and there are null values and redundant data not related to SOC. So, in order to accurately predict the battery SOC, a lot of data preprocessing is needed before the model prediction, where 80% of the work is carried out in the process of cleaning and preparing the data [27]. By using Pandas and Numpy tools to analyze the parsed battery data completely, we first filter the data for null values and outliers (combined with visual analysis) and then use the describe() function to calculate the number (count), mean, standard deviation (std), minimum (min), maximum (max), and median of battery data. Based on this, the battery data are segmented and analyzed as a whole to preliminarily understand the driving state and charging state of the car, so as to pave the way for subsequent visual analysis.

In the data cleaning stage, it is necessary to understand the missing values, duplicate values, and abnormal points in the data. The first is missing values, and there are three common data-missing situations: complete random missing, random missing, and non-random missing [28,29]. There are two types of missing values:

1. Since the missing values account for less than 10% of the total data, they can be deleted directly.
2. If the missing values account for a larger proportion of the total data, the missing data need to be filled in. The common ways of filling in are mean interpolation and regression replacement methods.
3. Outliers are missing.

Through experiments in this paper, it is found that there are zero missing values in the data set used, and the number of abnormal values is very small. However, due to different criteria for determining outliers, there will be deviations in the identification of outliers. At the abstract level, exceptions are defined as patterns that do not conform to the expected normal behavior, so a simple exception detection method is to define an area that represents the normal behavior and declare any observations in the data that do not belong to the normal area as exceptions [30].

Table 5. Data retained after pretreatment.

Index	Illustrate
Vehicle status	01: Vehicle start; 02: Turn off; 03: Other; 254: Abnormal; 255: Invalid
Charge status	01: Parking and charging; 02: Driving and charging; 03: Not charging; 04: Charging completed; 254: Abnormal; 255: Invalid
Speed	Valid value range: 0∼220 km/h, 65,534 means abnormal, 65,535 means invalid
Sum mileage	Valid value range: 0∼99,999.9 km, 4,294,967,294 means abnormal. 4,294,967,295 means invalid
Sum voltage	Valid value range: 0∼1000 V, FFFE means abnormal, FFF means invalid
Sum current	Valid value range: −1000∼+1000 A, 65,534 means abnormal, 65,535 means abnormal
Soc	Valid value range: 0∼100%, 254 means abnormal, 255 means invalid
Gearnum	Binary bits, 0–6 binary bits represent neutral gear–sixth gear, 1101 reverse gear, 1110 D gear, 1111 parking P gear
Maxbatterysinglevoltageval	Valid value range: 0∼15 V, minimum measurement unit: 0.001 V, 65,534 means abnormal, 65,535 means invalid
Minbatterysinglevoltageval	Valid value range: 0∼15 V, minimum measurement unit: 0.001 V, 65,534 means abnormal, 65,535 means invalid
Maxtmpval	Valid value range: −40∼+210 °C, minimum measurement unit: 1 °C, 254 means abnormal, 255 means invalid
Mintmpval	Valid value range: −40∼+210 °C , minimum measurement unit: 1 °C, 254 means abnormal, 255 means invalid

4.3. K-Nearest Neighbor Algorithm

From the visual analysis, it can be seen that the SOC value has a linear regression relationship with some indicators, and the correlation is high, while the KNN algorithm is very effective for classification and regression problems. So, we use the KNN algorithm to predict SOC simply and quickly.

KNN was originally an intuitive classification method and has been widely used in pattern recognition. With a little modification, it can also be effectively applied for regression purposes. However, because different models have different requirements for data, KNN can ensure better prediction results only by selecting appropriate models in combination with the characteristics of the data themselves [31]. The core idea of the KNN nearest neighbor algorithm is that if most of the K nearest samples in the feature space belong to a certain category, the sample also belongs to this category. K is usually an integer no greater than 20. The implementation of KNN classification prediction is divided into the following steps:

1. Randomly select K tuples from the training tuples as the initial nearest neighbor tuples, and calculate the distance from the test tuples to the K tuples, respectively;
2. Sort according to the increasing relationship of distance;
3. Select the K points with the minimum distance;
4. Determine the occurrence frequency of the category of the first K points;

5. The category with the highest frequency among the first K points is returned as the prediction classification of test data [32].

In this paper, the training data are divided 8:2 for training and testing, and KNN constructs a range instead of setting the most important n_neighbors, weights and P as default. Dividing the training and test sets, a grid search method is used to let the KNN algorithm itself find the optimal parameters and the highest overlap according to the data set assignment. The setting of hyperparameters affects the selection of K and P values, and here we will analyze and compare different ranges to obtain the optimal solution. The distance P is calculated by (1) and (2), which reflects the similarity of the two points before. The feature space of the K-nearest neighbor method is generally the n-dimensional real vector space Rn, and the Euclidean distance and Manhattan distance [33] are used in the distance corresponding to Equations (1) and (2), respectively.

$$d(Xi, Yi) = \left(\sum_{i=1}^{n} |xi - yi|^2 \right)^2 \quad (1)$$

$$d(Xi, Yj) = \sum_{i=1}^{n} |xi - yi| \quad (2)$$

where xi denotes the predicted value, yi is the sample value, and $|xi - yi|$ denotes the absolute value between the predicted value and the sample value [34].

To avoid the distance deviation caused by the different sizes of different features, standardization is first required in data preprocessing. The prediction process of the KNN algorithm is shown in Figure 8.

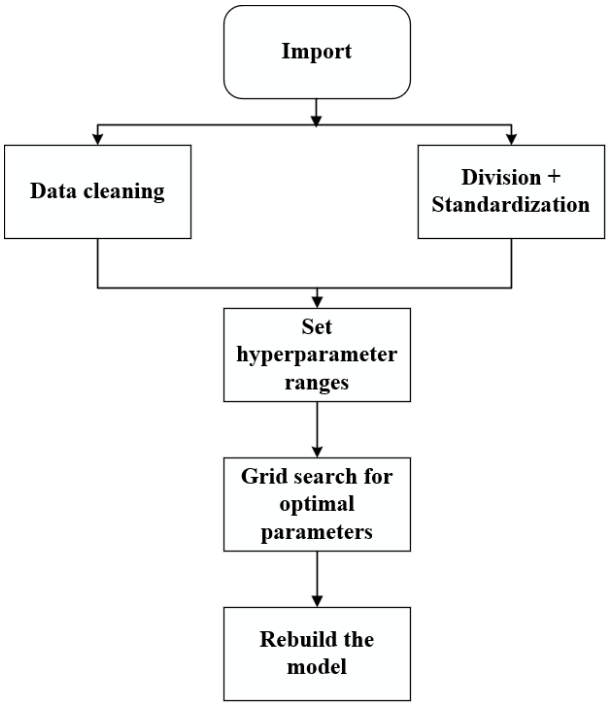

Figure 8. KNN algorithm flow.

5. Results and Discussion

5.1. Visualization

In this experiment, according to the method in Section 4.2, first we filtered and eliminated missing values and outliers, secondly, combined with correlation functions for the overall analysis of battery data, the screening analysis results are shown in Table 6, and

finally, the visual analysis results are gradually carried out by using the correlation function. Figure 9 shows the battery-related data and the visual analysis of vehicle status in turn.

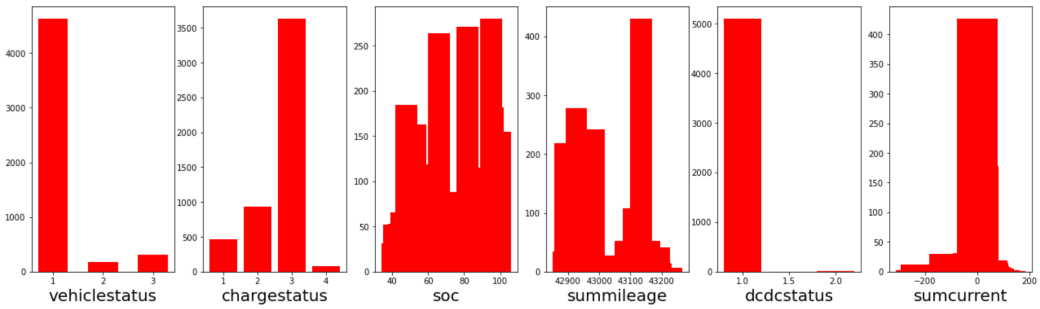

Figure 9. Battery data visualization.

Table 6. Mean, variance, maximum value, and minimum value of battery data.

	Vehiclestatus	Chargestatus	Speed	Summileage	Sumcurrent	Sumvoltage	Soc	Maxtmpval	Mintmpval
Count	5111.0000	5111.0000	5111.0000	5111.0000	5111.0000	5111.0000	5111.0000	5111.0000	5111.0000
Mean	1.1545	2.6497	23.8796	43,048.9649	−0.7088	377.0951	73.1655	25.5437	23.8385
Std	0.5004	0.6657	25.1968	101.3028	42.3631	16.3452	17.7004	3.7593	1.9647
Min	1.0000	1.0000	0.0000	42,886.0000	−233.700	347.5000	40.0000	23.0000	22.0000
Max	3.0000	4.0000	85.8000	59,776.0000	107.0000	405.0000	100.0000	255.0000	34.0000

According to the results shown in Table 6 and combined with Table 5, we can roughly learn that the vehicle is divided into three states, vehicle start, off, and other, mostly using start for driving. The vehicle charging state is for parking charging, driving charging, not charging and charging is complete. The maximum speed is 85.80 km/h, and the cumulative mileage ranges from 42,886.00 km to 59,776.00 km (16,890 km in total). SOC is more than 30%, with an average of 73%. The car's maximum temperature value is mostly below 100 °C. Minimum temperature's highest value is also 35 °C, and fluctuates within −40∼42 °C. This means that the whole vehicle is in a good driving condition and the battery is in a healthy state. Basically, it is intuitive to understand the overall data of each indicator of the car. First, we grasp the data distribution as a whole so as to prepare for the next step of detailed visualization work. Combined with Figure 10, from (a), there is no direct relationship between SOC and speed, and the speed is basically below 80 km/h. However, it can be seen that the battery goes through two processes of discharging and charging, and the vehicle goes through two states of parking charge and discharging. This can also be derived from (b) the SOC versus time graph, and (c) shows the SOC in turn with sumvoltage, minbatterysinglevoltageval and sumcurrent. It can be seen that the SOC is almost the same as the sumvoltage and minbatterysinglevoltageval, in the stage of charging the current displays a process of first falling and then rising. Then, from (d), it is seen that there is no substantial pattern between the maximum, minimum temperature and SOC curves, and the correlation is low.

Finally, combined with the thermodynamic diagram, as shown in Figure 11, the correlation between these 15 battery data indicators is further intuitively obtained, in which the correlation between minbatterysinglevoltageval, sumvoltage and SOC is 0.98, basically close to 1, showing a high correlation. Through the analysis, it can be seen that the SOC of the battery has the highest linear correlation with the minbatterysinglevoltageval and sumvoltage, indicating that they have the greatest impact on the SOC in the operation of the battery, and can better predict the SOC value. Based on the above analysis, we take these two indicators as samples for reference, use the KNN algorithm for prediction, and then add a sumcurrent as a reference to eliminate contingency.

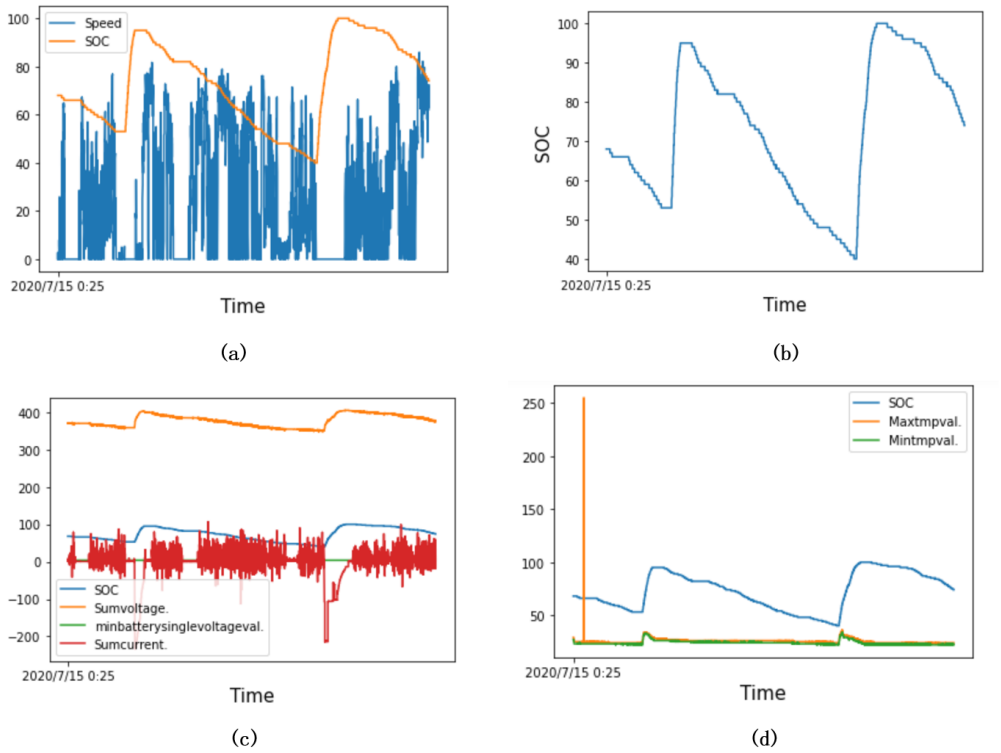

Figure 10. (**a**) SOC and speed; (**b**) SOC and time; (**c**) SOC and sumvoltage, sumcurrent, and minbatterysinglevoltageval; (**d**) SOC and maximum temperature and minimum temperature.

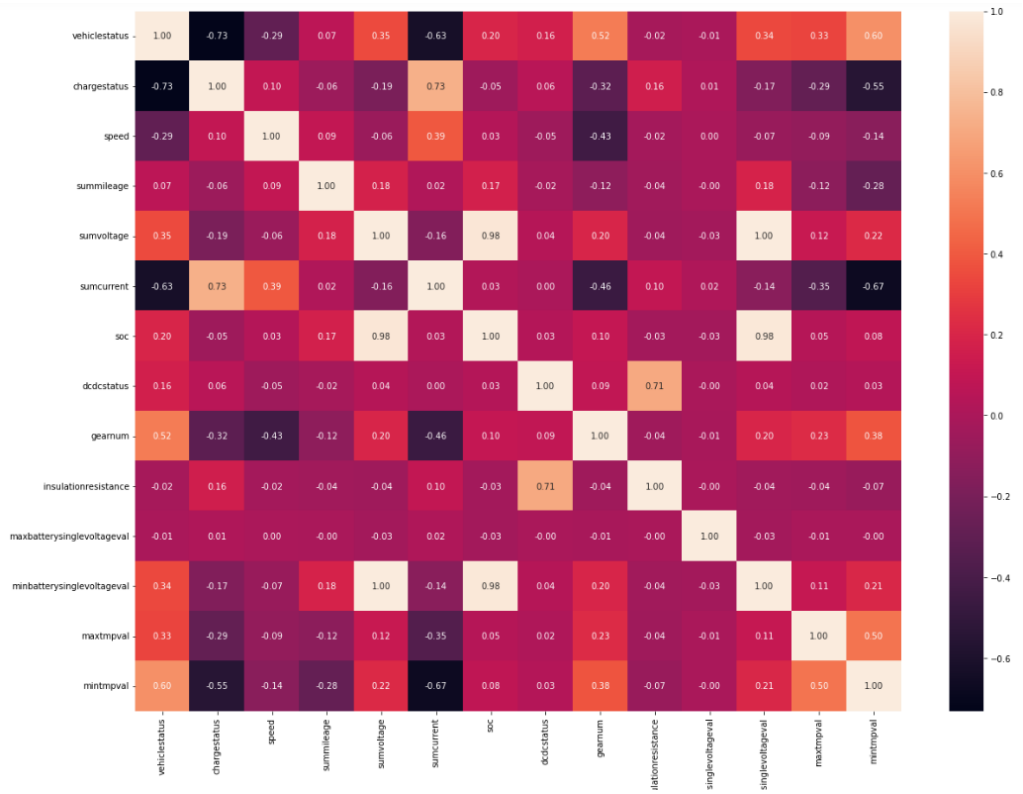

Figure 11. Heat map of each data of battery.

5.2. KNN Prediction

The article divides the data set and filters the hyperparameters by the method in Section 4.3, and selects the sumcurrent, sumvoltage and minbatterysinglevoltageval as

the prediction index to construct the model to predict SOC. The two most important hyperparameters of the KNN algorithm are K (the number of specified nearest neighbor samples) and P (the selected distance), and we design several sets of value ranges to compare this aspect of hyperparameters and use the mean square error (see Equation (3)) as the evaluation index, where y_test is the tested data and y_pre is the predicted data, so as to find the optimal parameters. See Table 7 for details.

Table 7. Comparison of results for different values of K and P.

Hyperparameter Range	K, P Value Selected	MSE	Test Score
(3, 18), (2, 8)	$K = 3$, $P = 2$	0.6257625297381677	0.9987738423863558
(5, 18), (1, 8)	$K = 5$, $P = 1$	0.5864392223608215	0.9989231053759856

It can be seen that the selected $K = 5$ and $P = 1$ ($P = 1$ is the Manhattan distance and $P = 2$ is the Euclidean distance), which are the hyperparameters of KNN and have an MSE of 0.5864. The test score arrives at 0.9989, which is 0.0394 higher than the MSE (0.6258) and test score (0.9988) with the hyperparameters of $K = 3$ and $P = 2$, so the predicted distributions fit the original data better and the accuracy of predicted SOC is slightly improved. Figure 12 below shows the prediction results obtained with different K and P values.

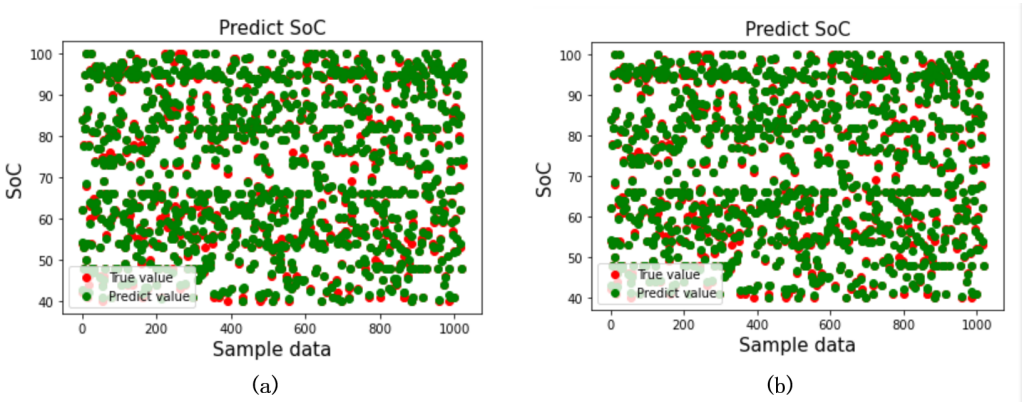

Figure 12. (**a**) $K = 3$, $P = 2$; (**b**) $K = 5$, $P = 1$.

$$MSE = \frac{SSE}{n} = \frac{1}{N}\sum_{i=1}^{N}(y_{\text{test}} - y_{pre})^2 \quad (3)$$

Red represents the SOC data in the original data and the green represents the SOC value predicted according to the key factors visually analyzed. It can be seen that regardless of the selection range of K and P values, the sumcurrent, sumvoltage, and minbatterysinglevoltageval analyzed in Section 5.1 can accurately predict the SOC value of a lithium battery.

6. Conclusions

In this paper, a new analytical method based on the original data of lithium batteries is proposed. This method analyzes the abstract hexadecimal message data generated by the lithium battery at the source end, parses it into intuitive and understandable data according to the GB/T32960 standard design algorithm, and gradually analyzes the key factors in the original data that are highly linear with SOC by using the visual method. Finally, the KNN algorithm is used to model, the key factors are taken as input, and the range of hyperparameters is set, so that the KNN algorithm can independently select parameters according to the characteristics of the data set distribution. The test score of SOC prediction will reach 0.9988, improving the accuracy of SOC prediction. After verification, the parsing method can completely parse the original battery data of the GB/T32960 standard, including three different types of messages: vehicle login, real-time information reporting and vehicle logout, with a correct parsing rate of over 95%. This

offers a fresh approach to the study of battery data. It can also be connected to the current BMS in the future for intuitive data processing and early warning judgment.

Author Contributions: W.R. and X.B. conceived and designed the experiment; W.R. and X.B. performed experiments; M.L., Z.X. and J.W. surveyed data; W.R analyzed the data and wrote papers; J.G. and A.C. supervised and operated the project. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Hubei University of Automotive Technology Doctor's Research Start-Up Fund (Grant NO. BK201604).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

SOC	State of Charge
VIN	Vehicle Identification Number
BMS	Battery Management System
KNN	K Nearest Neighbors
SSE	The Sum Of Squares Due To Erro
MSE	Mean square error

References

1. Hu, Z.; Song, Y.; Xu, Z.; Luo, Z.; Zhan, K.; Jia, L. lmpacts and Uilization of Electric Vehicles Integration into Power Systems. *Chin. Soc. Electr. Eng.* **2012**, *32*, 1–10.
2. Nicholas, L.; Daniel, S. Regulatory adaptation: Accommodating electric vehicles in a petroleum world. *Energy Policy* **2012**, *45*, 308–316.
3. Yabe, K.; Shinoda, Y.; Seki, T.; Tanaka, H.; Akisawa, A. Market penetration speed and effects on CO2 reduction of electric vehicles and plug-in hybrid electric vehicles in Japan. *Energy Policy* **2012**, *45*, 529–540. [CrossRef]
4. Shao, C.; Wang, X.; Shahidehpour, M. Partial decomposition for distributed electric vehicle charging controlconsidering electric power grid congestion. *IEEE Trans. Smart Grid* **2016**, *8*, 75–83. [CrossRef]
5. Xiangning, X.; Jianfeng, W.; Shun, T.; Qiushuo, L. Study and Recommendations of the Key Isues in Planning of Electric Vehicles'Charging Facilities. *China Electrotech.* **2014**, *29*, 1–10.
6. Al-Ogaili, A.S.; Aris, I.B.; Sabry, A.H.; Othman, M.L.B.; Azis, N.B.; Isa, D.; Hoon, Y. Design and development of three levels universal electric vehicle charger based on integration of VOC and SPWM techniques. *J. Comput. Theor. Nanosci.* **2017**, *14*, 4674–4685. [CrossRef]
7. Mehigan, L.; Deane, J.P. A review of the role ofdistributed generation (DC) in future electricitysystems. *Energy* **2018**, *163*, 822–836. [CrossRef]
8. Wang, Q.; Li, S.; Li, R. China's dependency onforeign oil will exceed 0.8 by 2030: Developing anovel NMGM-ARIMA to forecast China's foreign oildependence from two dimensions. *Energy* **2018**, *163*, 151–167. [CrossRef]
9. Li, M.; Xu, H.; Li, W. The structure andcontrol method of hybrid power source for electricvehicle. *Energy* **2016**, *112*, 1273–1285. [CrossRef]
10. Smiley, A.; Plettg, L. An adaptive physics-based reduced-order model of an aged lithium-ioncell, selected using an interacting multiple-modelKalman filter. *J. Energy Storage* **2018**, *19*, 120–134. [CrossRef]
11. Uddin, K.; Jackson, T.; Widanage, W.D. On the possibility of extending the lifetime of lithium-ion batteries through optimal V2G facilitated by an integrated vehicle and smart-grid system. *Energy* **2017**, *133*, 710–722. [CrossRef]
12. Hu, X. *Power Battery Technology and Application*; Chemical Industry Press: Beijing, China, 2012; pp. 125–258.
13. Cheng, K.W.E.; Divakar, B.P.; Wu, H. Battery-Management System (BMS) and SOC Development for Electrical Vehicles. *EEE Trans. Ions Veh.* **2011**, *1*, 76–88. [CrossRef]
14. Zhan, D.; Huang, L.; Lu, X. BMS-based control of electric vehicle battery management system. *Spec. Purp. Veh.* **2022**, *2*, 18–21.
15. Battery Management System BMS Knowledge and Functions. Available online: https://zhuanlan.zhihu.com/p/403671105 (accessed on 12 July 2022).
16. Williamson, S.S.; Rathore, A.K.; Musavi, F. Industrial Electronics for Electric Transportation: Current State-of-the-Art and Future Challenges. *IEEE Trans. Ind. Electron.* **2015**, *62*, 3021–3032. [CrossRef]

17. Zhang, R.; Xia, B.; Li, B.; Cao, L.; Lai, Y.; Zheng, W.; Wang, H.; Wang, W. State of the art of lithium-ion battery soc estimation for electrical vehicles. *Energies* **2018**, *11*, 1820. [CrossRef]
18. Rui, Z.; Qiang, L.; Sanfu, W. Research on battery data analysis of pure electric vehicles. *Autom. Instrum.* **2017**, *11*, 106–108.
19. Hongyang, L.; Lin, Y.; Jilin, L. Fault diagnosis of electric vehicle battery based on long short-term memory network. *Mechatronics* **2020**, *26*, 17–23.
20. Jaidi, J.; Chitta, S.D.; Akkaldevi, C.; Panchal, S.; Fowler, M.; Fraser, R. Performance Study on the Effect of Coolant Inlet Conditions for a 20 Ah LiFePO4 Prismatic Battery with Commercial Mini Channel Cold Plates. *Electrochem* **2022**, *3*, 259–275. [CrossRef]
21. Ma, D.; Gao, K.; Mu, Y.; Wei, Z.; Du, R. An Adaptive Tracking-Extended Kalman Filter for SOC Estimation of Batteries with Model Uncertainty and Sensor Error. *Energies* **2022**, *15*, 3499. [CrossRef]
22. Hasan, A.S.M.J.; Yusuf, J.; Faruque, R.B. Performance comparison of machine learning methods with distinct features to estimate battery SOC. In Proceedings of the 2019 IEEE Green Energy and Smart Systems Conference (IGESSC), Long Beach, CA, USA, 4–5 November 2019; pp. 1–5.
23. National Automotive Standardization Technical Committee. *GB/T 32960-2016 Technical Specification for Electric Vehicle Remote Service and Management System*; China Standard Press: Beijing, China, 2019; pp. 1–5.
24. Dou, H.; Xul, B.; Shen, F. V-SOINN: A topology preserving visualization method for multidimensional data. *Neurocomputing* **2021**, *449*, 280–289. [CrossRef]
25. Paspatis, I.; Tsohou, A.; Kokolakis, S. AppAware: A policy visualization model for Mobile applications. *Inf. Comput. Secur.* **2020**, *28*, 116–132. [CrossRef]
26. Huang, Q. Data visualization method and system implementation based on Python. *Inf. Comput.* **2019**, *14*, 137–140.
27. Dasu, T.; Johnson, T. *Exploratory Data Mining and Data Cleaning*; Wiley-InterScience: Hoboken, NJ, USA, 2003.
28. Mavridis, D.; White, I.R. Dealing with missing outcome data in meta-analysis. *Res. Synth. Methods* **2020**, *11*, 2–13. [CrossRef]
29. Golden, R.M.; Henley, S.S.; White, H.; Kashner, T.M. Consequences of Model Misspecification for Maximum Likelihood Estimation with Missing Data. *Econometrics* **2019**, *7*, 37. [CrossRef]
30. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 58. [CrossRef]
31. Min, J. Analysis and Prediction of Flight Delay Based on Data Minings. Ph.D. Thesis, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 2018.
32. Liu, J.; Yang, G. Short-term delay risk prediction of airport flights based on KNN. *J. Chongqing Jiaotong Univ.* **2021**, *40*, 12–18.
33. Wu, D.; Wang, Y.; Wu, X.; Jin, A. Research and application of K-mean-square clustering algorithm based on Euclidean distance. *Digit. Technol. Appl.* **2017**, *4*, 148–150.
34. Machine Learning Based SOC Prediction. Available online: https://blog.csdn.net/abc1234598/article/details/120027974 (accessed on 12 May 2022).