IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 15, 2018, accepted October 5, 2018, date of publication October 11, 2018, date of current version October 31, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2875514

Lithium-Ion Cell Screening With Convolutional Neural Networks Based on Two-Step Time-Series Clustering and Hybrid Resampling for Imbalanced Data

CHENGBAO LIU[1,2], JIE TAN[1], HEYUAN SHI[3], AND XUELEI WANG[1]
[1]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]School of Software, Tsinghua University, Beijing 100084, China

Corresponding author: Jie Tan (tan.jie@tom.com)

This work was supported in part by the National Nature Science Foundation of China under Grant U1701262 and the Intelligent Manufacturing New Model Application Project of the Ministry of Industry and Information Technology of the People's Republic of China under Grant 2016ZXFM06005.

ABSTRACT Due to the material variations of lithium-ion cells and fluctuations in their manufacturing precision, differences exist in electrochemical characteristics of cells, which inevitably lead to a reduction in the available capacity and premature failure of a battery pack with multiple cells configured in series, parallel, and series–parallel. Screening cells that have similar electrochemical characteristics to overcome the inconsistency among cells in a battery pack is a challenging problem. This paper proposes an approach for lithium -ion cell screening using convolutional neural networks (CNNs) based on two-step time-series clustering (TTSC) and hybrid resampling for imbalanced data, which takes into account the dynamic characteristics of lithium-ion cells, thus ensuring that the screened cells have similar electrochemical characteristics. In this approach, we propose the TTSC to label the raw samples and propose the hybrid resampling method to solve the sample imbalance issue, thereby obtaining labeled and balanced datasets and establishing the CNN model for online cell screening. Finally, industrial applications verify the effectiveness of the proposed approach and the inconsistency rate of the screened cells drops by 91.08%.

INDEX TERMS Lithium-ion cell screening, time-series clustering, resampling, convolutional neural networks.

I. INTRODUCTION

In recent years, lithium-ion batteries, which feature high energy densities, high power densities, long lives and environmental friendliness, have increasingly found widespread applications in the area of consumer electronics, such as electric vehicles (EVs), hybrid electric vehicles (HEVs) and portable power systems (PPSs) [1]–[3]. In particular, lithium-ion battery packs for EVs consist of multiple cells in series, parallel, and series-parallel configurations to satisfy sufficient energy and voltage requirements. However, because of manufacturing variability, the restrictions of production technology and tolerances, cell architecture, material defects, and degradation with use [4], [5], individual cells in a battery pack exhibit some variation in performance, even cells from the same batch that are manufactured under similar conditions, causing inhomogeneity among cells in the pack [6], [7]. Such inhomogeneity within the battery pack is embodied by intrinsic factors, including mismatches in capacity, internal resistance and self-discharge, and by extrinsic factors, such as differences in state of charge (SOC) and working voltage [5], [8]. These factors may give rise to some problems during the use of lithium-ion battery packs, such as overcharge or overdischarge of the battery pack, different attenuation velocities in the performance of individual cells, and different heat generation rates among cells [9]–[13]. Then, the problems will lead to a reduction in the available capacity and premature failure of the battery pack, which shortens its lifespan and may cause safety problems such as battery explosions [14], [15]. To relieve the inconsistency among cells in battery packs, two methods can generally be used:

a preliminary screening of homogeneous cells for a battery pack with cells configured in series and the cell balancing techniques, as well as thermal management in battery management system (BMS) [16]–[19]. Lithium-ion cell screening is a precondition for guaranteeing consistent performance among cells in the battery pack; otherwise, other methods cannot play a better role. Therefore, screening individual cells with similar electrochemical characteristics for a battery pack with cells configured in series is a crucial issue.

Lithium-ion cell screening is currently receiving considerable attention [20]–[22], which can be classified into three methods: single-parameter sorting, multiparameter sorting, and dynamic characteristics sorting. Single-parameter sorting is only based on a characteristic parameter of the cells, such as capacity screening, voltage screening or internal resistance screening. Although it is applicable to large-scale lithium-ion cell screening, the method cannot reflect the overall battery performance, and inconsistent cells may be mistakenly screened, thus reducing the performance of the configured battery packs. Multiparameter sorting uses multiple characteristic parameters to sort cells, such as static capacity, voltage, self-discharge rate, and battery internal resistance. Kim *et al.* [5], [23], [24] and Kim and Cho [14] presented a new approach based on two types of screening processes, capacity screening and resistance screening, for improved voltage/SOC balancing of a lithium-ion series battery pack. Zhang *et al.* [25] used electrochemical impedance spectroscopy (EIS) and the equivalent circuit parameters to screen the cells. He *et al.* [26] compared and analyzed the inconsistencies among commercial 18650 lithium-ion cells from five manufacturers to propose a facile consistency screening approach, which comprehensively considers the weight, size, and electrochemical performance of the cells to provide high performance consistency and reduce screening time. However, multiparameter sorting only considers the characteristic parameters of the cells under stable working conditions; without the dynamic parameters, it fails to guarantee consistent electrochemical characteristics in the screened cells. In addition, dynamic characteristics sorting evaluates the similarity of battery charge and discharge curves (voltage curves, current curves, SOC curves, internal resistance curves, and so forth) in the same environment for cell screening. Raspa *et al.* [4] proposed using self-organizing map (SOM) neural networks based on the SOC variability within each screened group of cells for the classification of homogeneous cells. Du *et al.* [27] extracted the curvature of the charge-discharge voltage curve as a feature vector to screen cells. Kim [28], [29] presented an approach to screen lithium-ion cells with similar electrochemical characteristics by two-level basis pattern recognition combined with discrete wavelet transform (DWT)-extracted features of experimental voltage signals. In fact, the voltage curves in the charge and discharge process of the cells not only directly indicate the voltage variation law while working but also indirectly reflect the variation law of the battery capacity, internal resistance, and temperature. Wu *et al.* [30] discussed the relationship between a lithium-ion battery's remaining useful life (RUL) and battery terminal voltage curves during charge process and presented an online method for lithium-ion battery RUL estimation using feed forward neural network (FFNN) and importance sampling (IS). Therefore, a screening method combining multiparameter sorting with dynamic characteristics sorting is the current research trend.

In this paper, a new approach for lithium-ion cell screening using convolutional neural networks based on two-step time-series clustering and hybrid resampling for imbalanced data (TTSCHR-CNN) is proposed. A two-stage cell screening method, namely, capacity screening and discharge voltage curve screening, is implemented in an orderly manner. First, we initially divide the cells into six classes according to their capacity range. Then, the cells in each class are used by TTSCHR-CNN to screen for defective cells in the discharge voltage curves. Generally, the discharge voltage curve of a cell can be considered as a time series. Due to the higher time and space complexity of unsupervised time-series clustering in a large number of samples, we first consider labeling the discharge voltage time-series dataset of cells as the majority class with the label ''1'' (negative samples) and the minority class with the label ''0'' (positive samples, the discharge voltage time series of defective cells) offline by using a two-step time-series clustering method and then use a hybrid resampling method to obtain a dataset with balanced positive and negative samples to train the convolutional neural network (CNN) model for online cell screening. The proposed method has been verified to effectively screen cells with similar electrochemical characteristics.

The remainder of this paper is organized as follows. Section II shows the basic concept of the problems studied in this paper and presents the screening process for screening cells with similar electrochemical characteristics. In section III, TTSCHR-CNN for lithium-ion cell screening is described in detail. Section IV analyzes and discusses the experimental results and presents an application architecture for lithium-ion cell screening in industrial production. In the final section, some conclusions and future works are presented.

II. PROBLEM DEFINITION

The manufacturing process of lithium-ion battery packs for EVs, as shown in Figure 1, can be largely divided into the cell manufacturing process, formation and screening process, and configuring battery pack process. First, the cell manufacturing process produces and assembles individual cells, including batching, coating, slitting, winding, assembling, and injecting electrolyte. Second, the formation and screening process activates the individual cells and screens for cells with consistent performance in electrochemical characteristics, including activation, removal of out-of-spec cells, and cell screening. Third, the configuring battery pack process screens a fixed number of cells with similar electrochemical characteristics to assemble battery packs for EVs or other consumer electronics. In particular, for the formation and

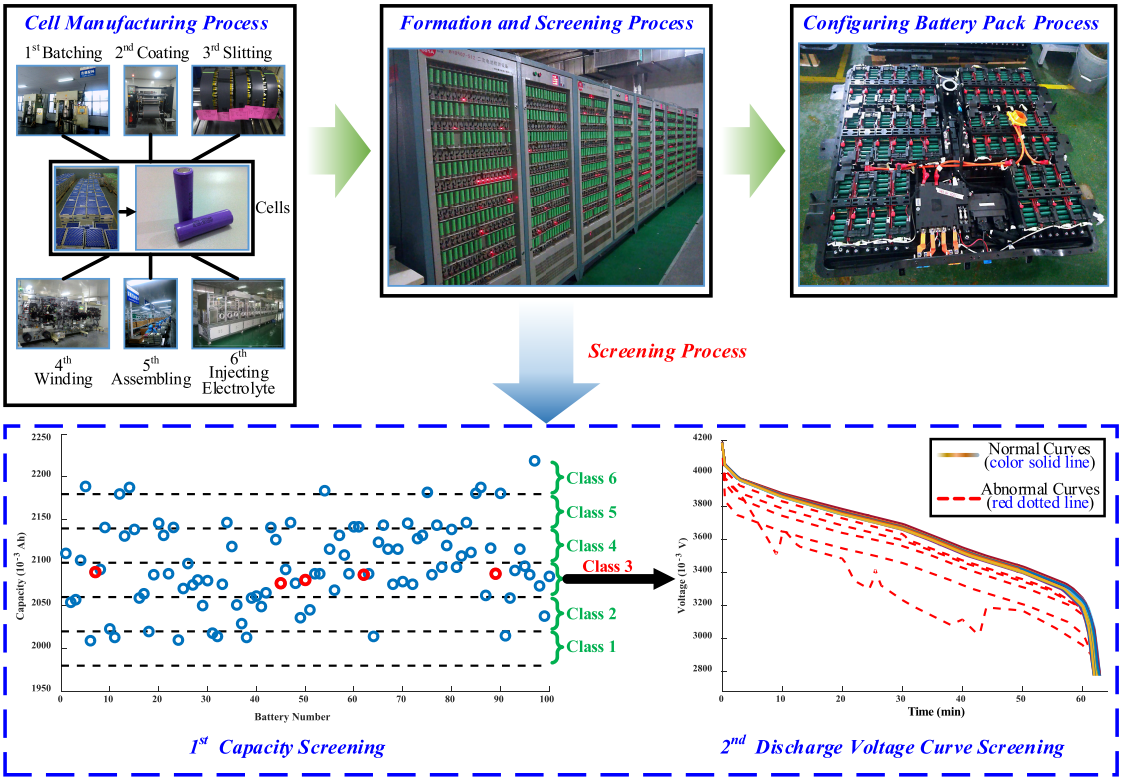

FIGURE 1. The manufacturing process of lithium-ion battery packs for EVs. A two-stage cell screening method, namely, capacity screening and discharge voltage curves screening, is implemented in an orderly manner.

screening process, the individual cells are first activated through charging, aging and discharging to form a solid electrolyte interface (SEI) film on the anode surface and have stable performance; then, faulty cells caused by internal short circuits that reduce the open-circuit voltage and discharge capacity are discarded to avoid performance or safety issues; finally, cells are screened for similar electrochemical characteristics for configuring lithium-ion battery packs.

In this paper, a two-stage cell screening method, including capacity screening and DVC screening, is implemented in an orderly manner. The capacity is defined as the maximum total electrical charge, expressed in ampere-hour (Ah), that the cells can deliver from the fully charged state to the fully discharged state [5], as expressed in

$$C_n = \int_{SOC0\%}^{SOC100\%} i\mathrm{d}t. \tag{1}$$

where C_n denotes the capacity, i denotes the discharge current, and t denotes the discharge time. In the first capacity screening process, as shown in Figure 1, 100 cells are divided into six classes according to their capacity range, every 0.04 Ah for a class from 1.98 Ah to 2.22 Ah. In the second DVC screening process, the cells in each class (class 3 in Figure 1) continue to be screened by evaluating the similarity of the discharge voltage curves, and defective cells with large deviations in their discharge voltage curves, the 5 red dotted lines in Figure 1 (their capacity distribution is indicated by the red circles), are determined to be the abnormal DVCs. This paper is focused on the second screening process: DVC screening.

Generally, the DVC of a cell can be considered a time series, which is a sequence of real-value data points with timestamps. Therefore, DVC screening is a time-series classification problem. In this paper, we denote a discharge voltage time series as $V = \{v_1, v_2, \ldots, v_n\}$, where v_i is the discharge voltage value at time stamp t_i, and there are n timestamps for each discharge voltage time series. We denote a labeled discharge voltage time series dataset as $D = \{V_i, y_i\}_{i=1}^{N}$, which contains N discharge voltage time series and their associated labels. For each $i = 1, 2, \ldots, N$, V_i represents the i^{th} discharge voltage time series, and its label is y_i. Herein, y_i is a class label in $\mathcal{C} = \{1, 0\}$, where "1" represents the normal discharge voltage time series and "0" represents the abnormal discharge voltage time series. For addressing online discharge voltage time-series classification of cells in industrial processes, the following four questions are considered in this paper.

- **How are the discharge voltage time series from industrial processes aligned in time?** Because the discharge voltage time series of cells are from a considerable amount of lithium-ion cell formation equipment, their sampling periods are inconsistent, and the discharge voltage time series of cells are not aligned in time. The time series for classification are required to be synchronized to the same reference time. Therefore, data alignment is the first problem to be addressed.

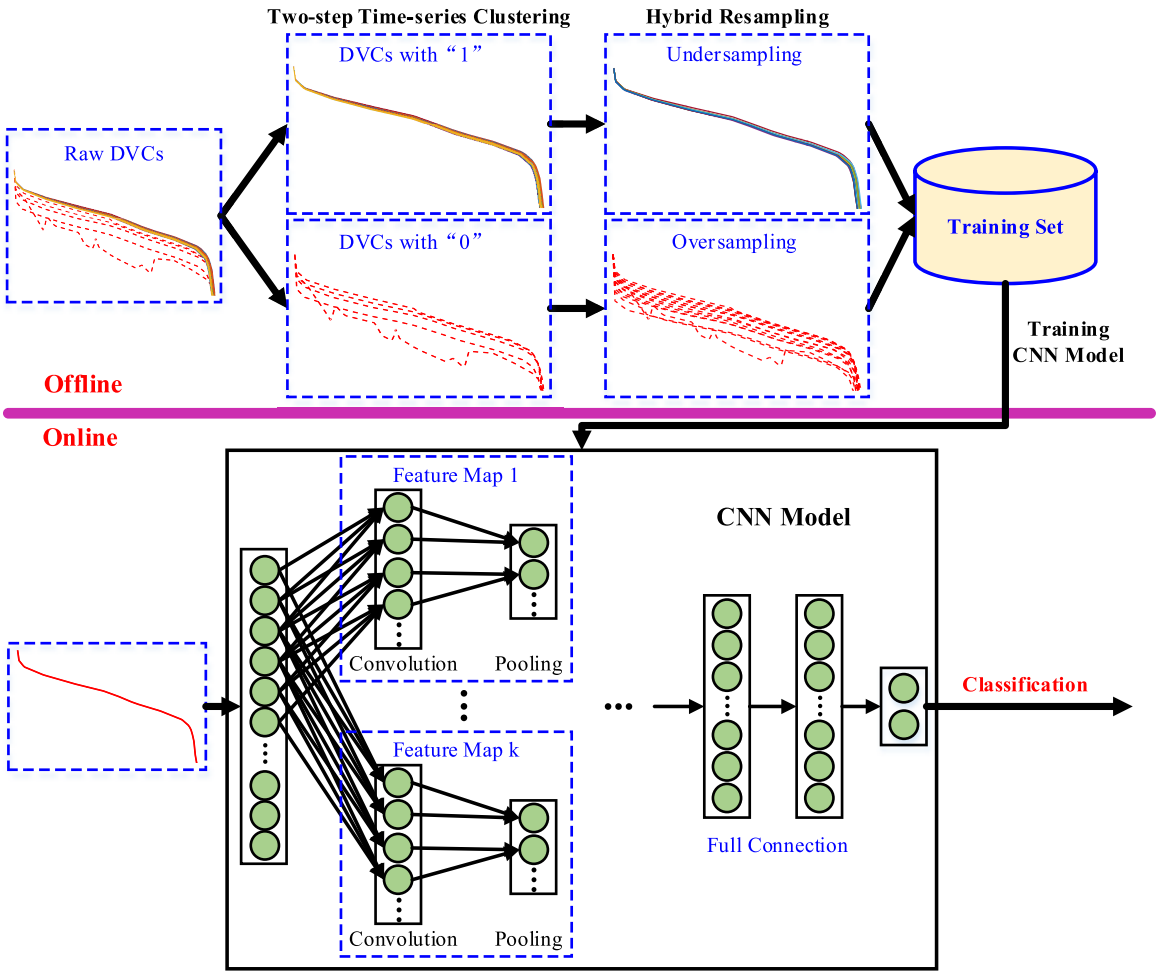

FIGURE 2. Overall architecture of TTSCR-CNN for the discharge voltage curve (DVC) screening. A new approach for DVC screening using CNN based on two-step time-series clustering and hybrid resampling for imbalanced data (TTSCHR-CNN) is proposed, where we first label the discharge voltage time-series dataset of cells as the majority class and the minority class offline by using a two-step time-series clustering method and then use a hybrid resampling method to obtain a balanced dataset to train the CNN model for online cell screening.

- **How are unlabeled discharge voltage time series labeled?** The raw discharge voltage time series are often unlabeled, and clustering techniques are considered to classify them into two classes with the label "1" or "0". However, time-series clustering objectively organizes a large amount of historical data into similar groups, which has high computational complexity and is not conducive to online screening. Thus, clustering the discharge voltage time series offline to obtain the labeled dataset to train supervised classification algorithms for online cell screening is feasible.
- **How is the imbalanced discharge voltage time-series data for classification algorithms learned?** In industrial processes, the discharge voltage time series with the label "0" are the minority compared with those with the label "1". Training classification algorithms from such imbalanced data is a formidable challenge and may lead to a minority class of samples being misjudged as majority classes, thereby reducing the classification accuracy. A resampling method for imbalanced discharge voltage time-series data is proposed.
- **Which method is suitable for online discharge voltage time-series classification of cells in industrial processes?** The discharge voltage time series are naturally high dimensional and large in data size. A highly accurate and efficient online time-series classification is challenging.

III. TTSCHR-CNN FOR LITHIUM-ION CELL SCREENING

A. TTSCHR-CNN FRAMEWORK

The overall architecture of TTSCHR-CNN for DVC screening is depicted in Figure 2 and consists of two stages: offline and online. In the offline stage, we can classify the raw DVCs into two classes by using a two-step time-series clustering method: the normal DVCs with the label "1" and the abnormal DVCs with the label "0". Due to the imbalance of DVCs with labels "1" and "0" in industrial processes, a hybrid resampling method is proposed to generate a balanced dataset

to train the CNN model for online cell screening. In the online stage, a trained CNN model is used to screen for defective cells online, where the input is a discharge voltage time series to be predicted and the output is its label.

B. DATA ALIGNMENT

The raw discharge voltage time-series data from a large amount of lithium-ion cell formation equipment in industrial processes are not aligned in time. Obtaining a time series for classification synchronized to the same reference time $T = \{t_1, t_2, \ldots, t_n\}$ is first considered. Data smoothing and interpolation are effective methods for data alignment. Specifically, interpolation methods include piecewise linear interpolation, Lagrange interpolation, Newton interpolation and Hermite interpolation, among which piecewise linear interpolation has the advantages of being simple and having only a small amount of calculations. Suppose that a raw discharge voltage time series in $\hat{T} = \{\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_l\}$ is represented by $\hat{V} = \{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_l\}$. We use piecewise linear interpolation to establish an interpolation function on each adjacent interval $(\hat{t}_i, \hat{t}_{i+1})$. The interpolation function is given by

$$V(t) = \frac{t - \hat{t}_{i+1}}{\hat{t}_i - \hat{t}_{i+1}}\hat{v}_i + \frac{t - \hat{t}_i}{\hat{t}_{i+1} - \hat{t}_i}\hat{v}_{i+1}. \tag{2}$$

where $\hat{v}_i$ denotes the discharge voltage value at time stamp $\hat{t}_i$. In this way, we can obtain consistent discharge voltage time series $V = \{v_1, v_2, \ldots, v_n\}$ in the same reference time $T = \{t_1, t_2, \ldots, t_n\}$.

C. TWO-STEP TIME-SERIES CLUSTERING

For the discharge voltage time series $V = \{v_1, v_2, \ldots, v_n\}$ obtained through data alignment processing, we denote an unlabeled discharge voltage time-series dataset as $D^* = \{V_i\}_{i=1}^N$, which contains N discharge voltage time series. The dataset D^* can be divided into two classes (the majority class and the minority class) by using time-series clustering techniques: the majority class is labeled "1", which represents the discharge voltage time series of normal cells; the minority class is labeled "0", which represents the discharge voltage time series of abnormal cells. As a challenging issue, time-series clustering has been increasingly considered in recent years, and related works are classified into three categories [31]: whole time-series clustering, subsequence time-series clustering, and time point clustering. One of the most used algorithms based on whole time-series clustering is k-means [32], where the main idea is minimizing the total distance, such as the Euclidean distance, between all time series in a cluster and their cluster centers.

In general, the DVC of a cell reflects the operating voltage characteristics of the cell to a certain extent. Each cell in the formation process is discharged with the same discharge rate C/i $(i = 0.5, 1, 2, 5, \ldots)$, where C/i represents the discharge current calculated from the nominal capacity divided by the discharge duration n in hours (h) [5]. As shown in Figure 3, the DVCs can generally be divided into three stages [33], [34],

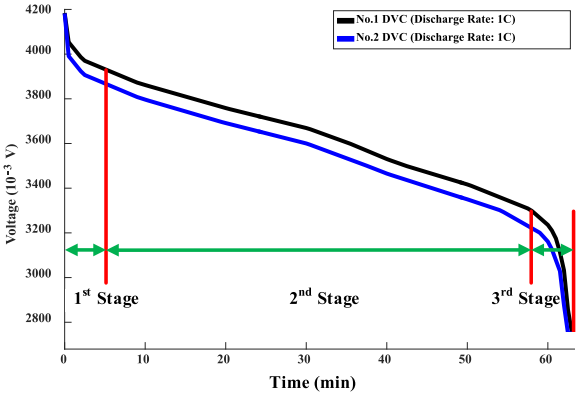

FIGURE 3. Three-stage analysis based on constant current discharge voltage curves (DVCs) of 18650 cells at a rate of 1 *C* (1 *C* = 2.0 *A*).

where the discharge voltage of the cells decreases rapidly in the first and third stages and the DVCs tend to be gentle in the second stage. In particular, in the second stage, the gentle duration depends on factors such as battery voltage, ambient temperature, battery discharge rate, battery life, and battery quality. Thus, it can be observed that a cell has better operating voltage characteristics when its DVC in the second stage is gentler and its discharge voltage value at each sampling time is higher than other cells. For example, cell No. 1 has better performance than cell No. 2 in Figure 3.

Therefore, this paper presents a two-step time-series clustering (TTSC) for labeling the discharge voltage time-series dataset D^* of cells: calculating a dividing line and redividing the dataset D^*. In the first step, we use the k-means algorithm to divide the raw discharge voltage time-series dataset D^* into k classes, where $C = \{c^{(i)}\}(i = 1, 2, \ldots, k)$ represents a set of the k cluster centers, and the value of k is selected based on experience. Based on the above three-stage DVC analysis, the k cluster centers are then divided into two parts, $C_{maj} = \{c_{maj}{}^{(i)}\}$ and $C_{min} = \{c_{min}{}^{(j)}\}$, where $c_{maj}{}^{(i)} \in C$, $c_{min}{}^{(j)} \in C$, and $|C_{maj}| + |C_{min}| = k$. Generally, the classes corresponding to C_{maj} contain most of the samples, and the classes corresponding to C_{min} contain only a few samples. Subsequently, we select the two most adjacent cluster centers that are close to the boundaries of C_{maj} and C_{min}, which are denoted as $c_{maj}{}^{(border)}$ and $c_{min}{}^{(border)}$, $c_{maj}{}^{(border)} \in C_{maj}$ and $c_{min}{}^{(border)} \in C_{min}$. A dividing line is defined as the center line of $c_{maj}{}^{(border)}$ and $c_{min}{}^{(border)}$ and is given by

$$c^{(dividing-line)} = \frac{1}{2} \times (c_{maj}{}^{(border)} + c_{min}{}^{(border)}). \tag{3}$$

In the second step, we use the dividing line to redivide the raw discharge voltage time-series dataset D^* into two parts: S_{maj} and S_{min}. The discharge voltage time-series sample V_i is classified into S_{maj} when the Euclidean distance between V_i and $c^{(dividing-line)}$ is greater than zero; otherwise, V_i is classified into S_{min}. For S_{maj}, we consider further filtering and manually screening a few abnormal discharge voltage time-series samples to add to the set S_{min}, and then the samples in S_{maj} are reduced to form a new set $S^*{}_{maj}$ and the samples

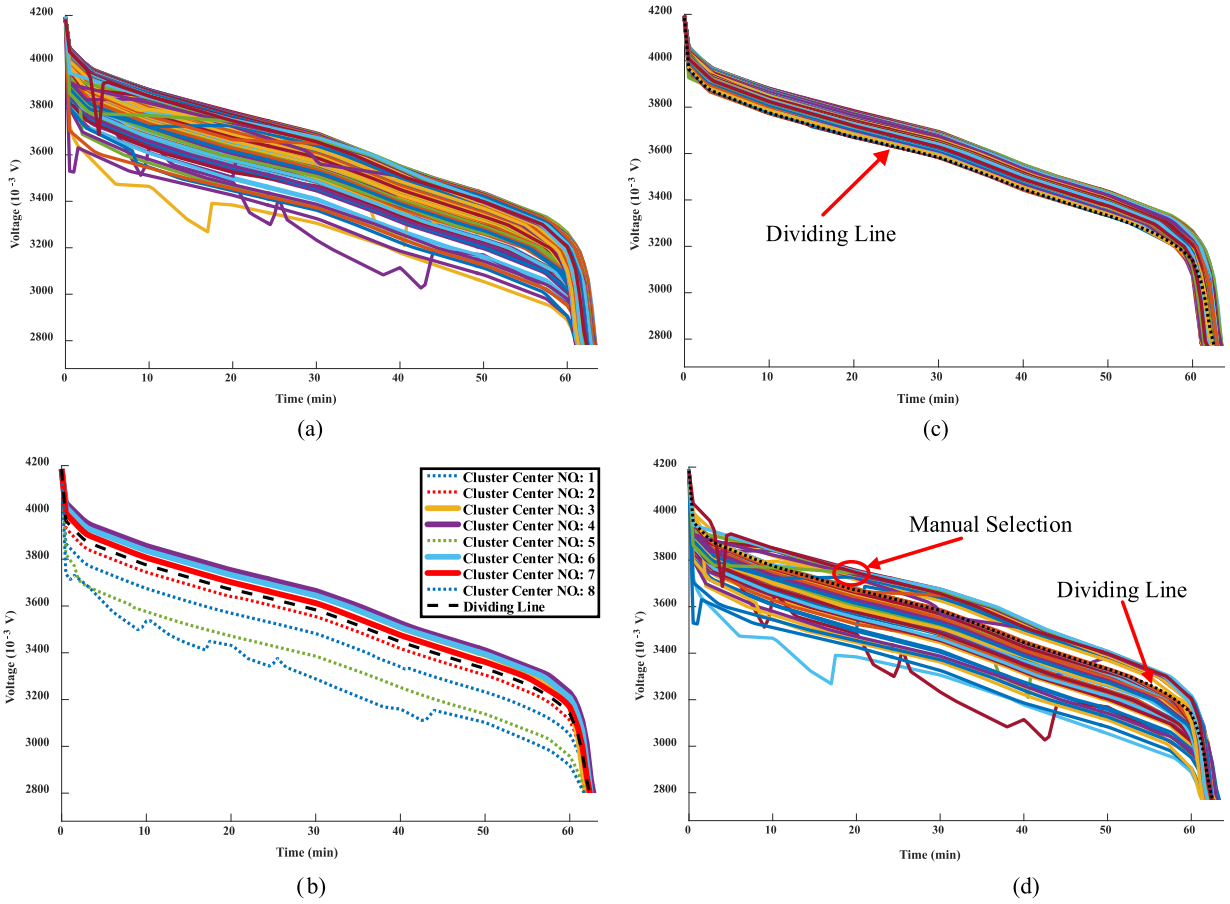

FIGURE 4. An example of the process of dividing the discharge voltage time-series dataset into the majority class with the label "1" and the minority class with the label "0". (a) DVCs of History Samples, Sample Size: 3640. (b) Cluster Center Curves of History Samples Based on K-means. (c) DVCs of the Majority Class, Sample Size: 3542. (d) DVCs of the Minority Class, Sample Size: 98.

in S_{min} are increased to form a new set S^*_{min}. Consequently, we classified the dataset D^* into the majority class S^*_{maj} and the minority class S^*_{min}, where the discharge voltage time-series samples in S^*_{maj} are labeled ''1'' and those in S^*_{min} are labeled ''0''.

An example of the process of dividing the discharge voltage time-series dataset into the majority class with the label ''1'' and the minority class with the label ''0'' is shown in Figure 4. Specifically, the DVCs of cells for 3640 history samples are shown in Figure 4(a). The k-means algorithm is used to cluster the dataset into 8 classes, and the curves of their cluster centers are shown in Figure 4(b), where the classes corresponding to four solid lines (the number of cluster centers is 3, 4, 6, and 7) contain the most samples, and the classes corresponding to the other four dotted lines (the number of cluster centers is 1, 2, 5, and 8) only contain a few samples. We select cluster centers No. 2 and No. 7 in Figure 4(b) as the two most adjacent cluster centers to define the dividing line by formula (3), which is shown in Figure 4(b) with a black dotted line. The dividing line is used to redivide the dataset, where the samples above the dividing line are classified as the majority class and the samples below the dividing line are classified as the minority class. In addition, a few abnormal DVCs of cells in the majority class are manually screened to add to the minority class. A new majority class (its sample size is 3542) and a new minority class (its sample size is 98) are shown in Figure 4(c) and (d), respectively. In Figure 4(d), the abnormal discharge voltage curves above the dividing line are screened by manual selection and circled with a red circle.

D. HYBRID RESAMPLING FOR IMBALANCED DATA

Typically, the discharge voltage time series with the label ''0'' represent defective cells and are always a minority class. Training classification algorithms from such an imbalanced dataset may lead to a small number of samples in minority classes being misclassified as majority classes, thereby reducing the classification accuracy. Consequently, resampling techniques [35], which are used to rebalance the sample space for an imbalanced dataset to alleviate the effect of the skewed class distribution in the learning process, are often performed to address imbalanced learning to obtain a balanced training set. Resampling techniques can be divided into three groups [35], [36]: oversampling methods [37], undersampling methods, and hybrid methods. First, oversampling methods are used to eliminate the risk of an imbalanced

distribution by creating synthetic minority samples. The synthetic minority oversampling technique (SMOTE) [38], [39], which creates artificial data based on the feature space similarities between existing minority samples, is a powerful method that has shown great success in various fields [35]. Second, undersampling methods [40] are used to eliminate the risk of an imbalanced distribution by discarding similar samples in the majority class. Third, hybrid methods are a combination of oversampling and undersampling methods. In this paper, a hybrid resampling method, which creates some synthetic discharge voltage time-series samples with the label "0" by using SMOTE and discards a number of the similar samples with the label "1" by using undersampling based on clustering, is proposed to obtain a balanced training set for the CNN model.

Given the discharge voltage time-series samples with the label "0" as the positive learning dataset (the minority class) and those with the label "1" as the negative learning dataset (the majority class), $P = \{V_{11}, V_{12}, \ldots, V_{1|P|}\}$ and $N = \{V_{01}, V_{02}, \ldots, V_{0|N|}\}$, where $|N| \gg |P|$, $V_{ij} \in \mathbb{R}^{n\times 1}$ and n denotes the time-series length or dimension.

For the minority class dataset P, we use an oversampling algorithm based on SMOTE [38] to create new synthetic discharge voltage time-series samples. Specifically, for each minority class sample $V_{0j} \in P$, we use the k-nearest neighbors algorithm to calculate the Euclidean distance between sample V_{0j} to all samples in P to obtain k_P nearest samples for V_{0j}; then, we randomly select n_P samples from the k_P nearest samples, and for each sample $\hat{V}_{0j}$ in n_P samples, we create a new synthetic sample according to the following formula:

$$V_{0new} = V_{0j} + \Delta \times (\hat{V}_{0j} - V_{0j}). \tag{4}$$

where $\Delta = [\delta_1, \delta_2, \ldots, \delta_n]$ and $\delta_i \in [0, 1]$ is a random number. Thus, the number of samples in the positive learning dataset is increased from $|P|$ to $n_P \times |P|$, and the new positive learning dataset is denoted by $\bar{P}$. The pseudocode of the oversampling algorithm based on SMOTE is described in algorithm 1.

For the majority class dataset N, we propose a new undersampling algorithm based on clustering to discard a portion of the similar samples to obtain a new majority class dataset $\bar{N}$. The pseudocode of the undersampling algorithm based on clustering is described in algorithm 2. Finally, there is a similar proportion in the number of $\bar{P}$ and $\bar{N}$, and a balanced training set $\{\bar{P}, \bar{N}\}$ combining $\bar{P}$ and $\bar{N}$ together is used to train the CNN model.

Algorithm 1 Oversampling algorithm based on SMOTE

Input: A monitory class dataset P and its number $|P|$, $P = \{V_{01}, V_{02}, \ldots, V_{0|P|}\}$, the number of nearest neighbors k_P, the random sampling number from the k_P nearest samples n_P;
Initialize an extended minority class dataset $\bar{P}$, $\bar{P} = P$;
$i \leftarrow 0$;
repeat
  $i \leftarrow i + 1$;
  Calculate k_P nearest neighbors of V_{0i} and randomly select n_P samples from the k_P nearest samples to form a temporary set $S_{nearest}$, where $S_{nearest} = \{\hat{V}_{01}, \hat{V}_{02}, \ldots, \hat{V}_{0|S_{nearest}|}\}$;
  $j \leftarrow 0$;
  repeat
    $j \leftarrow j + 1$;
    Generate a random vector Δ between [0,1];
    Synthesize a new sample according to the formula $V_{0new} = V_{0i} + \Delta \times (\hat{V}_{0j} - V_{0i})$;
    Add V_{0new} to $\bar{P}$;
  until $j > |S_{nearest}|$
until $i > |P|$
Output: The extended minority class dataset $\bar{P}$.

Algorithm 2 Undersampling algorithm based on clustering

Input: A majority class dataset N, k_N, the cluster number to be divided from N, and $N_N = \{n_1, n_2, \ldots, n_{k_N}\}$, a set of sample numbers to be discarded from each cluster;
Cluster all samples in the dataset N into k_N clusters, and the i^{th} cluster center is denoted by c_N^i, where $i = 1, 2, \ldots, k_N$;
$i \leftarrow 0$;
repeat
  $i \leftarrow i + 1$;
  Calculate the Euclidean distance between all samples in the i^{th} cluster with its cluster center c_N^i separately and sort them in ascending order to obtain a set S^i;
  Discard the first n_i even samples in S^i, and the remaining samples in S^i form a new set $\bar{S}^i$;
until $i > k_N$
Output: A new majority class dataset $\bar{N}$ through undersampling, where $\bar{N} = \{\bar{S}^1, \bar{S}^2, \ldots, \bar{S}^{k_N}\}$.

E. CONVOLUTIONAL NEURAL NETWORKS FOR ONLINE DISCHARGE VOLTAGE TIME-SERIES CLASSIFICATION OF CELLS

The discharge voltage time series are naturally high dimensional and large in data size. A highly accurate and efficient online time-series classification is challenging. Recently, there have been active studies on deep neural networks for time-series classification tasks. Fully convolutional network (FCN) [41], LSTM fully convolutional network (LSTM-FCN) [42], and multi-scale convolutional neural network (MCNN) [43] take advantage of CNN to address univariate time series, and a multichannel CNN [44] has been proposed to solve multivariate time series. In particular, LSTM-FCN improved the performance of FCN by augmenting the FCN module and has high performance for time-series classification. In this paper, we establish a CNN model to predict the discharge voltage time series of cells online, as shown in Figure 2. The CNN model consists of an input layer, convolutional layer, pooling layer, fully connected layer and output layer. Specifically, the convolutional

layer and pooling layer constitute a feature map, in which the convolutional layer extracts different features of the input (the discharge voltage time series) by using a convolution, and the pooling layer can reduce the size of the convolutional layer by using max pooling to avoid overfitting and improve computational efficiency [43]. Furthermore, the output of all pooling layers in k feature maps as an input continues executing the same operation to extract more advanced features [45]. Finally, all these features are integrated together through one or more fully connected layers, which is followed by a softmax classifier to predict the label distribution. In addition, the convolutional activation function in the CNN model is the rectified linear unit (ReLU) function [45], which is an unsaturated nonlinear function defined as

$$f_{cov}(x) = max(0, x). \tag{5}$$

To evaluate the effectiveness of classifiers, researchers have proposed many performance metrics [36], [46], such as *precision* or *recall*. However, for imbalanced learning problems, *precision* or *recall* cannot effectively assess classification performance, and *F-value* and *G-mean*, well-reported performance metrics, are frequently adopted to provide comprehensive assessments for imbalanced learning [46]. These metrics can be formulated by a confusion matrix, as shown in Table 1, and are defined as (6) below.

$$\begin{aligned} Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ F - value &= \frac{2 \times Recall \times Precision}{Recall + Precision} \\ G - mean &= \sqrt{\frac{TP \times TN}{(TP + FN) \times (TN + FP)}}. \end{aligned} \tag{6}$$

TABLE 1. Confusion Matrix for Performance Evaluation.

		Predicted Class	
		Positive	***Negative***
Actual Class	***Positive***	*TP (True Positives)*	*FN (False Negatives)*
	Negative	*FP (False Positives)*	*TN (True Negatives)*

IV. EXPERIMENTS AND APPLICATION

A. EXPERIMENTS

To test the effectiveness of the proposed TTSCHR-CNN method, we collect one month of discharge voltage time-series data of cells (the nominal capacity is 2.0 Ah) from a lithium battery production company in China, where the cells are manufactured under a similar environment and discharged in the same type of lithium-ion cell formation equipment to generate discharge voltage time-series data. In addition, we select 131096 cells with capacities between 1.98 Ah and 2.22 Ah as experimental samples. First, the cells are preliminarily screened to divide them into six classes according to the capacity range, every 0.04 Ah for a class from 1.98 Ah to 2.22 Ah. Specifically, there are only a few cells with capacities between 2.18 Ah and 2.22 Ah in a batch production, and the cells are manually screened by their discharge voltage curves in industrial production. Thus, this experiment is focused on five classes of cells with capacities between 1.98 Ah and 2.18 Ah, and five CNN models for online cell screening are respectively established.

Since the discharge voltage time-series data of cells directly collected from industrial processes are generally not aligned in time, they are considered to synchronize to the same reference time $t = \{0, 0.5, 1, 1.5, \ldots, 80\}$ by data alignment techniques in section III(B), where the sampling period is set to 30 seconds and the dimension of the discharge voltage time series is 161.

After data alignment, five classes of discharge voltage time series of the cells with capacities between 1.98 Ah and 2.18 Ah are labeled as the majority class samples with the label "1" or the minority class samples with the label "0" by the two-step time-series clustering (TTSC) method in section III(C), where k in the k-means algorithm is set to 10. Specifically, each class of discharge voltage time series in the five classes is classified into 10 clusters by using the k-means algorithm, and five dividing lines are defined by the two most adjacent cluster centers of each class, as shown in Figure 5(a.x). Ten clusters in each class are divided into two parts: minority clusters containing few samples and majority clusters containing most of the samples. For example, in Figure 5(a1), seven dotted lines (the number of cluster centers is 1, 4, 5, 6, 8, 9, and 10) are defined as the minority clusters, three solid lines (the number of cluster centers is 2, 3, and 7) are defined as the majority clusters, and No. 3 and No. 9 as the two most adjacent cluster centers are selected to define the dividing line. The samples below the dividing lines and the abnormal samples manually screened above the dividing lines are defined as the minority class samples, and the DVCs of the minority class are shown in Figure 5(b.x).

In addition, due to the imbalance of the majority class and the minority class samples, the hybrid resampling method in section III(D) is used to obtain balanced training sets to train the CNN models. Specifically, for the minority class samples, the abnormal samples screened by manual selection above the dividing line and a certain proportion of randomly selected minority class samples are considered to synthesize new samples for extending the minority class by using the oversampling method in algorithm 1. The majority class samples are used by the undersampling method in algorithm 2 to discard a portion of the similar samples to obtain a new majority class dataset.

Hence, the new datasets with balanced positive and negative samples are used to train the CNN models for online lithium cell screening, where the datasets are divided into training sets for learning the CNN models and test sets for evaluating the CNN models. The numbers of samples in different data preprocessing steps are shown in Table 2. The negative and positive samples of training sets for the five

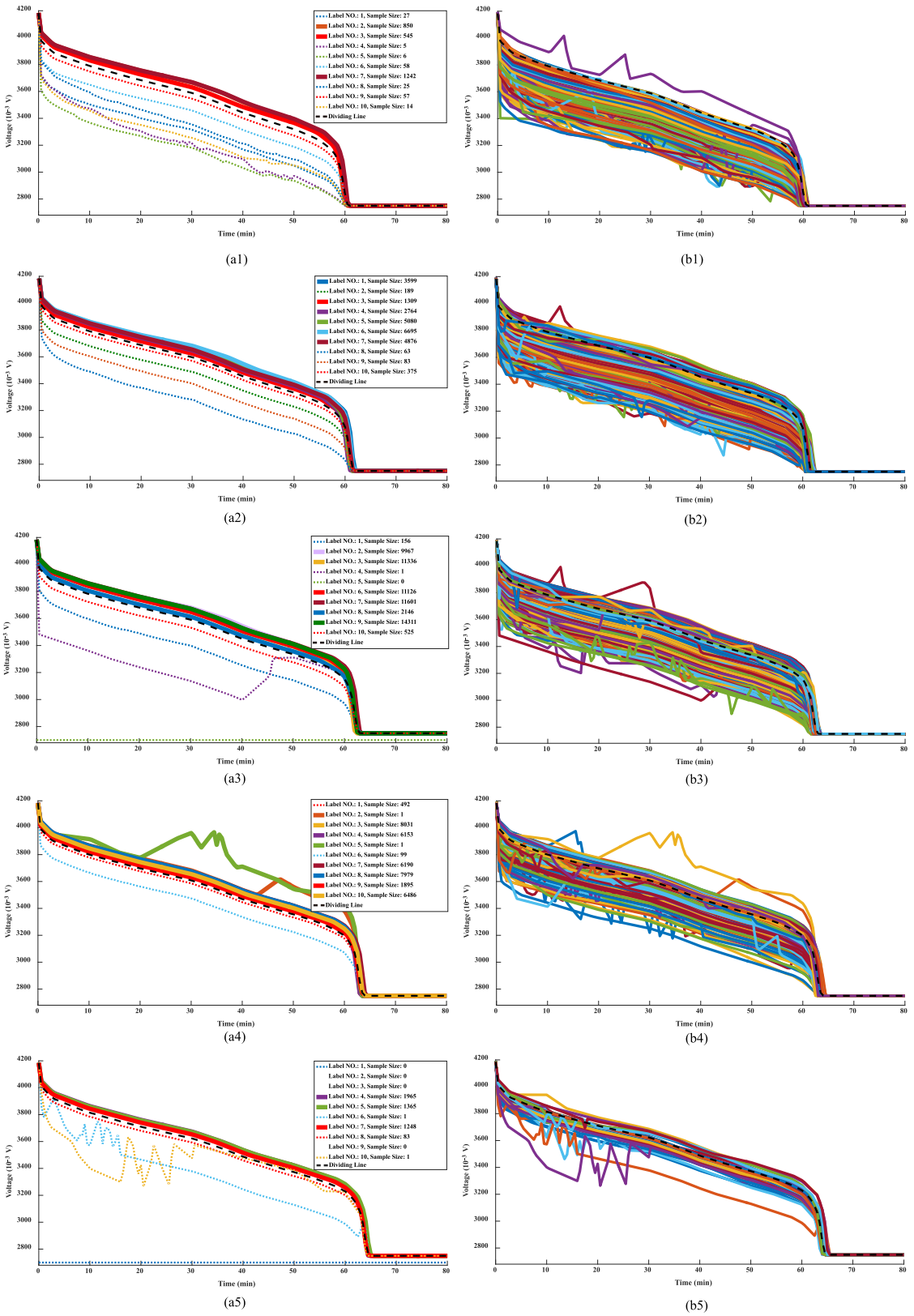

FIGURE 5. Five classes of cluster center curves and discharge voltage curves (DVCs) of the minority class divided by a two-step time-series clustering. (a1) Cluster Center Curves of cells with capacities between 1.98 and 2.02 Ah. (a2) Cluster Center Curves of cells with capacities between 2.02 and 2.06 Ah. (a3) Cluster Center Curves of cells with capacities between 2.06 and 2.10 Ah. (a4) Cluster Center Curves of cells with capacities between 2.10 and 2.14 Ah. (a5) Cluster Center Curves of cells with capacities between 2.14 and 2.18 Ah. (b1) DVCs of the Minority Class, Sample Size: 193. (b2) DVCs of the Minority Class, Sample Size: 722. (b3) DVCs of the Minority Class, Sample Size: 1063. (b4) DVCs of the Minority Class, Sample Size: 634. (b5) DVCs of the Minority Class, Sample Size: 91.

TABLE 2. The Number of Samples in Different Data Preprocessing Steps.

Capacity Range (Ah)	Capacity Screening	TTSC		Hybrid Resampling		Datasets for Training CNN			
						Training		Test	
		Majority	Minority	Negative	Positive	Negative	Positive	Negative	Positive
1.98 - 2.02	2829	2636	193	1200	309	200	209	1000	100
2.02 - 2.06	25033	24311	722	2200	1230	1200	1130	1000	100
2.06 - 2.10	61177	60114	1063	2600	1717	1600	1617	1000	100
2.10 - 2.14	37331	36697	634	2000	1116	1000	1016	1000	100
2.14 - 2.18	4663	4572	91	1120	213	120	113	1000	100
2.18 - 2.22	63	–	–	–	–	–	–	–	–

TABLE 3. The Performance Metrics of Five CNN Models on the Test Sets.

Capacity Range(Ah)	Precision	Recall	F-value	G-mean
1.98 - 2.02	0.9434	1.0000	0.9709	0.9970
2.02 - 2.06	0.9238	0.9700	0.9463	0.9809
2.06 - 2.10	0.9406	0.9500	0.9453	0.9718
2.10 - 2.14	0.9091	0.9000	0.9045	0.9444
2.14 - 2.18	0.9367	0.7400	0.8268	0.8581

CNN models in Table 2 are basically balanced. The five test sets contain 1000 negative samples and 100 positive samples. We train the CNN models using TensorFlow [47], [48], and the performance metrics of five CNN models on the test sets are summarized in Table 3. As shown, the F-value and G-mean for the first four discharge voltage time-series datasets exceed 90%. However, due to insufficient samples for the discharge voltage time-series dataset of cells with capacities between 2.14 Ah and 2.18 Ah, the F-value and G-mean are relatively low. In general, the results show that the CNN models for lithium-ion cell screening can obtain better performance for imbalanced discharge voltage time-series datasets of cells.

To further test the online prediction performance of the CNN models, we collect another month of discharge voltage time-series samples of cells (the nominal capacity is 2.0 Ah) from the same company under similar production conditions. The samples are first divided into five classes according to the capacity range, every 0.04 Ah for a class from 1.98 Ah to 2.18 Ah. We randomly select 1000 samples from five classes of samples as unlabeled test sets. The DVCs of abnormal cells predicted by the trained CNN models on the unlabeled test sets are shown in Figure 6. The results show that the CNN models can effectively predict the discharge voltage time series of abnormal cells on the unlabeled test sets.

However, a few discharge voltage time series of abnormal cells manually screened above the dividing lines (e.g., the red dotted line in Figure 6(a1) and four red dotted lines in Figure 6(a2)) are not accurately predicted due to insufficient training samples. Therefore, a high-efficiency oversampling method for abnormal samples manually screened above the dividing lines to synthesize more minority class samples can be studied further in future work.

In addition, we compare the TTSCHR-CNN method with the traditional method that currently uses capacity screening in industrial processes. An inconsistent rate of the screened cells is defined as follows:

$$Inconsistent\ rate = \frac{Number\ of\ inconsistent\ cells}{Total\ number\ of\ cells} \times 100\%. \tag{7}$$

Where inconsistent cells are abnormal cells that have not been screened by two screening methods. To show the advantages of our TTSCHR-CNN screening method, the dropped rate of inconsistent cells is given as follows:

$$Dropped\ rate = \frac{|IR_{tradition} - IR_{TTSCHR-CNN}|}{IR_{tradition}} \times 100\%. \tag{8}$$

Where $IR_{tradition}$ denotes the inconsistent rate of screened cells by using the traditional method, and $IR_{TTSCHR-CNN}$ denotes the inconsistent rate of screened cells by using the TTSCHR-CNN screening method. The results presented in Table 4 show that our method greatly improves the inconsistent rate and that the average dropped rate for five datasets drops by 91.08%, indicating that the TTSCHR-CNN method is significantly better than the traditional screening method.

B. APPLICATION

An online lithium-ion cell screening system is implemented on a big data analysis platform based on Cloudera's Distribution Including Apache Hadoop (CDH) [49], [50], which is an open-source big data analysis tool that provides the Hadoop Distributed File System (HDFS) for storing data, Hadoop MapReduce for writing parallel-processing jobs, HBase for providing real-time read/write access to huge data sets with billions of rows and millions of columns, and so on. The application architecture of online lithium-ion cell screening in industrial production is shown in Figure 7. The architecture consists of three layers: data layer, computing layer and visualization layer. First, abundant data (including the discharge voltage time series of cells) generated by lithium-ion cell formation equipment in the data layer are transmitted

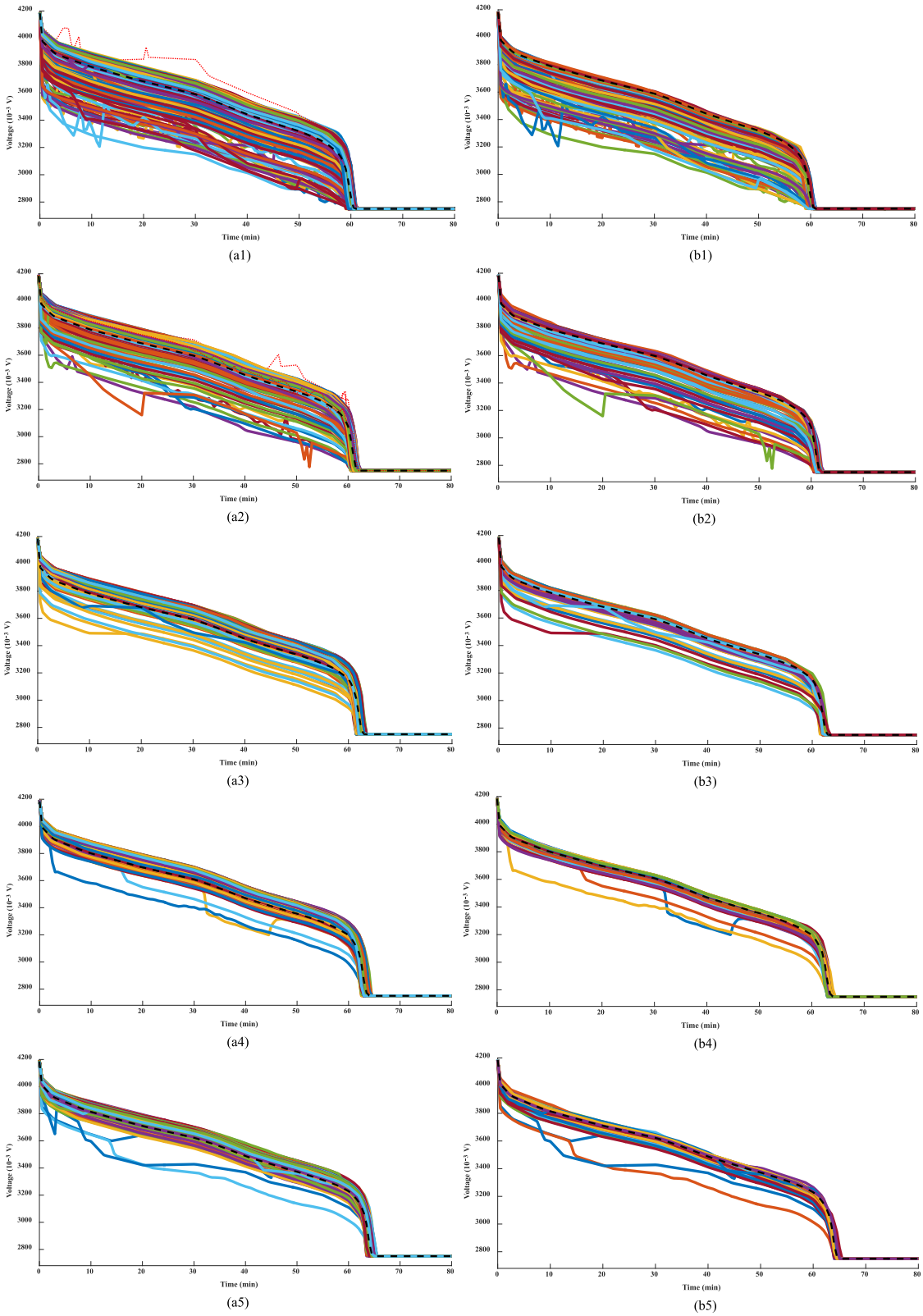

FIGURE 6. The discharge voltage curves (DVCs) of abnormal cells predicted by the trained CNN models on the unlabeled test sets. (a1) DVCs of unlabelled test set with capacities between 1.98 and 2.02 Ah. (a2) DVCs of unlabelled test set with capacities between 2.02 and 2.06 Ah. (a3) DVCs of unlabelled test set with capacities between 2.06 and 2.10 Ah. (a4) DVCs of unlabelled test set with capacities between 2.10 and 2.14 Ah. (a5) DVCs of unlabelled test set with capacities between 2.14 and 2.18 Ah. (b1) DVCs of predicted abnormal cells, Abnormal Rate: 191/1000. (b2) DVCs of predicted abnormal cells, Abnormal Rate: 84/1000. (b3) DVCs of predicted abnormal cells, Abnormal Rate: 42/1000. (b4) DVCs of predicted abnormal cells, Abnormal Rate: 47/1000. (b5) DVCs of predicted abnormal cells, Abnormal Rate: 53/1000.

TABLE 4. Performance Comparison for Two Screening Methods.

Capacity Range (Ah)	Total numbber of cells	Traditional Screening Method		TTSCHR-CNN		Dropped rate
		Number of inconsistent cells	Inconsistent rate	Number of inconsistent cells	Inconsistent rate	
1.98 - 2.02	2829	193	6.82%	0	0	100%
2.02 - 2.06	25033	722	2.88%	22	0.0879%	96.95%
2.06 - 2.10	61177	1063	1.74%	54	0.0882%	94.93%
2.10 - 2.14	37331	634	1.70%	64	0.1714%	89.92%
2.14 - 2.18	4663	91	1.95%	24	0.5147%	73.6%

to HBase for sorting in the computing layer. Second, the discharge voltage time series of cells in HBase are preprocessed by the methods proposed in this paper, including data alignment, two-step time-series clustering and hybrid resampling methods, to obtain training sets for the CNN models of online cell screening. The promising results predicted by the CNN models are used to update the training database when each result has to be tested by an evaluation system, which is used by some algorithms to verify or experts to distinguish that the predicted results are reliable. Finally, the predicted defective cells are displayed on the Web, where the cells are represented by the red rectangles in the visualization layer in Figure 7.

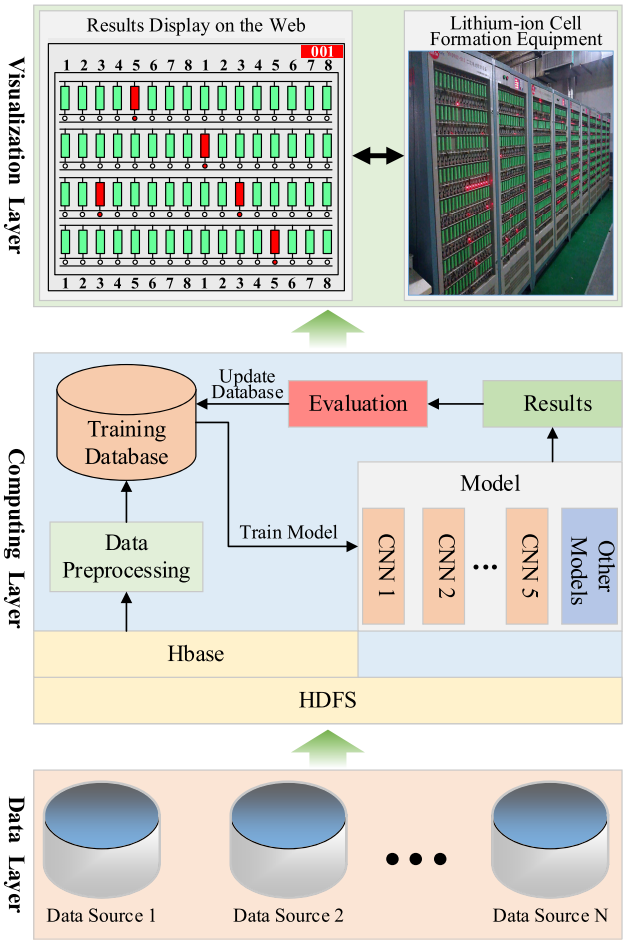

FIGURE 7. The application architecture of online lithium-ion cell screening in industrial production.

V. CONCLUSION

Screening lithium-ion cells that have similar electrochemical characteristics to configure battery packs can effectively alleviate the performance, lifespan and safety problems caused by the inconsistency among cells in a battery pack. The paper proposed the TTSCHR-CNN approach for lithium-ion cell screening.

The paper makes the following contributions:

- It screened cells with the static and dynamic characteristics of lithium-ion cells taken into consideration, thus ensuring that the screened cells have consistent electrochemical characteristics.
- It proposed the TTSCHR-CNN approach, which is an end-to-end "offline training and online screening" data-driven method for screening cells that have abnormal discharge voltage curves, providing a reference framework for time-series classification problems in similar industrial processes.
- It proposed a two-step time-series clustering method to label the raw no-label samples, providing labeled training samples for a supervised screening algorithm.
- It proposed a hybrid resampling method to solve the training sample imbalance, ensuring that the lithium-ion cell screening models based on CNN are effectively trained so as to enhance its screening precision.

We collected the data of 131096 cells manufactured by a certain lithium-ion cell manufacturer as experimental samples to verify the effectiveness of the approach proposed in the paper. The results show that the inconsistency rate of the screened cells by using the TTSCHR-CNN approach drops by 91.08%.

In the future, we attempt to use the more data generated in lithium-ion cell manufacturing processes to screen cells and ensure that the screened cells have consistent manufacturing processes and performance parameters. For the sample imbalance caused by the minority class sample scarcity in industrial processes, we will study the oversampling method that can effectively predict the minority class sample probability density space, thereby generating the more diverse and accurate minority class samples.

REFERENCES

[1] M. Mathew, Q. H. Kong, J. Mcgrory, and M. Fowler, "Simulation of lithium ion battery replacement in a battery pack for application in electric vehicles," *J. Power Sources*, vol. 349, pp. 94–104, May 2017.

[2] M. Al-Zareer, I. Dincer, and M. A. Rosen, "Novel thermal management system using boiling cooling for high-powered lithium-ion battery packs for hybrid electric vehicles," *J. Power Sources*, vol. 363, pp. 291–303, Sep. 2017.

[3] X. Zhou, L. Cheng, Y. Tang, Z. Pan, and Q. Sun, "Model identification of lithium-ion batteries in the portable power system," in *Proc. IEEE Int. Conf. Prognostics Health Manage.*, Ottawa, ON, Canada, Jun. 2016, pp. 1–5.

[4] P. Raspa *et al.*, "Selection of lithium cells for EV battery pack using self-organizing maps," *J. Automot. Safety Energy*, vol. 2, no. 2, pp. 157–164, Sep. 2011.

[5] J. Kim, J. Shin, C. Chun, and B. H. Cho, "Stable configuration of a Li-ion series battery pack based on a screening process for improved voltage/SOC balancing," *IEEE Trans. Power Electron.*, vol. 27, no. 1, pp. 411–424, Jan. 2012.

[6] M. Dubarry, N. Vuillaume, and B. Y. Liaw, "Origins and accommodation of cell variations in Li-ion battery pack modeling," *Int. J. Energy Res.*, vol. 34, no. 2, pp. 216–231, 2010.

[7] S. F. Schuster, M. J. Brand, P. Berg, M. Gleissenberger, and A. Jossen, "Lithium-ion cell-to-cell variation during battery electric vehicle operation," *J. Power Sources*, vol. 297, pp. 242–251, Nov. 2015.

[8] H. Dai, N. Wang, X. Wei, Z. Sun, and J. Wang, "A research review on the cell inconsistency of li-ion traction batteries in electric vehicles," *Automot. Eng.*, vol. 36, no. 2, pp. 181–188 and 203, Feb. 2014.

[9] S. J. Harris, D. J. Harris, and C. Li, "Failure statistics for commercial lithium ion batteries: A study of 24 pouch cells," *J. Power Sources*, vol. 342, pp. 589–597, Feb. 2017.

[10] Y. Zheng, M. Ouyang, L. Lu, and J. Li, "Understanding aging mechanisms in lithium-ion battery packs: From cell capacity loss to pack capacity evolution," *J. Power Sources*, vol. 278, pp. 287–295, Mar. 2015.

[11] S. Paul, C. Diegelmann, H. Kabza, and W. Tillmetz, "Analysis of ageing inhomogeneities in lithium-ion battery systems," *J. Power Sources*, vol. 239, no. 10, pp. 642–650, Oct. 2013.

[12] M. Ouyang, X. Feng, X. Han, L. Lu, Z. Li, and X. He, "A dynamic capacity degradation model and its applications considering varying load for a large format Li-ion battery," *Appl. Energy*, vol. 165, pp. 48–59, Mar. 2016.

[13] X. Han, M. Ouyang, L. Lu, J. Li, Y. Zheng, and Z. Li, "A comparative study of commercial lithium ion battery cycle life in electrical vehicle: Aging mechanism identification," *J. Power Sources*, vol. 251, no. 2, pp. 38–54, Apr. 2014.

[14] J. Kim and B. H. Cho, "Screening process-based modeling of the multi-cell battery string in series and parallel connections for high accuracy state-of-charge estimation," *Energy*, vol. 57, no. 8, pp. 581–599, Aug. 2013.

[15] S. Abada, G. Marlair, A. Lecocq, M. Petit, V. Sauvant-Moynot, and F. Huet, "Safety focused modeling of lithium-ion batteries: A review," *J. Power Sources*, vol. 306, pp. 178–192, Feb. 2016.

[16] X. Cao, Q.-C. Zhong, Y.-C. Qiao, and Z.-Q. Deng, "Multilayer modular balancing strategy for individual cells in a battery pack," *IEEE Trans. Energy Convers.*, vol. 33, no. 2, pp. 526–536, Jun. 2018.

[17] M. Kokila, P. Manimekalai, and V. Indragandhi, "Design and development of battery management system (BMS) using hybrid multilevel converter," *Int. J. Ambient Energy*, to be published, doi: 10.1080/01430750.2018.1492440.

[18] Q. Ouyang, J. Chen, J. Zheng, and H. Fang, "Optimal cell-to-cell balancing topology design for serially connected Lithium-Ion battery packs," *IEEE Trans. Sustain. Energy*, vol. 9, no. 1, pp. 350–360, Jan. 2018.

[19] N. Yang, X. Zhang, G. Li, A. Cai, and Y. Xu, "Effects of temperature differences among cells on the discharging characteristics of lithium-ion battery packs with series/parallel configurations during constant power discharge," *Energy Technol.*, vol. 6, no. 6, pp. 1067–1079, Jun. 2018.

[20] Y. H. Chen, W. U. Wei-Jing, H. W. Liu, and Z. Z. Zhang, "Study on sorting technology for lithium-ion power battery of electric vehicle," *J. Hunan Univ.*, vol. 43, no. 10, pp. 23–31, Oct. 2016.

[21] K. Fang, S. Chen, D. Mu, B. Wu, and F. Wu, "Investigation of nickel–metal hydride battery sorting based on charging thermal behavior," *J. Power Sources*, vol. 224, no. 224, pp. 120–124, Feb. 2013.

[22] K. Lee and D. Kum, "Development of cell selection framework for second-life cells with homogeneous properties," *Int. J. Elect. Power Energy Syst.*, vol. 105, pp. 429–439, Feb. 2019.

[23] J.-H. Kim, J.-W. Shin, C.-Y. Jeon, and B.-H. Cho, "Screening process of Li-Ion series battery pack for improved voltage/SOC balancing," in *Proc. IEEE Int. Conf. Power Electron.*, Sapporo, Japan, Jun. 2010, pp. 1174–1179.

[24] J. Kim, J. Shin, C. Jeon, and B. Cho, "High accuracy state-of-charge estimation of li-ion battery pack based on screening process," in *Proc. IEEE Appl. Power Electron. Conf. Expo.*, Mar. 2011, pp. 1984–1991.

[25] J. Zhang, J. Huang, L. Chen, and L. I. Zhe, "Lithium-ion battery discharge behaviors at low temperatures and cell-to-cell uniformity," *J. Automot. Safety Energy*, vol. 5, no. 4, pp. 391–400, Jun. 2014.

[26] X. He, G. Zhang, X. Feng, L. Wang, G. Tian, and M. Ouyang, "A facile consistency screening approach to select cells with better performance consistency for commercial 18650 lithium ion cells," *Int. J. Electrochem. Sci.*, vol. 12, no. 11, pp. 10239–10258, Dec. 2017.

[27] D. U. Chang-Qing, D. Luo, C. Zhang, D. Guo, and Y. H. Wang, "Study on screening method of lithium ion power battery," *Chin. J. Power Sources*, vol. 41, no. 7, pp. 997–980, Jun. 2017.

[28] J. Kim, "Cell seletion through two-level basis pattern recognition with low/high frequency components decomposed by DWT-based MRA," in *Proc. IEEE Energy Convers. Congr. Expo. (ECCE)*, Pittsburgh, PA, USA, Sep. 2014, pp. 906–911.

[29] J. Kim, "Discrete wavelet transform-based feature extraction of experimental voltage signal for Li-ion cell consistency," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1150–1161, Mar. 2016.

[30] J. Wu, C. Zhang, and Z. Chen, "An online method for lithium-ion battery remaining useful life estimation using importance sampling and neural networks," *Appl. Energy*, vol. 173, pp. 134–140, Jul. 2016.

[31] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering—A decade review," *Inf. Syst.*, vol. 53, pp. 16–38, Oct./Nov. 2015.

[32] X. Huang, Y. Ye, L. Xiong, R. Y. K. Lau, N. Jiang, and S. Wang, "Time series k-means: A new k-means type smooth subspace clustering for time series data," *Inf. Sci.*, vols. 367–368, pp. 1–13, Nov. 2016.

[33] K. Honkura and T. Horiba, "Study of the deterioration mechanism of $LiCoO_2$/graphite cells in charge/discharge cycles using the discharge curve analysis," *J. Power Sources*, vol. 264, no. 9, pp. 140–146, Oct. 2014.

[34] J. J. Liu, "Research on lithium battery charge and discharge model and influence of key parameters," *Chin. J. Power Sources*, vol. 41, no. 11, pp. 1546–1547 and 1574, Jun. 2017.

[35] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[36] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[37] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Syst. Appl.*, vol. 91, pp. 464–471, Jan. 2018.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[39] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.

[40] M. A. Tahir, J. Kittler, K. Mikolajczyk, and F. Yan, "A multiple expert approach to the class imbalance problem using inverse random under sampling," in *Proc. Int. Workshop Multiple Classifier Syst.*, 2009, pp. 82–91.

[41] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1578–1585.

[42] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2018.

[43] Z. Cui, W. Chen, and Y. Chen. (May 2016). "Multi-scale convolutional neural networks for time series classification." [Online]. Available: https://arxiv.org/abs/1603.06995

[44] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Time series classification using multi-channels deep convolutional neural networks," in *Web-Age Information Management*. Cham, Switzerland: Springer, 2014, pp. 298–310.

[45] F. Y. Zhou, L. P. Jin, and J. Dong, "Review of convolutional neural network," *Chin. J. Comput.*, vol. 40, no. 6, pp. 1229–1251, Jun. 2017.

[46] H. Cao, X.-L. Li, D. Y.-K. Woon, and S.-K. Ng, "Integrated oversampling for imbalanced time series classification," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 12, pp. 2809–2822, Dec. 2013.
[47] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement.*, Savannah, GA, USA, Nov. 2016, pp. 265–283.
[48] A. M. Taqi, A. Awad, F. Al-Azzo, and M. Milanova, "The impact of multi-optimizers and data augmentation on TensorFlow convolutional neural network performance," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Miami, FL, USA, Apr. 2018, pp. 140–145.
[49] R. Menon, "Cloudera's distribution including apache Hadoop," in *Cloudera Administration Handbook*. Birmingham, U.K.: Packt, 2014, pp. 57–86.
[50] S. Maneas and B. Schroeder, "The evolution of the Hadoop distributed file system," in *Proc. 32nd Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA)*, Krakow, Poland, May 2018, pp. 67–74.

CHENGBAO LIU received the B.S. degree in automation from the Xi'an University of Architecture and Technology, Xi'an, China, in 2012, and the M.S. degree in control theory and control engineering from the Automation Research and Design Institute of Metallurgical Industry, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Science. His main research interests include knowledge automation, intelligent manufacturing, industrious big data analysis, and time series analysis.

JIE TAN received the B.S. degree in automatic control from the Xi'an University of Technology, Xi'an, China, in 1986, and the M.S. degree in industrial automation from Central South University, Changsha, China, in 1992. He is currently a Professor with the Institute of Automation, Chinese Academy of Sciences. He has authored over 50 papers in international journals and conferences, and holds 60 patents for inventions. His research interests are in industrial intelligence technology and systems, artificial intelligence, big data analysis, deep learning, and intelligent factories. He has served as a reviewer for several journals and conferences, such as *Computer Integrated Manufacturing Systems*. He has hosted more than 30 scientific research projects, such as the National Natural Science Foundation of China, the National 863 Program, and the National Science and Technology Support Program.

HEYUAN SHI received the B.S. degree from the School of Information Science and Engineering, Central South University, Changsha, China, in 2015. He is currently pursuing the Ph.D. degree in software engineering with Tsinghua University, Beijing, China. His current research interests include embedded systems, Internet of Things, and operating systems.

XUELEI WANG received the M.S. degree in control theory and control engineering from the Dalian University of Technology, Dalian, China, in 1999, and the Ph.D. degree in control theory and control engineering from Shanghai Jiao Tong University, Shanghai, China, in 2002. He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. He has authored over 30 papers in international journals and conferences. His research interests are in industrial process modeling and optimization control, intelligent control, complex industrial process assisted analysis, and decision support.