Article

Cloud-Based Deep Learning for Co-Estimation of Battery State of Charge and State of Health

Dapai Shi [1,†]**, Jingyuan Zhao** [2,*,†]**, Zhenghong Wang** [1]**, Heng Zhao** [3]**, Chika Eze** [4]**, Junbin Wang** [5]**, Yubo Lian** [5] **and Andrew F. Burke** [2,*]

1   Hubei Longzhong Laboratory, Hubei University of Arts and Science, Xiangyang 441053, China
2   Institute of Transportation Studies, University of California-Davis, Davis, CA 95616, USA
3   College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China
4   Department of Mechanical Engineering, University of California, Merced, CA 94720, USA
5   BYD Automotive Engineering Research Institute, Shenzhen 518118, China
*   Correspondence: jyzhao@ucdavis.edu (J.Z.); afburke@ucdavis.edu (A.F.B.)
†   These authors contributed equally to this work.

Abstract: Rechargeable lithium-ion batteries are currently the most viable option for energy storage systems in electric vehicle (EV) applications due to their high specific energy, falling costs, and acceptable cycle life. However, accurately predicting the parameters of complex, nonlinear battery systems remains challenging, given diverse aging mechanisms, cell-to-cell variations, and dynamic operating conditions. The states and parameters of batteries are becoming increasingly important in ubiquitous application scenarios, yet our ability to predict cell performance under realistic conditions remains limited. To address the challenge of modelling and predicting the evolution of multiphysics and multiscale battery systems, this study proposes a cloud-based AI-enhanced framework. The framework aims to achieve practical success in the co-estimation of the state of charge (SOC) and state of health (SOH) during the system's operational lifetime. Self-supervised transformer neural networks offer new opportunities to learn representations of observational data with multiple levels of abstraction and attention mechanisms. Coupling the cloud-edge computing framework with the versatility of deep learning can leverage the predictive ability of exploiting long-range spatio-temporal dependencies across multiple scales.

Keywords: lithium-ion battery; state of charge; state of health; deep learning; cloud; field application

1. Introduction

With increased concerns about global warming, transportation electrification has recently emerged as an important step across the world. In electrified vehicles, rechargeable lithium-ion batteries are currently the most widely used systems for electrochemical energy storage and powering electric vehicles (EVs) due to their relatively high specific energy, acceptable cost and cycle life [1]. However, degradation and aging during the system's operational lifetime is still one of the most urgent and inevitable problems, especially under realistic conditions [2]. In field applications, such as an EV, an online battery management system (BMS) offers tools to monitor cell behavior under dynamic operating conditions. However, predicting real-life battery performance in field applications only using the online BMS is either difficult or impossible due to the limited data computing and storage ability of the onboard chips.

Over the past decade, scientists and researchers are increasingly storing and analyzing their big datasets by using remote 'cloud' computing servers [3]. On the cloud, researchers can interact with field data more flexibly and intelligently. Migrating observational data from custom servers to the cloud opened up a new world of opportunities to both assimilate the data sensibly and explore it in depth.

Several international companies have recognized this and have recently launched their cloud-based software, including Bosch [4], Panasonic [5] and Huawei [6]. Such public-cloud services are also termed software as a service (SaaS). The SaaS provided by Bosch—battery in the cloud—claimed that it is possible to improve the cycle life of batteries by 20% through the development of digital twins by using the big datasets from vehicle fleets. The universal battery management cloud (UBMC) service developed by Panasonic aims to identify the cell state and optimal battery operation. The SaaS launched by Huawei aims to provide a public cloud computing and storage service for EV companies. By learning from the historical battery data, the purely data-driven model embedded on its cloud monitoring system is applied to predict cell fault by discovering intricate structure in large EV-battery datasets. Beyond enterprise-level cloud services, a national-level big-data platform was built in 2017 in China, named the National Monitoring and Management Platform for New Energy Vehicles (NMMP-NEV) [7]. Up to now, the NMMP-NEV has provided remote fault diagnosis for more than six million EVs.

1.1. Literature Review

1.1.1. Modelling and Predicting Battery States

A battery is a sophisticated material system, with its functionality reliant on the transport of charge and ions through distinct phases and across interfaces, as well as both reversible and irreversible chemical reactions, among other material-dependent factors. The performance of a cell can be influenced by variations in components such as electrodes, electrolytes, interfaces, microstructures, current collectors, separators, binders, and cell or pack designs, as well as environmental factors and operating conditions. While significant progress has been made in first-principles, atomistic, and physics-based electrochemical modeling of battery systems, the absence of comprehensive predictive models remains a limiting factor for advancement. The battery management system (BMS) plays a pivotal role in maintaining the safe and reliable operation of battery systems for EV applications. Battery modelling is the core function of a BMS. Over the past few years, a variety of estimation techniques have been developed for the determination of the state of batteries in terms of two important parameters: SOC [8] and SOH [9]. In the literature, the most-studied methods in this regard for Li-ion batteries are equivalent circuit models (ECMs), physics-based models (PBMs), the observational filter model, and, more recently, data-driven, machine learning-based techniques. Each method has its own advantages and challenges. But there is always a trade-off between model accuracy and computational cost (Figure 1). For example, ECMs offer an effective tool to identity the cell states with low computational cost. Such a simple method has been widely used in onboard BMS for the last decade. However, it cannot provide accurate cell parameter values due to the simplification and assumptions in battery behaviors. Compared to ECMs, PBMs can approximate the physico-chemical processes that take place inside the cell during the system's operation, which provides accurate and physically consistent predictions. This method requires detailed information on cell specifications, including the materials and chemistry of the electrode, electrolyte, separator, current collectors, and so on. However, it is impossible to obtain the evaluation of these parameters during the operational lifetime under realistic conditions. In addition, battery problems in this case are governed by highly parameterized partial differential equations (PDEs). Solving the governing PDEs faces severe challenges and introduces multiple sources of uncertainty, especially in real-life physical problems with missing and noisy data and uncertain boundary conditions. Filter-based models are the most-commonly used methods for battery parameter estimation in the existing studies. Two issues constrain the wide application of filtering algorithms: (i) model parameters need to be updated and (ii) the algorithms may have poor generalization performance.
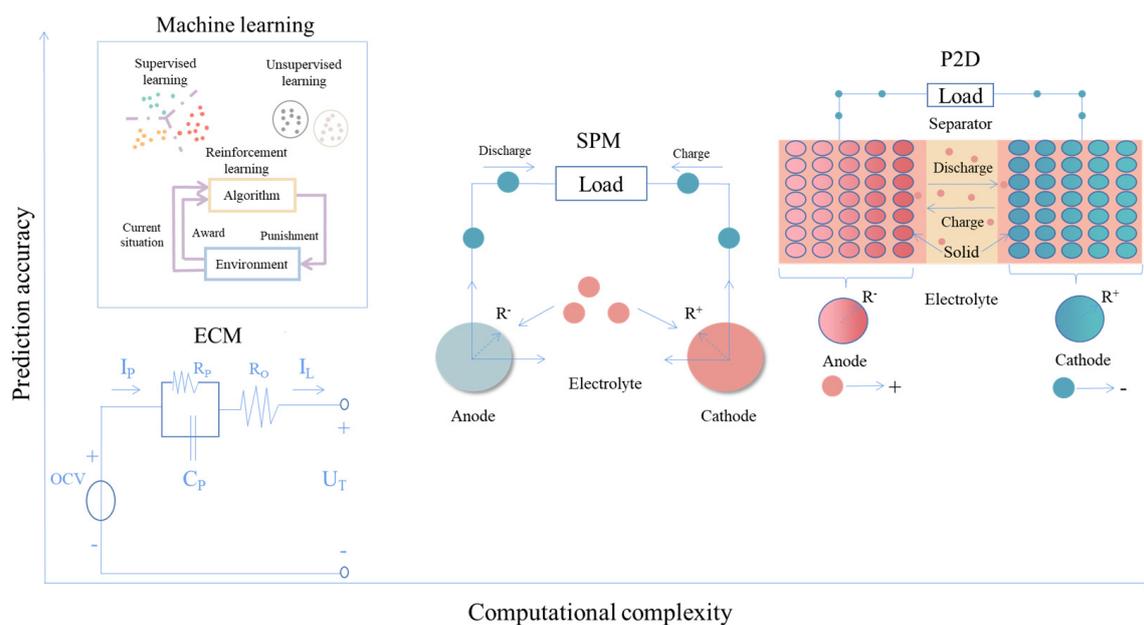
Figure 1. The trade-off between computational cost and model accuracy.

Conversely, the data-driven approach, especially machine-learning-based techniques, displays superior advantages in applications to materials and batteries, from the characterization of the material properties to the non-destructive evaluation of cell performance [10]. Machine learning allows computational models to discover intricate structure in the dataset and capture the statistics of the observational data [11].

The machine learning techniques used to predict the evolution of the battery can be classified into two main categories: traditional machine learning such as kernel-based approaches, and deep learning approaches such as deep neural networks. Conventional machine-learning techniques can be applied to process observational data in their raw form. The learning subsystem in widespread use in the machine learning community, deep or not, is supervised learning—that is, classification and regression. Such practical applications of machine learning use hand-engineered features or raw data for almost all recognition and predictive tasks. For example, extreme gradient boosting (XGBoost) was used to estimate the battery SOC of Li-ion batteries under dynamic loading conditions [12]. The XGboost technique offers a tool to improve the predictive performance by leveraging a set of weak learners and aggregating the outputs of each base model. The application of the multi-step forecasting strategy using XGBoost has been demonstrated to work in capturing, in real-time, the cell dynamics and predicting the terminal voltage and SOC under WLTP driving cycles. Moreover, gradient boosting also shows good performance in battery lifespan prediction [13]. Using the openly shared dataset provided by MIT/Toyota [14], a variety of features including voltage-related, capacity-related and temperature-related features were extracted and constructed for the gradient-boosting regression tree (GBRT) model. Gaussian process regression (GPR) is another common machine learning technique for battery SOC and SOH estimation. For example, battery SOC is identified by using electrochemical impedance spectroscopy (EIS) measurements based on GPR [15]. Through the feature selection from EIS data over frequencies of 1 mHz to 6 kHz, the GPR model can be trained to establish the mapping relationship between the selected features and the SOC under various temperatures. In another study, the GPR technique has been demonstrated as an effective tool to learn nonlinear battery systems and predict capacity fade in a variety of loading scenarios [16]. By introducing a Bayesian non-parametric transition, the model can incorporate estimates of uncertainty into predictions, allowing the determination of varying probabilities of the ranges of possible future health values across a long-term timescale.

Conventional machine learning offers a straightforward and effective tool for classification and regression tasks. However, constructing such a machine-learning system in general requires careful feature engineering and considerable domain-specific expertise to design a feature extractor that can transform the raw data (such as battery voltage, current, etc.) into suitable vector representations from which the learning algorithm could classify or predict patterns in the input.

In recent years, a new learning philosophy is the family of deep learning, which enables a machine to be fed with raw observations in mathematically useful latent spaces and to discover intricate structure in datasets automatically. One popular type of deep learning model is recurrent neural networks (RNNs) and their popular variants, including long-short term memory (LSTM) and gated recurrent units (GRU).

For example, a single hidden-layer GRU-RNN model was designed to estimate battery SOC by using the measured voltage and current [17]. In the proposed gradient method, the weight change direction takes a compromise of the gradient direction at current instant and at historical time to prevent the oscillation of the weight shift and to improve the training speed. Moreover, artificial noise was added to the observational data to improve the generalization and robustness of the neural networks. Recently, a hybrid neural network model was developed for SOC estimation of batteries at low temperatures by coupling a convolutional neural network (CNN) and GRU [18]. The CNN module was applied to learn the feature parameters of the inputs, while the bidirectional weighted GRU offers tools to improve the fitting performance of the network at low operating temperatures by tuning the weights.

In application to battery SOH estimation, a dynamic RNN model with good mapping ability was established for co-estimation of SOC and SOH for a lithium-ion battery [19]. The dynamic RNN model was suitable for estimating the nonlinear and dynamic cell behaviors. Meanwhile, self-adaptive weight particle swarm optimization was applied to improve the performance of the networks. Compared with the traditional gradient descent algorithm, particle swarm optimization offers an opportunity to improve the error convergence speed and avoid local optima. In a recent study, an encoder–decoder model based on the GRU was developed to be suitable for time series prediction of a Li-ion battery. The GRU-based encoder–decoder model has demonstrated its ability to predict the dynamic cell voltage response under complex current load profiles. In contrast to a conventional ECM model, the data-driven deep neural network does not require domain-specific knowledge and time-consuming tests under a well-controlled laboratorial environment.

Collectively, the results from these works demonstrate that RNNs and their variants are effective in modelling and predicting nonlinear battery systems [20]. However, they suffer from limitations due to the sequential processing and challenges related to back-propagation through time, particularly in the modelling of long-range connections across multiple timescales. These are manifested as training instabilities leading to vanishing and exploding back-propagated gradient problems [21]. The transformer model, primarily utilized for natural language processing, has recently achieved remarkable advancements in time series forecasting [22]. The transformer model allows for parallel processing, enabling efficient utilization of computing resources and faster training. Consequently, this methodology can be a promising option for battery state estimation. For example, one study proposed Dynaformer, a new deep learning architecture based on a transformer, which can predict the aging state and full voltage discharge curve for real batteries accurately, using only a limited number of voltage/current samples [23]. The study shows that the transformer-based model is effective for different current profiles and is robust to various degradation levels. Transformers tackle these obstacles by employing self-attention and positional encoding methods that simultaneously focus on and encode the order information while analyzing current data points within the sequence. These methods preserve the sequential information essential for learning while eliminating the traditional concept of recurrence. Transformers are capable of capturing such information through

the utilization of multiple attention heads. In addition, the model bridge the gap between simulations and real data, enabling accurate planning and control over missions.

1.1.2. Cloud-Based Progress on Battery State Prediction

Engineers are increasingly exploring the option of outsourcing computing needs to the cloud, which has become popular among researchers dealing with large amounts of data, due to its rapid improvements [24]. For example, one study proposed a cloud-based approach to estimate battery life by analyzing charging cloud data, which includes capacity and internal resistance estimations [25]. The capacity estimation relied on the ampere-hour integral method, which was further improved using temperature data and optimized with the Kalman filter (KF). To increase the precision of the estimation results, the study also implemented fuzzy logic (FL) to manage observation noise. Finally, the battery life was predicted using the Arrhenius empirical model. Experimental results demonstrated that the proposed method exhibited a low error rate of less than 4% in estimating battery life. In another study, the authors proposed a cloud-assisted online battery management method based on AI and edge computing technologies for EVs [26]. A cloud-edge battery management system (CEBMS) was established to integrate cloud computation and big data resources into real-time vehicle battery management. The proposed method utilized a deep-learning-algorithm-based cloud data mining and battery modeling method to estimate the battery's voltage and energy state with high accuracy. The effectiveness of the proposed method was verified by experimental tests, demonstrating its potential for more effective battery use and management in EVs.

Solving real-life physical problems in practical applications can be a daunting task, especially when dealing with multiple sources of uncertainty and imperfect data, including missing or noisy data, and outliers. This study focuses on the potential of using an AI-powered cloud-based framework to predict a nonlinear multiphysics and multiscale electrochemical systems' evolution in real-world applications. The cloud-based, closed-loop framework utilizes machine learning models that can learn seamlessly from field battery data in EV applications [27]. The concept of establishing digital twins for battery systems is an innovative approach to generating longitudinal electronic health records in cyberspace. By creating a digital replica of the battery system and continually training it on a stream of field data, it allows for continual lifelong learning. This approach can lead to significant benefits such as improved robustness, higher accuracy and faster training times. The continual training on a stream of field data enables the digital twin to adapt to changing conditions and learn from real-world experiences. This leads to improved accuracy in predicting the behavior of the battery system, which can help optimize its performance and extend its lifespan. In summary, the establishment of digital twins for battery systems offers a powerful approach to achieving continual lifelong learning, improving the robustness and accuracy of the system and reducing downtime.

1.2. Major Challenges Involved

Onboard BMS has long been an important component for EVs in the monitoring and controlling of battery systems. Despite relentless progress, solving real-life battery problems with noisy data and uncertain boundary conditions through traditional approaches remains challenging. Modelling and predicting multiscale and multi-physics battery problems for EV applications require further developments. Challenges specific to battery SOC and SOH estimation will further stimulate the development of new methodologies and frameworks, and we identify four major issues why collaboration between onboard BMS and cloud BMS (Figure 2) is of great importance in achieving the task, as follows.
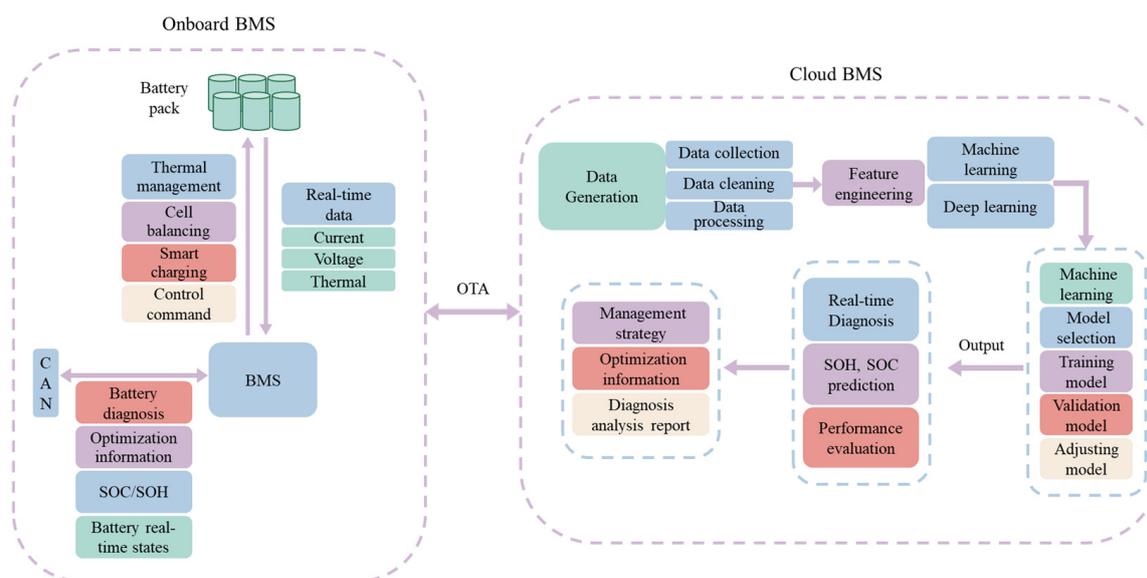
Figure 2. Synergy between onboard BMS and cloud BMS.

(1) Although data-driven machine learning techniques introduce considerable savings in computational cost compared with the traditional numerical methods (e.g., solving PDEs using finite elements), it still requires complex formulations and elaborate computer codes. Performing such tasks requires scheduling the training of computational algorithms in a more powerful computing environment. This is where cloud systems have come into play.

(2) Upon identifying cell conditions in real-world applications, there will be cell-to-cell, pack-to-pack and batch-to-batch variation, even with the most state-of-the-art manufacturing techniques. These cells would exhibit distinct states after long-term incubation. The specific approach to the predictive modelling of such battery systems significantly relies on the amount of data available and on the cell itself.

(3) In field applications such as EVs, the operation of the batteries depends not only on user driving patterns but also on environmental factors. Lab tests cannot incorporate diverse driving cycles and resting periods. Uncertainty arising from the randomness of high-dimensional parameter spaces make it difficult to perfectly match lab experiments to field applications.

(4) Last, but perhaps most important: even with open sharing of test data, reproducibility and generalization issues make it rather challenging to transfer academic progress to industry. However, the cloud-computing system provides a very flexible platform for analyzing, training and developing new frameworks and standardized benchmarks, which can be leveraged to improve our observational, empirical and physical understanding of real-life battery systems in a more intelligent manner.

1.3. Contributions of the Work

Despite relentless progress, predicting the dynamics of nonlinear battery systems by using traditional physical models inevitably faces severe challenges and introduces multiple sources of uncertainty. First-principle, phase-field, atomistic simulations may lead to insights into fundamental battery charging–discharging mechanisms, but they cannot truly predict cell performance for real-world applications [28]. We are well aware of the many benefits from cloud computing and storage. However, obviously, it is not enough to migrate data to a cloud platform; researchers need to be able to interact with the data more seamlessly and intelligently. Contributions of this study are as follows:

(a) Field data, which exhibits irregular loading conditions, dynamic operating scenarios, and path-dependent deterioration processes, is generated and uploaded to the cloud, reflecting real-world usage and making reliable predictions meaningful.

(b)　A specialized attention-based transformer neural network model is designed to learn parameters in the high-dimensional stochastic thermodynamic and kinetic battery system. The proposed transformer model has the advantage of strong generalization and robustness in the small data regime.

(c)　We examine the evolution of batteries using deep learning approaches in the time-resolved context and demonstrate how transformer neural networks, which automatically extract useful features, have the potential to overcome the limitations that have hindered the widespread adoption of data-driven machine learning-based techniques to date.

(d)　The designed cloud-based data-driven framework provides a highly flexible digital solution for a wide range of diverse physical, chemical and electrochemical problems in a way that produces promising results for the target outputs.

2. Key Components of Cloud BMS

Machine learning has emerged as a promising technique, but training an intelligent machine requires plentiful, high-quality and relevant training data. Onboard BMS cannot competently perform this task due to the high computational complexity of the data-driven models [29]. To make the best use of such a flexible technique, a cloud-based BMS provides complementary skills and opportunities to improve our understanding and evaluate comprehensive battery behaviors. Sensor data can be transmitted continuously to the cloud, where machine learning models can learn seamlessly from labeled samples while exploiting the wealth of information in the observations (Figure 3). With the advancement of sensor networks, it is now feasible to monitor the battery system across multiple spatial and temporal scales.
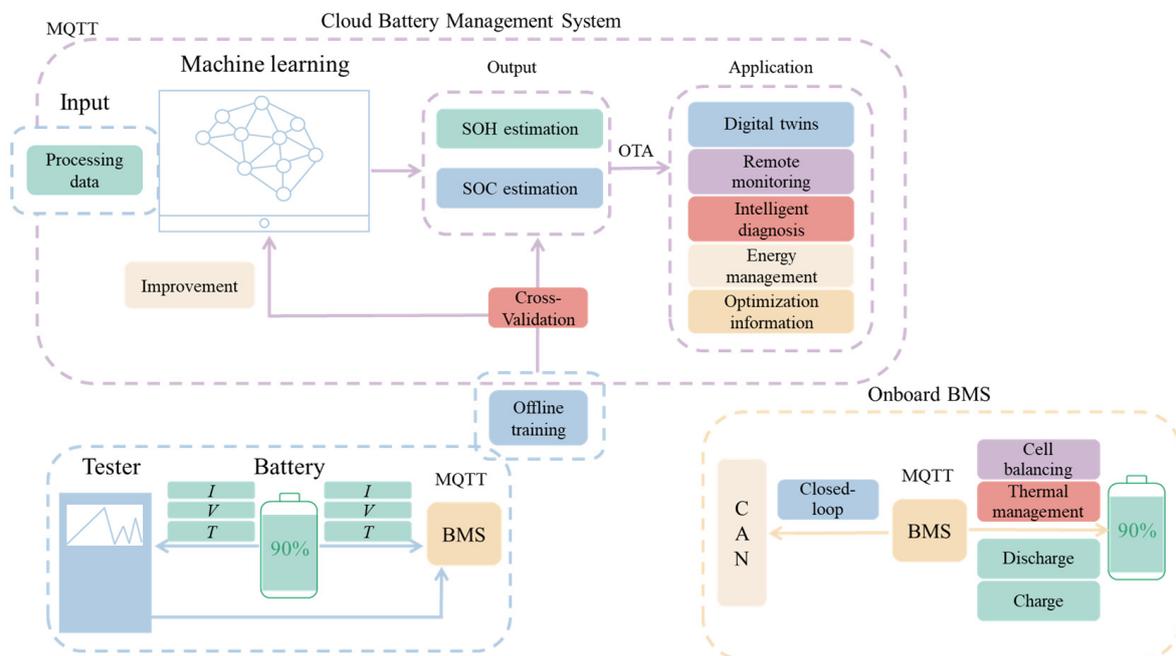
Figure 3. Cloud-based framework for commercial EV applications.

2.1. Physical Entity

In EV applications, lithium-ion batteries encounter complex operating conditions, including stochastic discharging processes for driving, dynamic charging processes for "refilling", and resting processes when parked. While most existing research in battery modeling has focused on either one cell or a specific, well-designed test, these efforts produce insights and improve our understanding of physical systems but cannot fully reflect the real-life situations with diverse aging mechanisms, significant cell-to-cell variability,

and complex loading scenarios. The huge gap between lab tests and practical applications makes it challenging to transfer academic progress to engineering. However, by assimilating real sensor measurements to optimize computational models, a digital twin can be used to replicate the behavior of a physical entity in silico. Focusing on analyzing time-resolved battery data such as voltage, current and temperature can directly contribute to meeting certain goals. Ultimately, what matters is the predictive ability under realistic conditions. These three fundamental parameters are the only information that we can obtain from an operating a battery using the onboard BMS.

2.2. IoT

The widespread use of the internet of things (IoT) in end-use devices such as EVs enables a wealth of multi-fidelity observations to be explored across several spatial and temporal scales [30]. There is a growing realization that terminal devices embedded with electronics and connected to networks play a crucial role in monitoring the evolution of complex digital and physical systems. With the prospect of trillions of sensors in the coming decade, it will be possible to seamlessly incorporate multi-fidelity data streams from real-world cases into physical models. In electric vehicle (EV) applications, battery performance, states and mechanical properties can vary greatly with dynamic loading conditions such as charging–discharging current rate, operating voltage window, frequency of usage and temperature. This calls for sophisticated and continuous monitoring throughout the operational lifetime.

Sensor measurements of battery cells can be transmitted to IoT components by the onboard BMS using the Controller Area Network (CAN) protocol. A special IoT protocol, message queuing telemetry transport (MQTT), allows for dual-direction messaging between the device and cloud and requires minimal resources. A large amount of sequential data are generated and collected from both private and fleet vehicles, which can be easily scaled to connect with millions of IoT devices. Data stored in onboard memory can be seamlessly uploaded to the cloud using TCP/IP protocols. The IoT wireless system in modern cities provides infrastructure for real-time data transmission using IoT actuators and onboard sensors.

2.3. Cloud

Cloud storage and computing have been demonstrated as powerful tools for remote monitoring and diagnosis. For automotive industry uses, researchers and engineers can configure their cloud environment and infrastructure to suit their requirements. The cloud-based BMS can seamlessly learn the stream of time-series battery data and produce electronic health records in the cloud. The most popular programming languages for cloud development include Java and Go. In addition, PHP offers a simple, effective, and flexible tool for web developers to create dynamic interfaces and interact with data deluge. The servers in these systems should have high-performance CPUs, plenty of RAM, and fast storage such as solid-state drives (SSDs). Additionally, the storage arrays should have high capacity, high performance and redundancy features such as RAID or replication to ensure data availability and durability. Backup and recovery systems are also critical for protecting customer data in case of disasters or system failures, and they should have high capacity and reliability. Cloud-based digital twins have demonstrated practical value in closed-loop full-lifespan battery management, including material design, cell performance evaluation and system optimization [31].

2.4. Modelling

Despite the progress made on the electrochemical modeling of battery systems using first-principle, atomistic or physics-based methods, the lack of canonical predictive models that can associate cell properties and mechanisms underlying their behavior with cell states has been a bottleneck for widespread adoption. Mathematical and computational tools have been developing rapidly, yet the multiscale and multiphysics battery system behavior

dominated by the spatial or temporal context underscores the need for a transformative approach. Machine learning technology is now a successful part of data-driven approaches, addressing a wide range of problems that have resisted the best efforts of the artificial intelligence (AI) community for many years [32]. The availability of shared data and open-source software, along with the ease of automation of materials tools, has brought machine learning into computational frameworks. Several software libraries, including TensorFlow, PyTorch and JAX, are contributing to the determination of cell performance by using various data modalities, such as time series, spectral data, lab tests, field data and more.

3. Methodologies

A learning algorithm that can seamlessly combine data and abstract mathematical operators plays a crucial role in discovering the representations needed for regression or classification. Deep learning techniques, in particular, naturally offer tools for extracting features and patterns from data automatically. To explore the observational data (which are uploaded to a private cloud system) that are characterized by multiple spatial and temporal coverages, a specialized self-attention transformer-based neural network model is designed in this study. Transformer-based deep learning (bidirectional encoder representations from transformers, known as BERT) has received a lot of attention since it was proposed in 2017 [33], particularly in natural language processing [34] and computer vision [35]. In comparison with recurrent neural networks (RNNs), transformer neural networks perform parallelization and solve the long-term dependencies problem and thus can process the observations much more quickly. Inspired by the successful operation, recently, various transformer-based models have been designed in the aspects of time-series prediction and analysis. The core idea of transformer networks is the self-attention mechanism, which belongs to a variant of the attention mechanism that can discover intricate structure in large time-series datasets and reduces dependence on the unimportant information across multiple timescales. In this study, we investigate the use of a transformer and design specialized network architectures that automatically satisfy the physical system for multivariate time series predictive tasks, as shown in Figure 4.
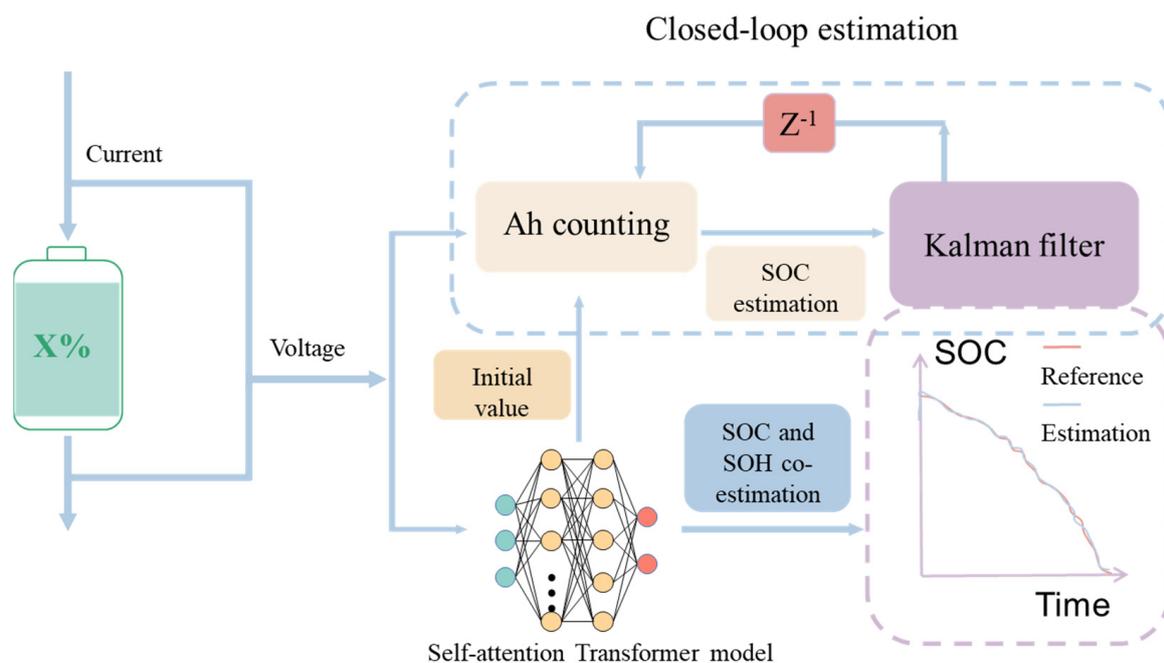
Figure 4. Specialized transformer architecture for the prediction of the battery system.

3.1. Transformer Neural Networks for Co-Estimation

In this study, the transformer deep learning model includes embedding, dual-encoder architecture and a gating mechanism. Time-series data are a collection of samples, observations and features recorded in a sequential manner over time. Self-supervised techniques are utilized to enhance transformers by providing them with the capability to learn, classify and forecast unlabeled data. The embedding vector replaces the original time-series data, i.e., the entire feature vector at a given time step. The specialized dual-encoder architecture is applied to battery prognostic and health management, which shows a higher test accuracy than a typical encoder–decoder architecture. A gating mechanism is used for coupling the predictive results of the two encoders. The success of a data-driven approach for predictive modeling of such real-life battery systems depends heavily on the amount of available data and the complexity of the model. Therefore, the model can be trained offline, and during the system's operational lifetime, online prediction only requires some sampled data points after data preprocessing (Figure 5).
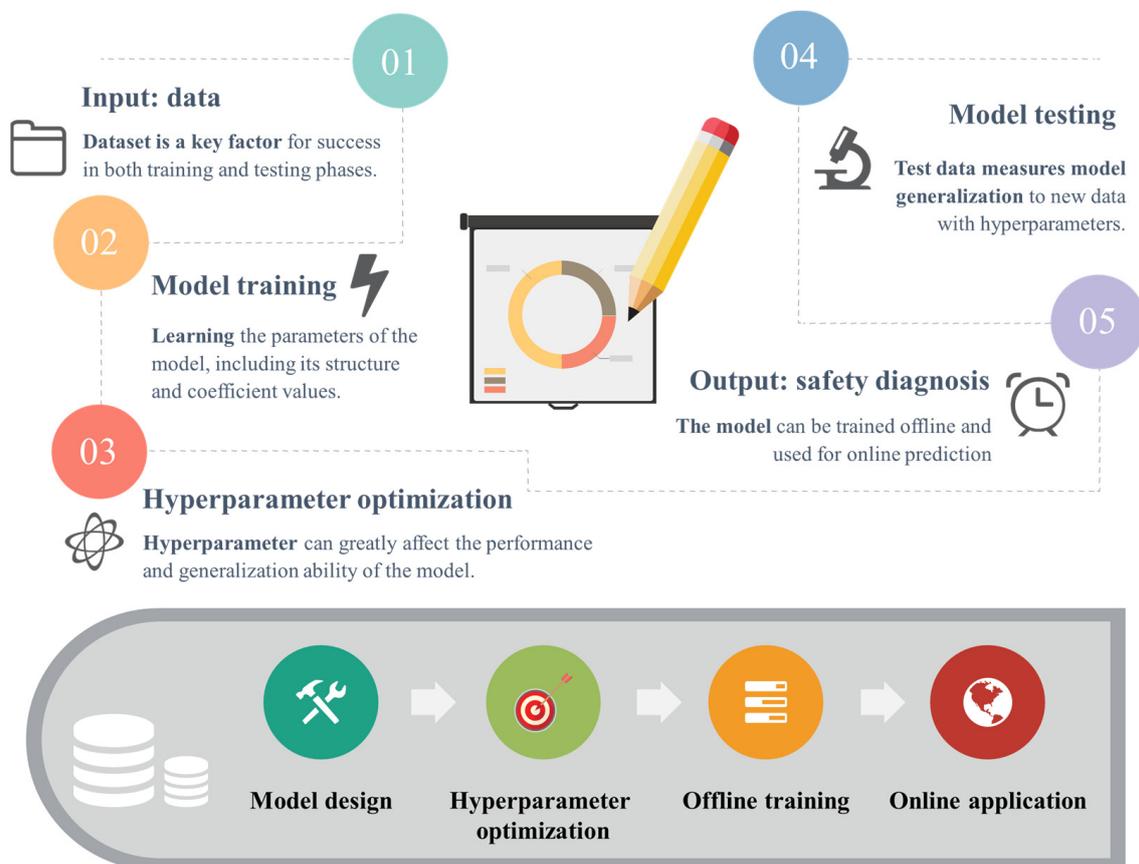
Figure 5. Architecture of the deep learning for co-estimation of battery states.

3.1.1. Data Processing

Data normalization has been demonstrated as a crucial step to a training process in a data-driven machine learning manner. A classical method of data normalization, named Z-score transformation, is used to normalize all the parameters of Li-ion batteries in the dataset into the vectors, characterized by the standard normal distribution with a mean of zero and variance of one. The Z-score method for each feature is calculated as:

$$z_{feature} = \frac{x_{obs.} - \mu}{\sigma} \tag{1}$$

where $x_{obs.}$ is the raw observational data, and μ and σ are the mean and standard deviation of the complete population, respectively.

The Transformer is an encoder-decoder structured sequence-to-sequence model, designed to accept a sequence of observational data as input. Transformers capture this information by employing multiple attention heads. To implement the transformer model efficiently, time-series data are separated into different segments based on the charging and discharging processes using an adaptive-length sliding window. Each time, the deep learning explores the long-range correlations across multiple timescales based on the sensor data inside the sliding window (Figure 6). It follows two rules: (a) the model outputs the SOC every 1 sample points (i.e., 10 s) when the direction of the current flow is constant. (b) the model re-starts the prediction process when the direction of the current flow changes. The sliding window offers a simple and effective tool to capture the structured relationships between the input and output under the charging and discharging processes.
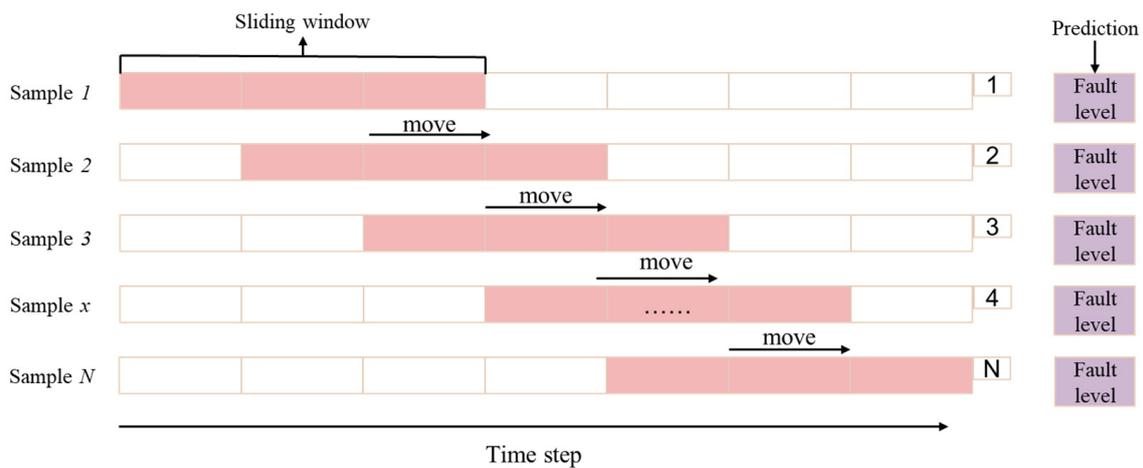
Figure 6. Sliding window mechanism of the transformer model.

3.1.2. Embedding

Unlike the LSTM or RNN models, The transformer model has no recurrence and no convolution. Instead, it models the sequence information using the encoding included in the input embeddings. The embedding of a typical BERT model includes token and position embeddings. By embedding time (seconds, minutes, hours, weeks, years, etc.) into the input, the model can effectively analyze time-series data while utilizing the computational advantages of modern hardware such as GPUs, TPUs and others. In some ways, our embedding strategy is analogous to BERT, but it has unique capabilities and merits for leveraging physical information. The token embedding of the original BERT is a discrete variable (word), while the observational data of our model is a time-series variable (cell parameters) with missing data and sensor noise. Moreover, fine-tuning ensures that the output embedding for each cell condition encodes contextual information that is more relevant to the multiscale and multiphysics battery system. The positional encoding applied to model the sequence information of the battery can be expressed as:

$$\text{PE}(t)_i = \left\{ \begin{array}{l} P_i = \sin(\dfrac{pos}{10,000^{\frac{2i}{d}}}) \\ P_{i+1} = \cos(\dfrac{pos}{10,000^{\frac{2i}{d}}}) \end{array} \right\} \tag{2}$$

where pos is the position in the time-step of the input, and i is the dimension of the embedding vector. It allows the learning algorithm to easily learn to attend by relative positions [33].

3.1.3. Dual Encoder

In a typical transformer model, the encoder block consists of multi-head self-attention modules and position-wise feedforward neural networks. To meet the needs of the battery system, a dual-encoder architecture is designed to produce predictions that respect the

physical invariants and principles. The transformer model has made breakthrough progress due to the self-attention mechanism, which offers an effective tool for the automatic extraction of abstract spatio-temporal features automatically. Such new paradigms of pretraining and fine-tuning enable large-scale scientific computations on long-range correlations across multiple timescales and thus enhance the generalization of neural network models. The multi-head attention mechanism allows the transformer model to extract information from different representation subspaces, which offers new opportunities for capturing the subtle differences between different battery cells within the pack.

In the self-attention module, multi-head self-attention sublayers simultaneously transform into query, key and value matrices. A sequence of vectors can be generated from the linear projections of the scaled dot-product attention:

$$\text{Attention}(q_h, k_h, v_h) = \text{softmax}\left(\frac{q_h k_h^T}{\sqrt{d_k}}\right) v_h \tag{3}$$

where $q_h \in R^{n*d_k}$, $k_h \in R^{m*d_k}$, $v_h \in R^{m*d_v}$ represents the query, key and value matrices, respectively; n and m denote the lengths of queries and keys/values, respectively; and d_k and d_v denote the dimensions of keys/queries and values, respectively. The multi-head attention mechanism with $h \in \{0, 1, \cdots, H\}$ different sets of learned projections can be expressed as:

$$\text{Multi-head\_Attention}(q_h, k_h, v_h) = \text{Concat}(\text{head}_1, \text{head}_h)\omega^o \tag{4}$$

where

$$\text{head}_h = \text{Attention}(q_h \omega_h^q, k_h \omega_h^k, v_h \omega_h^v) \tag{5}$$

3.2. Data Generation

Machine learning techniques encompass a collection of algorithms, techniques, normative structures and data that enable the derivation of a plausible model directly from observational data. The battery raw data under realistic conditions has been recorded and uploaded to a private cloud server, including cell voltage, current and temperature, which has been used to achieve a number of tasks for prognostics and health management, such as battery failure diagnosis [36] and battery SOH prediction [37]. In EV applications, multivariate time series represent the evolution of a group of variables: voltage, current and temperature over time. Table 1 lists the key cell specifications. The dataset is divided into training and test sets. The training set is utilized to learn the model for developing a base model and improve the accuracy and generalizability of predictions by fine-tuning the model under unseen battery charge-discharge protocols. The test set is used to quantitatively predict cell states: SOC and SOH.

Table 1. Cell chemistry and operating windows.

Parameter	Value
Cell type	Nickel Manganese Cobalt (NMC)
Nominal open circuit voltage	3.6 V
Nominal capacity	135 Ah
Operating voltage window	4.2 V to 2.5 V
Operating current density	Up to 1 C during charging vs. up to 6 C during discharging/driving
Operating temperature window	Less than 45 °C protected by thermal management

3.3. Evaluation Criteria

SOC and SOH are the two most important parameters in the prognostics and predictive health management, which are generally defined as:

$$\text{SOC} = \frac{Ah_{current}}{Ah_{full}} \times 100\% \tag{6}$$

$$\text{SOH} = \frac{Ah_{\text{full\_pre.}}}{Ah_{\text{full\_nom.}}} \times 100\% \tag{7}$$

where $Ah_{current}$ and Ah_{full} are the cell capacity in the present state and its full capacity, respectively, during the specific charging or discharging step, and $Ah_{\text{full\_pre.}}$ and $Ah_{\text{full\_nom.}}$ are the full capacity and nominal capacity, respectively.

The main output of the transformer model in this study is the prediction of SOC and SOH, which is compared with the observed values of the Li-ion cells. Three metrics are used to evaluate model performance, including root mean square error (RMSE), the mean absolute percentage error (MAPE) and the maximum absolute error (MAE). The inputs are the variables that follow a ground truth joint distribution. Specifically, RMSPE is defined as

$$\text{RMSPE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i{}^{*})^2} \tag{8}$$

where $\hat{y}_i$ and are $y_i{}^{*}$ the observed and predicted value of the *i*-th sample in the observational data.

MAPE can be expressed as

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i{}^{*}|}{y_i} \tag{9}$$

MAE can be given by:

$$\text{MAE} = \max_{1 \le i << n} |\hat{y}_i^{*} - y_i^{*}| \tag{10}$$

4. Performance of Cloud-Based BMS

4.1. SOC Estimation Results

The proposed transformer model was utilized to explore the intricate structures in battery time-series data and identify the representations necessary for predicting cell states. However, please note that solving real-life physical problems with missing, gappy or noisy boundary conditions requires pre-training of transformer models. In this study, tens of cells are randomly collected and used during their operational lifetime to pre-train the model initially. The observational data were fed into the transformer model, and its output was the SOC estimations corresponding to the sampling points (10 s sampling frequency using onboard sensor measurements). Due to the physico-chemical (thermodynamic and kinetic) principles, the model split the time-series data into several segments based on the charging and discharging processes. The model can thus discover intricate structures in two distinct operating conditions and then couple them together.

In the transformer model, multiple self-attention heads are introduced to operate on the same input in parallel. Each head uses distinct weight matrices to extract various levels of correlation between the input data. The transformer dual encoder offers predictive tools to extract vector representations of multivariate time-series, which can be considered as an autoregressive task of denoising the input [34]. The estimation results are independent in the charging and discharging conditions since the transformer model maps the input voltage and current sequences to SOC separately based on the direction of current flow. The initial SOC is calibrated at the time when the cell is fully charged or fully discharged during the system's operational lifetime, which depends on the usage behavior at an uncertain time. Once a precise value of the initial SOC is obtained, ampere-hour (Ah) counting

can be directly introduced to provide the ground truth for those observations. Therefore, the model estimates the SOC of the cell from voltage, current and temperature data by coupling the transformer model and the Ah counting method. The transformer-based model is initially trained for Cell_1, and the SOC estimation result is shown in Figure 7. The data-driven model achieves a MAPE of 0.76% and an RMSPE of 0.68%, with a maximum absolute error of less than 2%.
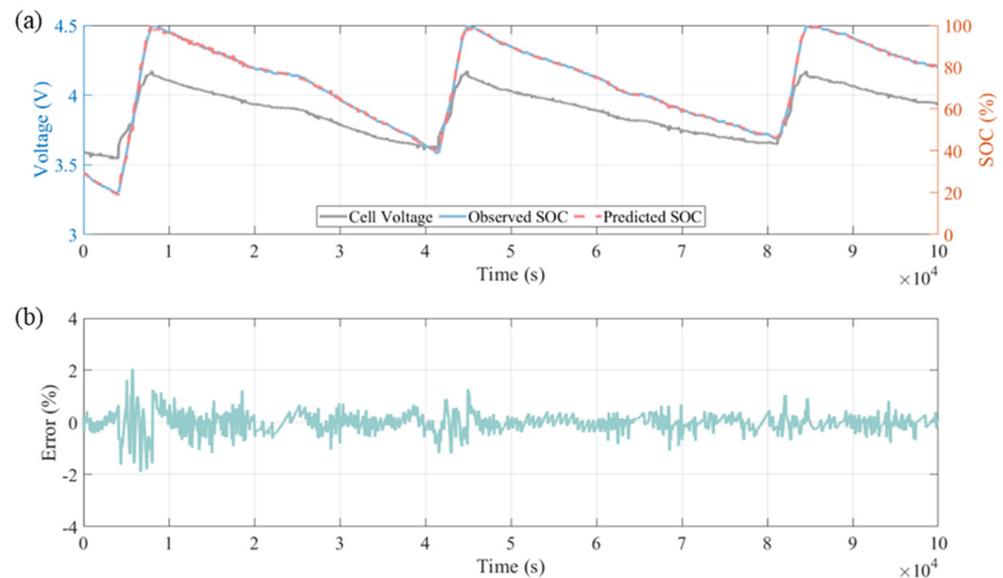
Figure 7. (**a**) SOC estimation results for Cell_1. (**b**) the prediction errors.

Subsequently, the self-attention transformer model is calibrated using another Cell_2, under totally different dynamic operating conditions. Regular calibration and maintenance of machine learning models require significant resources, including specialized personnel and technicians. Ensuring that the system remains accurate and effective necessitates constant attention and care, making it a crucial component of successful machine learning implementation. A good calibration process can be expressed as

$$\mathrm{P}(\hat{y} = y | \hat{p} = p) = p \tag{11}$$

where the probability $p \in [0, 1]$ is over the joint distribution, and $\hat{y}$ and $\hat{p}$ are the predictions and the associated confidence (probability of correctness). Let NN be a neural network, and thus it can be given by $NN(x) = (\hat{y}, \hat{p})$. As shown in Figure 8, the developed model can accurately estimate the SOC for the NMC battery (Cell_2) over both the charging and discharging processes with a MAE of less than 2.5%, a MAPE of 0.96% and an RMSPE of 0.81%. The proposed transformer approaches, in particular, provide reliable SOC estimations during the plateau in charge–discharge profiles. While accurate SOC estimation through machine learning modeling is possible, it should also focus on accounting for SOC errors induced by aging, temperature and hysteresis. Despite these factors, data-driven estimation remains a reliable SOC reference for other methods.
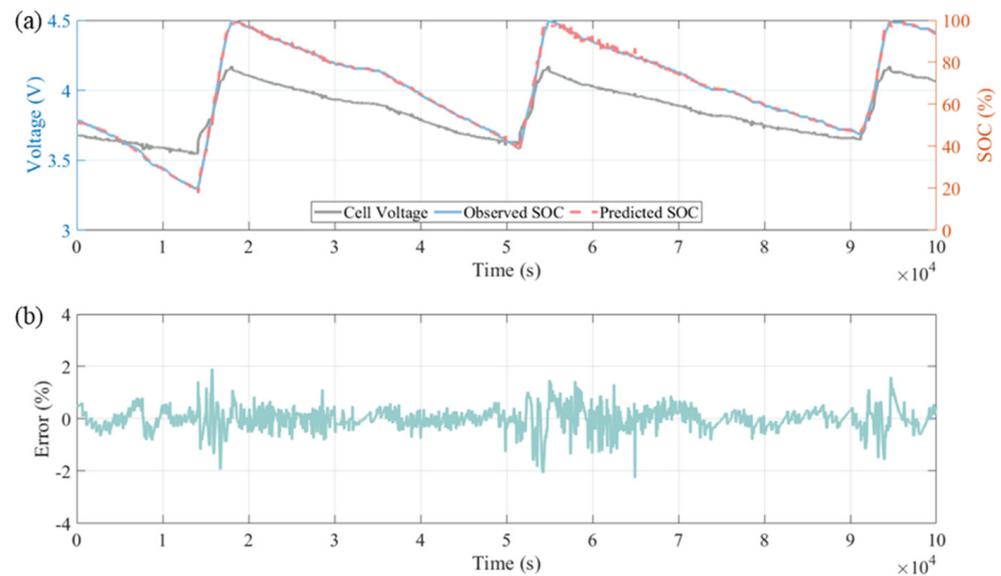
Figure 8. (**a**) SOC estimation results for Cell_2. (**b**) the prediction errors.

4.2. SOH Estimation Results

The definition, based on the Equation (12), is used to calculate the SOH in its current state. The estimation of the capacity can be expressed as:

$$\hat{C}_{mn} = transformer([x_{mn,t}, x_{mn,t+1}, \cdots, x_{mn,t+1+i}] | C_{mj}) \tag{12}$$

where $\hat{C}_{mn}$ is the estimated capacity of the m-th cell in the n-th cycle, $x_{mn,t}$ is the observational data in the n-th cycle, and C_{mj} is the observed value used as the ground truth. In field applications, the ground truth of the capacity cannot be obtained for every cycle. Therefore, $j << n$. The methodologies of the transformer model need to be revisited. For a complete explanation of the algorithm, refer to [22].

Herein, the loss value used to determine the hyperparameters of the self-attention transformer model can be given by:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i^*)^2 \tag{13}$$

Setting the hyperparameters for a transformer model can be a challenging task, which depends highly on the specific case, including the size and complexity of the training data and the available hardware. The model processes the encoder block's outputs for input into the linear layers. However, concatenation alone may yield poor prediction accuracy. Thus, a dense interpolation algorithm [38] with tunable hyperparameters is adopted to enhance performance. The validity of the trained transformer is demonstrated through the interpolated results in the time-space domain. Despite a decrease in accuracy with increasing feature differences between the test and training data, the proposed method still produces reasonable interpolation results. The trained transformer is then employed to reconstruct dense data with halved trace intervals for the field data. The reconstructed dense data exhibit greater spatial continuity, and the spatial aliasing effects disappear in the time domain. These reconstructed dense data hold the potential to enhance the accuracy of subsequent seismic data processing and inversion.

Hyperparameters are inherent in every machine learning system, and the fundamental objective of automated machine learning (AutoML) is to optimize performance by automatically setting these hyperparameters. Table 2 summarizes the hyperparameters used in this study.

Table 2. Hyperparameters of the transformer model.

Hyperparameter	Value/Method
Layers	10 layers with 8 heads per layer
Training	65 k gradient updates, 4096 sequence length, 64 batch size (262,144 tokens)
Learning Rate	Maximum of 1×10^{-3}, linear warmup of 500 steps
Dropout	0.2 rate
Optimizer	Adam with starting learning rate of 2
Implementation	Use TensorFlow and PyTorch for efficient implementation
Layer Normalization	Normalize input to each layer of the transformer
Weight Tying	Tie decoder and output layer weights
Label Smoothing	Apply label smoothing to target labels
Early Stopping	Stop training when validation loss stops improving

Figures 9 and 10 illustrates the performance of the transformer model in estimating the SOH of cell_1 and cell_2, which are used to train and test the model, respectively. In each group, we randomly selected 50 sampling points, and the results show that the proposed transformer-based model can achieve high predictive accuracy, with MAPE varying between ±2.5% within a 98% confidence interval during the system's operational lifetime. However, the model still needs further development in some field applications. Firstly, more efforts are required, such as hand-engineered feature extraction and establishing ground truth, to label the observational data for training. Secondly, more calibration work is needed to enhance the model's performance in short-length cycles, such as charging from 50% to 80% SOC. As the model lacks sufficient information to learn from, it may fail to provide accurate and physically consistent predictions for each field charging process, especially under random usage behaviors (e.g., charging for only a few minutes). This can lead to extrapolation or observational biases, which can negatively impact the model's performance.

(**a**)   (**b**)
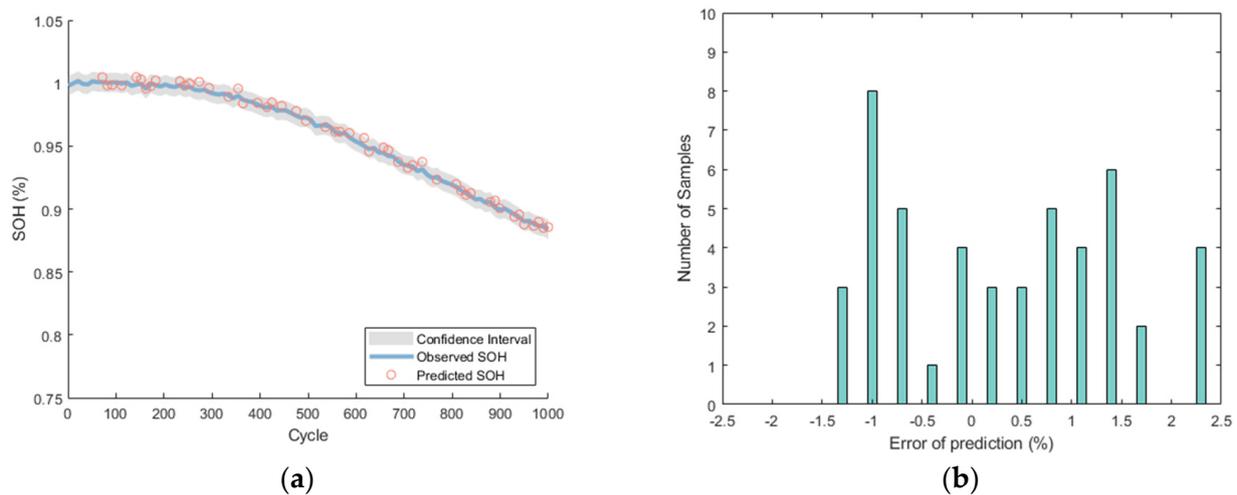
Figure 9. The training set. (**a**) SOH estimation for Cell_1. (**b**) the prediction errors.
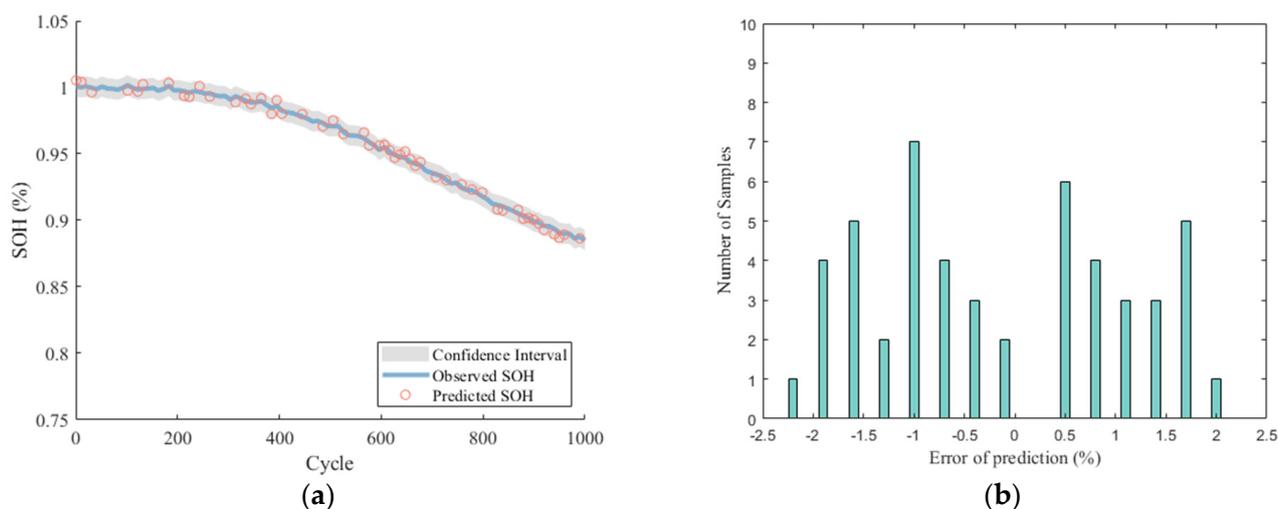
Figure 10. The test set. (**a**) SOH estimation for Cell_2. (**b**) the prediction errors.

5. Outlook

Developing cloud systems for battery and EV applications can pose challenges for many practitioners, but a user-friendly and accessible cloud development environment could help address some key issues. Observational data can be sparse and noisy, and may comprise vastly heterogeneous data modalities such as images, time series, lab tests, historical data, records, and more. The data for certain quantities of interest might not be readily available. To enhance the efficiency and accuracy of these systems, we propose five major recommendations:

(i) There is a significant opportunity for synergy between onboard-BMS and cloud-BMS technologies. Urgent and real-time tasks should be allocated to onboard BMS, while complex tasks that involve multiple scales and temporal dependencies should be distributed to cloud BMS.

(ii) Machine-learning models rely heavily on observational data, and new algorithms and mathematics are needed to yield accurate and robust methods that can handle high signal-to-noise ratios and outliers. These methods should also be able to generalize well beyond the training data. However, the model requires craftsmanship and elaborate implementations on different cell chemistries.

(iii) Battery behavior in EV applications is much more complex than in lab tests due to unprecedented spatial and temporal coverage. Working with noisy data and limited training sets and dealing with under-constrained battery problems with uncertain boundary conditions are major challenges that need to be addressed.

(iv) Developing deep learning architectures for modeling multiscale and multiphysics battery systems is currently done empirically, which is time-consuming. Training and optimizing deep neural networks can also be expensive. Emerging meta-learning techniques and transfer learning may offer promising directions to explore.

(v) Battery performance fluctuates unpredictably throughout its operational life. Precise forecasting and modeling of long-range spatio-temporal dependencies across cell, pack, and system levels are essential for efficient learning algorithms. A promising approach might involve hybrid modeling, combining physical process models with configurable, structured data-driven machine learning.

6. Conclusions

Field data have the potential to enhance the effectiveness of computational techniques developed for cloud-based battery management systems (BMS). In this study, we propose a cloud-based data-driven technique that utilizes state-of-the-art computational methods, specifically transformer neural networks, to accurately model cell behaviors for real-life electric vehicle (EV) applications. Our prediction model automatically extracts spatio-temporal

features using an attention-based deep learning approach, without relying on data from experimental test cycles or prior knowledge of cell chemistry and degradation mechanisms. By combining IoT devices to generate field data and machine-learning modeling on the cloud, our work underscores the potential for understanding and forecasting complex physical systems such as lithium-ion batteries. Overall, modeling and estimation using cloud-based BMS can complement other approaches based on simplified battery models (such as equivalent circuit models), physical and semi-empirical models, and specialized diagnostics embedded in the onboard BMS.

Author Contributions: Methodology, Supervision, A.F.B.; Software, J.Z. and H.Z.; Resources, Project administration and Funding acquisition, Y.L.; data curation, J.W.; writing—original draft, D.S. and J.Z.; writing—review & editing, C.E.; visualization, J.Z. and Z.W. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The data could not be shared due to confidentiality.

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

References

1. Li, M.; Lu, J.; Chen, Z.; Amine, K. 30 years of lithium-ion batteries. *Adv. Mater.* **2018**, *30*, 1800561. [CrossRef] [PubMed]
2. Sulzer, V.; Mohtat, P.; Aitio, A.; Lee, S.; Yeh, Y.T.; Steinbacher, F.; Khan, M.U.; Lee, J.W.; Siegel, J.B.; Howey, D.A. The challenge and opportunity of battery lifetime prediction from field data. *Joule* **2021**, *5*, 1934–1955. [CrossRef]
3. Gibney, E. Europe sets its sights on the cloud: Three large labs hope to create a giant public–private computing network. *Nature* **2015**, *523*, 136–138. [CrossRef] [PubMed]
4. Bosch Mobility Solutions: Battery in the Cloud. Available online: https://www.bosch-mobility-solutions.com/en/solutions/software-and-services/battery-in-the-cloud/battery-in-the-cloud/ (accessed on 19 August 2022).
5. Panasonic Announces UBMC Service: A Cloud-Based Battery Management Service to Ascertain Battery State in Electric Mobility Vehicles. Available online: https://news.panasonic.com/global/press/data/2020/12/en201210-1/en201210-1.pdf (accessed on 19 August 2022).
6. HUAWEI: CloudLi. Available online: https://carrier.huawei.com/en/products/digital-power/telecom-energy/Central-Office-Power (accessed on 19 August 2022).
7. National Monitoring and Management Platform for NEVs. Available online: http://www.bitev.org.cn/a/48.html (accessed on 19 August 2022).
8. Zheng, Y.; Ouyang, M.; Han, X.; Lu, L.; Li, J. Investigating the error sources of the online state of charge estimation methods for lithium-ion batteries in electric vehicles. *J. Power Sources* **2018**, *377*, 161–188. [CrossRef]
9. Hu, X.; Xu, L.; Lin, X.; Pecht, M. Battery lifetime prognostics. *Joule* **2020**, *4*, 310–346. [CrossRef]
10. Berecibar, M. Machine-learning techniques used to accurately predict battery life. *Nature* **2019**, *568*, 325–326. [CrossRef] [PubMed]
11. Ng, M.F.; Zhao, J.; Yan, Q.; Conduit, G.J.; Seh, Z.W. Predicting the state of charge and health of batteries using data-driven machine learning. *Nat. Mach. Intell.* **2020**, *2*, 161–170. [CrossRef]
12. Dineva, A.; Csomós, B.; Sz, S.K.; Vajda, I. Investigation of the performance of direct forecasting strategy using machine learning in State-of-Charge prediction of Li-ion batteries exposed to dynamic loads. *J. Energy Storage* **2021**, *36*, 102351. [CrossRef]
13. Yang, F.; Wang, D.; Xu, F.; Huang, Z.; Tsui, K.L. Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model. *J. Power Sources* **2020**, *476*, 228654. [CrossRef]
14. Severson, K.A.; Attia, P.M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M.H.; Aykol, M.; Herring, P.K.; Braatz, R.D. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **2019**, *4*, 383–391. [CrossRef]
15. Babaeiyazdi, I.; Rezaei-Zare, A.; Shokrzadeh, S. State of charge prediction of EV Li-ion batteries using EIS: A machine learning approach. *Energy* **2021**, *223*, 120116. [CrossRef]

16. Richardson, R.R.; Osborne, M.A.; Howey, D.A. Battery health prediction under generalized conditions using a Gaussian process transition model. *J. Energy Storage* **2019**, *23*, 320–328. [CrossRef]

17. Jiao, M.; Wang, D.; Qiu, J. A GRU-RNN based momentum optimized algorithm for SOC estimation. *J. Power Sources.* **2020**, *459*, 228051. [CrossRef]

18. Cui, Z.; Kang, L.; Li, L.; Wang, L.; Wang, K. A hybrid neural network model with improved input for state of charge estimation of lithium-ion battery at low temperatures. *Renew. Energy* **2022**, *198*, 1328–1340. [CrossRef]

19. Che, Y.; Liu, Y.; Cheng, Z. SOC and SOH identification method of li-ion battery based on SWPSO-DRNN. *IEEE J. Emerg. Sel. Top. Power Electron.* **2020**, *9*, 4050–4061. [CrossRef]

20. Schmitt, J.; Horstkötter, I.; Bäker, B. Electrical lithium-ion battery models based on recurrent neural networks: A holistic approach. *J. Energy Storage* **2023**, *58*, 106461. [CrossRef]

21. Sherstinsky, A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]

22. Ahmed, S.; Nielsen, I.E.; Tripathi, A.; Siddiqui, S.; Rasool, G.; Ramachandran, R.P. Transformers in time-series analysis: A tutorial. *arXiv* **2022**, arXiv:2205.01138.

23. Biggio, L.; Bendinelli, T.; Kulkarni, C.; Fink, O. Dynaformer: A Deep Learning Model for Ageing-aware Battery Discharge Prediction. *arXiv* **2022**, arXiv:2206.02555.

24. Drake, N. Cloud computing beckons scientists. *Nature* **2014**, *509*, 543–544. [CrossRef]

25. Li, K.; Zhou, P.; Lu, Y.; Han, X.; Li, X.; Zheng, Y. Battery life estimation based on cloud data for electric vehicles. *J. Power Sources* **2020**, *468*, 228192. [CrossRef]

26. Li, S.; He, H.; Wei, Z.; Zhao, P. Edge computing for vehicle battery management: Cloud-based online state estimation. *J. Energy Storage* **2022**, *55*, 105502. [CrossRef]

27. Zhao, J.; Nan, J.; Wang, J.; Ling, H.; Lian, Y.; Burke, A. Battery Diagnosis: A Lifelong Learning Framework for Electric Vehicles. In Proceedings of the 2022 IEEE Vehicle Power and Propulsion Conference (VPPC), Merced, CA, USA, 1–4 November 2022; IEEE: New York, NY, USA, 2022; pp. 1–6.

28. Aykol, M.; Herring, P.; Anapolsky, A. Machine learning for continuous innovation in battery technologies. *Nat. Rev. Mater.* **2020**, *5*, 725–727. [CrossRef]

29. Zhao, J.; Burke, A.F. Electric Vehicle Batteries: Status and Perspectives of Data-Driven Diagnosis and Prognosis. *Batteries* **2022**, *8*, 142. [CrossRef]

30. Mohammadi, F.; Rashidzadeh, R. An overview of IoT-enabled monitoring and control systems for electric vehicles. *IEEE Instru. Meas Mag.* **2021**, *24*, 91–97. [CrossRef]

31. Yang, S.; He, R.; Zhang, Z.; Cao, Y.; Gao, X.; Liu, X. CHAIN: Cyber hierarchy and interactional network enabling digital solution for battery full-lifespan management. *Matter* **2020**, *3*, 27–41. [CrossRef]

32. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]

33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Llion, J.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *Adv. Neural. Inf. Process Syst.* **2017**, *30*, 3–5.

34. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Rush, A.M. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 1 October 2020; EMNLP: New York, NY, USA, 2020; pp. 38–45.

35. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 10012–10022.

36. Zhao, J.; Ling, H.; Wang, J.; Burke, A.F.; Lian, Y. Data-driven prediction of battery failure for electric vehicles. *Iscience* **2022**, *25*, 104172. [CrossRef]

37. Zhao, J.; Ling, H.; Liu, J.; Wang, J.; Burke, A.F.; Lian, Y. Machine learning for predicting battery capacity for electric vehicles. *eTransportation* **2023**, *15*, 100214. [CrossRef]

38. Wang, B.; Zhang, N.; Lu, W.; Wang, J. Deep-learning-based seismic data interpolation: A preliminary result. *Geophysics* **2019**, *84*, 11–20. [CrossRef]