Energy Proceedings
Vol 45, 2024

ISSN 2004-2965

Estimation of the State of Charge Health of Electric Vehicle Batteries Through Machine Learning Approach

Capasso Clemente[1*], Chianese Giovanni[1], Veneri Ottorino[1], Patalano Stanislao[2], Vitolo Ferdinando[2]

1 National Research Council of Italy Institute of Sciences and Technologies for Sustainable Energy and Mobility

2 University of Naples Federico II, Department of Industrial Engineering
(*Corresponding Author: clemente.capasso@stems.cnr.it)

ABSTRACT

Ensuring high performances and lifetime of battery packs has critical importance, because of the transition toward electric mobility. Therefore, correct estimation of the battery state with ad-hoc designed Battery Management Systems (BMS) is pivotal to address this challenge. In this context, application of Machine Learning (ML) is gaining increasing research interest as it includes data-driven algorithms that enable accurate and fast predictions of the battery state. For this reason, this paper aims to contribute with: (i) a survey of the newest contributions to the prediction of the State of Charge/Health (SoC/SoH), and (ii) by schematizing a methodology that uses simulated data to train state-of-the-art types of neural networks (NNs) for SoC and SoH estimation of a LiNMC battery cell. Research papers considered in this review included applications of deep NN, and other ML algorithms. The impact of the training dataset on the performances of the ML models and their capability to generalize is remarked throughout the paper. For this reason, a validated electro-thermal model is used to generate data that accounts for different temperatures and current loads to simulate scenarios with different environmental conditions and driving cycles.

Keywords: Battery Management System (BMS), Machine Learning (ML), Neural Network (NN), State of Charge (SoC) estimation, State of Health (SoH) estimation.

NONMENCLATURE

Abbreviations

BMS	Battery Management System
CNN	Convolutional Neural Network
FCNN	Fully Connected Neural Network
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
MSE	Mean Square Error
NN	Neural Network
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RVM	Relevance Vector Machine
SoC	State of Charge
SoH	State of Health
SVR	Support Vector Regression

1. INTRODUCTION

The transition from fossil fuels to electric mobility is an important step toward reduction of the air pollution and climate change and is resulting in the continuous growth of the electric car share in the automotive market. For this reason, the sustainable lifecycle of battery packs for electric vehicles, which involves production, operation, and end-of-life, has fundamental importance for players involved in this field.

Battery packs have a modular structure, with cells grouped into modules and modules interconnected within the pack [1]. The design of a suitable BMS is essential to maintain optimum operating conditions at all these three levels. Indeed, BMSs enable cells monitoring and balancing, fault diagnosis, thermal management and estimation of charge level and health state [2]. Among these, SoC and SoH have a significant impact on the battery cell state. SoC and SoH are generally difficult to directly measure; therefore, the implementation of models and methods for their estimation is a topic of great interest in both the research and industrial communities [3].

# This is a paper for 15th International Conference on Applied Energy (ICAE2023), Dec. 3-7, 2023, Doha, Qatar.

SoC is defined as the residual capacity that can be supplied by the battery cell at its current operating state. It can be estimated with look-up tables, ampere-hour integration, model-based methods, and data-driven methods. The application of look-up tables is practical and immediate; however, they are characterized by low accuracy and robustness. The reliability of ampere-hour counting can be affected by measurement errors, accumulated during the time, due to sensors' accuracy. Model-based estimation can leverage equivalent circuit models and electrochemical models that can be coupled with adaptive filters, such as Kalman filters. In these cases, although the predictions are accurate, a high computational effort needs to be considered because of the iterative nature of these methods.

SoH measures the degradation of a battery cell, which is typically evaluated in terms of the actual capacity or of the internal resistance, according to equations (1) and (2):

$$\mathrm{SoH_C} = \frac{\mathrm{C_t}}{\mathrm{C_0}} \qquad (1)$$

$$\mathrm{SoH_R} = \left|\frac{\mathrm{R_f - R_t}}{\mathrm{R_f - R_0}}\right| \qquad (2)$$

where, $\mathrm{SoH_C}$ and $\mathrm{SoH_R}$ indicate the SoH in terms of capacity and resistance respectively, and the subscripts t, f, and 0 indicate the value of the capacity or resistance at the current time (t), at the end-of-life (f), and when the battery is new (0). Evaluation of both $\mathrm{SoH_C}$ and $\mathrm{SoH_R}$ can be performed with direct measurement or with indirect estimation. A direct measurement can be easily carried out in laboratory conditions but not onboard during operation, whereas, indirect estimation can be performed with model-based or data-driven methods. Once again, data-driven models enable a simpler black-box approach that does not involve the modelling of physical phenomena.

Increasing computational power, availability of big-data and capability to process them, and the possibility to combine multiple sensors enhanced data-driven methods involving various ML algorithms and deployment of NNs [4]. These methods enable real-time monitoring of the battery state by processing input data in very limited time, and, therefore, a big effort is devoted by researchers to investigate new applications and solutions. However, the development of a data-driven model involves collection of a large amount of data, which is time and resource consuming. Therefore, challenges arise in terms of affordable collection of high-quality training datasets, and implementation of new solutions for the overall architecture of the model. For these reasons, this paper aims to: (i) discuss latest contributions that have not yet been included in other reviews and (ii) to schematize a data-driven model that is trained on simulated data. Collection of high-quality data is essential in implementation of data-driven models. Indeed, dataset collected in scenarios with different load-history and temperature has higher quality and enables training of models with higher capability to generalize. However, the overall data collection process becomes more expensive and time-consuming as the scope of the data collection increases. In this context, high-fidelity physics-based models can provide data with a wide range of scenarios considered and this is the reason why a model based on the use of a simulation dataset was investigated in this work.

2. STATE OF THE ART

In this section, firstly a summary of typical workflow for the application of ML algorithms and NN in SoC and SoH estimation is provided, and then scientific advancements reported in the latest research papers, not included in previous reviews, are discussed.

2.1 *A workflow for data-driven applications in SoC and SoH estimation*

ML algorithms are typically divided into unsupervised, and supervised. Unsupervised ML algorithms deal with unlabeled data and are used for clustering, dimensionality reduction, and finding relations; supervised ML deals with classification and regression problems and involve respectively categorical and continuous labeled data. The choice of the appropriate ML algorithm depends on the *specific task* to be addressed. In particular, the estimation of SoC and SoH mainly involves regression algorithms.

For these kinds of data-driven methods, the use of a dataset with adequate quality and size plays a key role [3]. Indeed, data can be collected with dedicated experimental campaigns, simulations with high-throughput calculation, and mining from literature and datasets available from other studies [5]. Simulated or experimental data can account variability of many factors that impact the operating conditions of the batteries, such as the load profile and the temperature. Considering higher number of levels for each factor is essential to improve the capability of the trained model to generalize over a broader range of scenarios [3], and, therefore, leads to higher quality of the dataset.

Dedicated experimental campaigns either in a laboratory or in real vehicle operating conditions enable collection of high-fidelity data, however, they can be time-consuming and expensive [6]. On the other hand, the use of high-fidelity simulation models can enable the collection of larger structured datasets, with various simulated scenarios that may be difficult to be easily reproduced in a laboratory. Additionally, in the last decade, researchers have been storing open-source repository datasets of measurement campaigns to make them available to wider research community [7].

Once the *data collection* process is completed, the next step to be implemented is *feature engineering*, which enables a representation of the input data, using descriptors that are properly identified through preliminary analysis. For the application under study, these data typically consist of signals related to electric, mechanical, and thermal parameters. Typical processes of features' extraction from signals involve the calculation of statistical descriptors and either frequency or time-frequency domain analysis. Several types of NNs, such as CNNs, RNNs, and deep NNs, can successfully process portions of signals without previous extraction of features, avoiding the feature engineering step.

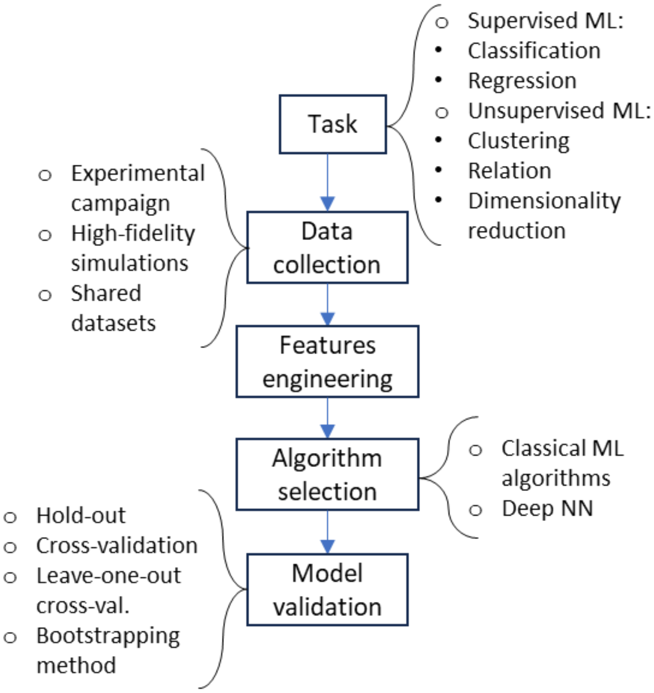

Fig. 1. An ML implementation workflow.

After completing the *feature engineering* step, proper *model choice*, based on ML and NNs algorithms can be performed in order to estimate SOC and SOH. These algorithms include linear regression, SVR, shallow NN, deep NNs, CNNs, and RNNs, which will be analyzed in detail in the next subsection along with studies that involved their deployment [4].

Model validation is finally performed for the evaluation of prediction performances. Typical metrics that are used to evaluate regression models are the mean square error (MSE), the root mean square error (RMSE), the mean absolute error (MAE), and the correlation coefficient R^2. Metrics used for classification tasks are the accuracy, the error rate, and the recall. As the model is trained to process data not included in the training set, it is important to test its capability to generalize. Indeed, an error in the model with data used in the training set (*training error*) is distinguished from a *generalization error*, which is the error that occurs when the model works on new samples. For this reason, the dataset is normally split into training, validation, and test sets. The validation and test sets are mutually exclusive with the training set, and to ensure this, four methods are typically used to split the dataset and to evaluate the selected metrics: the hold-out, the cross-validation, the leave-one-out cross-validation and the bootstrapping method [5].

2.2 Analysis of some of the latest contributions available in the literature

As mentioned in the previous subsection, there is a wide range of algorithms that have been considered in research papers available in the literature to train regression models for SoC and SoH estimation. They focus on algorithmic advances to improve performances of the models, on the inclusion of new information gathered with sensor fusion, and on novel solutions to overcome current challenges.

Guo and Ma [8] benchmarked different state-of-the-art types of deep NN, such as a simple FCNN, LSTM, GRU and temporal convolutional network, which is a particular type of CNN, for the SoC estimation. They trained these NNs with battery current and voltage, and evaluated the performances in terms of capability to generalize, computational efficiency and noise robustness. The results indicated that the FCNN showed the highest RMSE compared to the other algorithms; however, it was the most efficient in terms of computation time due to its simpler structure. On the other hand, the temporal convolutional network showed the best performance in terms of data noise robustness, thanks to its convolutional layers.

Besides current and voltage, Sulaiman et al. considered both ambient and battery cell temperature as input to train an FFNN for SoC estimation, using data collected during 70 trips of a BMW i3 [9]. Several training algorithms were benchmarked, such as Particles Swarm Optimization, Genetic Algorithm, Differential Evolution, Adaptive Moment Estimation, and Evolutionary Mating Algorithms. Results highlighted

Tab. 1. Summary of the research papers analyzed in this survey.

Ref.	ML algorithm	Input	Temperature	Load profile	Performance
[6]	DNN	Labeled and unlabeled current and voltage	10°C, 25 °C, 40°C	UDDS, LA92, and mixed.	-
[7]	Attention-based deep learning	Current, voltage, and temperature	≤ 45° C	Data collected during real trips	MAE_{SoC}≤ 2.5% $MAPE_{SoH}$≤ 2.5%
[8]	FCNN, LSTM, GRU and TCN	Current, voltage, and temperature	0 °C, 10°C, 25 °C, 40°C	US06 , HWFET, UDDS, LA92.	$RMSE_{GRU., LSTM, TCM}$ ≤ 3%
[9]	FFNN, different training algorithms	Current, voltage, and cell and ambient temperature	Ranging between 5° C and 35° C	Data collected during real trips	RMSE = 4.70%
[10]	LSTM	Current, voltage, and battery stress	25 °C	Constant, short and pulse conditions, NEDC and UDDS	RMSE = 1.88%
[11]	Moving window-incremental learning RLM+ Coulomb count.	Current and voltage	0 °C, 25 °C, 45°C	UDDS, NYCC, NEDC, and Japanese 1015.	RMSE ≤ 2%
[12]	MLR, SVR, and random forest	Partial charging times	Controlled with climate chamber	NASA Ames Prognostics Center of Excellence - data repository	R^2_{SVM}=97.2%

the lowest error and the second-lowest time required for training for evolutionary mating algorithm.

Jiang et al. also considered stress measurement to train LSTM for SoC estimation [10]. Preliminary results revealed that the integration of mechanical measurements with electrical ones leads to improved efficiency when ensuring accuracy, however, further developments need to be addressed, such as optimization of hyperparameters and the use of a dataset with multiple temperatures.

Wang et al. developed a novel hybrid model that uses moving mean and incremental approach in combined implementation of RVM and Coulomb counting to estimate SoC [11]. The dataset was obtained by merging experimental and simulated data, respectively from the Center for Advanced Life Cycle Engineering at the University of Maryland and the Advanced Vehicle Simulator developed by the Renewable Energy Laboratory. Validation and test revealed that the RMSE can stay below 2%.

To process both labeled and unlabeled data and, therefore, address data-hungriness of NN and costs of the labelling process, Ma and Zhang developed an input reconstruction-aided network to estimate SoC [6]. This network was made of an input reconstruction module and a linear unit, where the former was composed of an encoder and a decoder, both realized with LSTM layers. Datasets with different labeled/unlabeled data proportions have been compared in training of the proposed algorithm. Results showed an accuracy increase of more than 14% when large amounts of unlabeled observations were included, in comparison with training processes with no additional unlabeled observations.

With the purpose of exploiting data generated during on-field-operation, Shi et al. implemented a cloud-based data-driven architecture to model battery cell behavior for real-life-electric vehicle applications [7]. Such a framework consisted of sensors collecting data and continuously transmitting them to the cloud, to improve capability of SoC and SoH estimators, through offline procedures. Updated models were then used for onboard monitoring and diagnosis during operations.

Marri et al. benchmarked classical ML algorithms for SoH estimation, such as MLR, polynomial regression, SVR, and random forest, as they allow good accuracy and lower computation effort [12]. Features were selected based on analysis of partial charging times. Results indicated that partial charging times are well correlated with a number of charging/discharging cycles, when voltage is higher than 3.7. The overall best performance for all feature sets was achieved using SVR.

3. CASE STUDY

As discussed in the previous paragraphs, generation of a dataset that encompasses a wide range of load-profiles and environmental conditions is fundamental to train models with acceptable capability to generalize. However, this is an expensive and time-consuming process, especially in the case of SoH estimation, in which required data should involve the entire useful life of the battery system. For this reason, this study aims to contribute to this challenge with a methodology which is schematized in Figure 2 and involves the use of a validated physics-based model with lumped-parameters to generate a training dataset [2]. This case study considers a pouch Li-NMC battery cell with a capacity of 20 Ah and a rated voltage of 3.7 V, with a physical size (W x L x T) of 145 mm^3. To introduce a variety in the training dataset, during the model-based generation of data, a set of plausible numeric values is randomly selected based on the statistical distribution of typical values of each of the lumped parameters. Then, permutations can be carried out to simulate non-identical data. Different environmental conditions, driving styles and load profiles are set for the generation of the training dataset, which is used to train and benchmark different state-of-the-art types of deep NNs for the estimation of the SoC and SoH. Finally, validation of the trained model is carried out by regression of data that were collected during experiments to calibrate and validate the physics-based model with lumped-parameters. In this way, a NN is trained with synthetic data, and validated with real data.

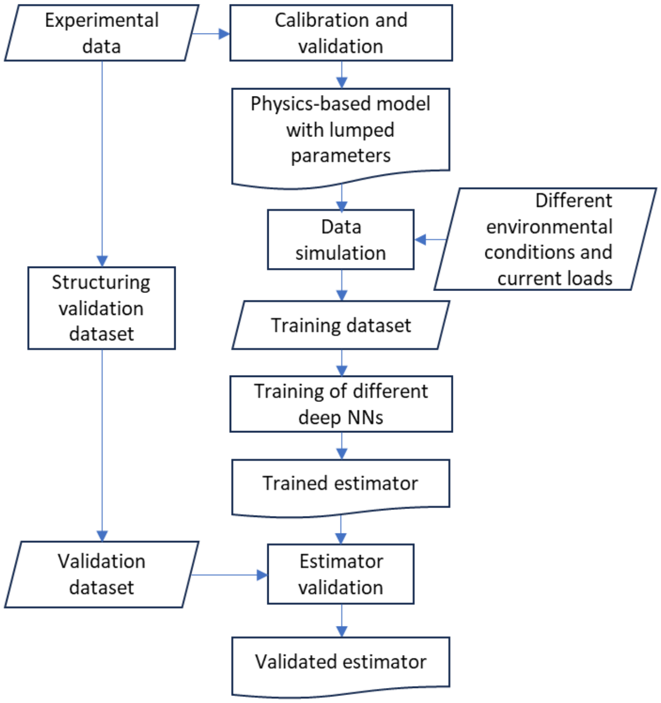
Experimental data
Calibration and validation
Physics-based model with lumped parameters
Data simulation
Different environmental conditions and current loads
Structuring validation dataset
Training dataset
Training of different deep NNs
Trained estimator
Validation dataset
Estimator validation
Validated estimator

Fig. 2. Schematic representation of the proposed methodology for SoC and SoH estimation.

4. CONCLUSIONS

In this study, a survey on the latest advancements in ML applications for SoC and SoH estimation is carried out. The present analysis highlighted that the latest research trends focus on both algorithmic improvement and benchmarking of the learning algorithms, and inclusion of as much data as possible to improve the prediction performances.

Two research streams for inclusion of larger amount of data can be observed: one focuses on inclusion of new information such as mechanical stress, and the other focuses on development of new architectures that integrate unlabeled datasets or cloud-based data collected during operation. Additionally, all the reviewed studies remarked the importance to collect data with as much environmental and load conditions as possible to ensure the highest capability of the trained model to generalize.

This is due to the need for a large dataset that encompasses as much operating scenarios as possible to achieve good generalization capability. However, this is a time and resource-consuming process, and, to cope with this challenge a methodology that exploits simulation of training dataset with a validated model is schematized and proposed. Full development and demonstration of such methodology as a valid and convenient technique will be addressed in future research.

DECLARATION OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors read and approved the final manuscript.

REFERENCE

[1] Zwicker MFR, Moghadam M, Zhang W, Nielsen C V. Automotive battery pack manufacturing – a review of battery to tab joining. J Adv Join Process 2020;1:100017.

[2] Capasso C, Iannucci L, Patalano S, Veneri O, Vitolo F. Battery Thermal Management Systems: A Case Study on Li-NMC storage systems for electric vehicles. 2023 IEEE Int Conf Electr Syst Aircraft, Railw Sh Propuls Road Veh Int Transp Electrif Conf ESARS-ITEC 2023 2023:1–7.

[3] Vidal C, Malysz P, Kollmeyer P, Emadi A. Machine Learning Applied to Electrified Vehicle Battery State of Charge and State of Health Estimation: State-of-the-Art. IEEE Access 2020;8:52796–814.

[4] Ren Z, Du C. A review of machine learning state-of-charge and state-of-health estimation algorithms for lithium-ion batteries. Energy Reports 2023;9:2993–3021.

[5] Wei Z, He Q, Zhao Y. Machine learning for battery research. J Power Sources 2022;549:232125.

[6] Ma L, Zhang T. Deep learning-based battery state of charge estimation: Enhancing estimation performance with unlabelled training samples. J Energy Chem 2023;80:48–57.

[7] Shi D, Zhao J, Wang Z, Zhao H, Eze C, Wang J, et al. Cloud-Based Deep Learning for Co-Estimation of Battery State of Charge and State of Health. Energies 2023;16:1–19.

[8] Guo S, Ma L. A comparative study of different deep learning algorithms for lithium-ion batteries on state-of-charge estimation. Energy 2023;263:125872.

[9] Sulaiman MH, Mustaffa Z, Zakaria NF, Saari MM. Using the evolutionary mating algorithm for optimizing deep learning parameters for battery state of charge estimation of electric vehicle. Energy 2023;279:128094.

[10] Jiang B, Tao S, Wang X, Zhu J, Wei X, Dai H. Mechanics-based state of charge estimation for lithium-ion pouch battery using deep learning technique. Energy 2023;278:127890.

[11] Wang C, Zhang X, Yun X, Fan X. A novel hybrid machine learning coulomb counting technique for state of charge estimation of lithium-ion batteries. J Energy Storage 2023;63:107081.

[12] Marri I, Petkovski E, Cristaldi L, Faifer M. Comparing Machine Learning Strategies for SoH Estimation of Lithium-Ion Batteries Using a Feature-Based Approach. Energies 2023;16.