Journal of Electronic Research and Application, 2024, Volume 8, Issue 6
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

An Ensemble Learning Method for SOC Estimation of Lithium-Ion Batteries Using Machine Learning

Yirga Eyasu Tenawerk[1]*, Linqing Xia[2], Jingfei Fu[1], Wanwen Wu[1], Zewei Quan[1], Wu Zhen[1]

[1]School of Information Engineering, Nanchang Hangkong University, Nanchang 330063, China
[2]Shanghai Dijietong Digital Technology Co., Ltd., Shanghai 200090, China

***Corresponding author:** Yirga Eyasu Tenawerk, eyutenaw243@gmail.com

Abstract: Accurately assessing the State of Charge (SOC) is paramount for optimizing battery management systems, a cornerstone for ensuring peak battery performance and safety across diverse applications, encompassing vehicle powertrains and renewable energy storage systems. Confronted with the challenges of traditional SOC estimation methods, which often struggle with accuracy and cost-effectiveness, this research endeavors to elevate the precision of SOC estimation to a new level, thereby refining battery management strategies. Leveraging the power of integrated learning techniques, the study fuses Random Forest Regressor, Gradient Boosting Regressor, and Linear Regression into a comprehensive framework that substantially enhances the accuracy and overall performance of SOC predictions. By harnessing the publicly accessible National Aeronautics and Space Administration (NASA) Battery Cycle dataset, our analysis reveals that these integrated learning approaches significantly outperform traditional methods like Coulomb counting and electrochemical models, achieving remarkable improvements in SOC estimation accuracy, error reduction, and optimization of key metrics like R^2 and Adjusted R^2. This pioneering work propels the development of innovative battery management systems grounded in machine learning and deepens our comprehension of how this cutting-edge technology can revolutionize battery technology.

Keywords: SOC; Lithium-ion batteries; Random Forest Regressor; Gradient Boosting Regressor; Machine Learning

Online publication: December 2, 2024

1. Introduction

Lithium-ion batteries, ubiquitous in electric vehicles, portable electronics, and large-scale stationary storage systems, necessitate precise State of Charge (SOC) determination to optimize battery lifespan and mitigate risks of failure or thermal runaway [1,2]. While widely adopted, traditional SOC estimation methods, including Coulomb counting and model-based procedures, often struggle with accuracy and efficiency issues, particularly under varying operational conditions [3]. SOC estimation is a pivotal aspect of battery management systems (BMS), crucial for assessing battery performance and safety in diverse applications like electric vehicles and

renewable energy systems. Lithium-ion batteries undergo intricate electrochemical processes during operation, with complex interdependencies among factors and inconsistent operating conditions, rendering SOC estimation a formidable challenge [4]. To address these challenges, researchers have turned to machine learning, which offers promising avenues to enhance SOC estimation accuracy. Recent advancements in machine learning capitalize on the non-linear and multi-process nature of battery behavior, eliminating the need for explicit modeling [5]. This makes machine learning particularly advantageous in refining SOC estimates compared to traditional methods, especially under diverse operational settings, thereby advancing the capabilities of BMS and ensuring reliable battery performance and safety.

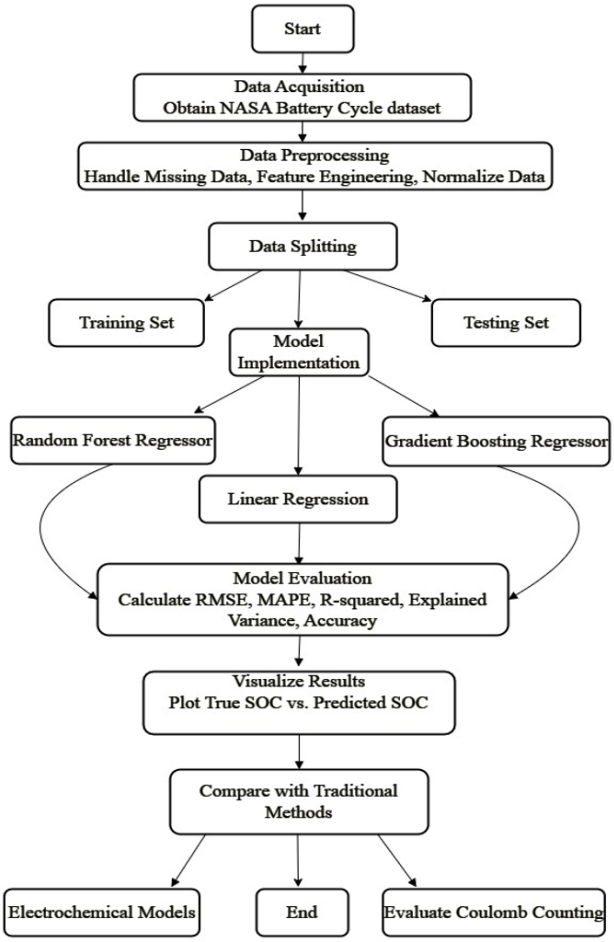

Figure 1. Methodology flowchart

The methodology flowchart is shown in **Figure 1**. This research evaluates the predictive capabilities of three machine learning algorithms, specifically the Random Forest Regressor, Gradient Boosting Regressor, and Linear Regression, in determining SOC. The present research will demonstrate the efficiency of these models in addressing the identified drawbacks of old approaches of SOC estimation. This will be done by evaluating their quantitative criteria, such as Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), R-squared, Explained Variance score, and accuracy within a 5% negligible difference. Ultimately, it can be affirmed that the utilization of machine learning methods for SOC assessment enhances the efficiency and longevity of lithium-ion batteries in various applications. Although current approaches have limitations, new models aim to address these difficulties by improving BMS and enabling additional innovation to enhance energy storage devices [6]. The study has the following contributions:

(1) To develop machine learning models for SOC estimation.

(2) The performance of these models is compared using various evaluation metrics.

(3) To visualize the SOC estimation performance through graphical analysis.

2. Ensemble learning approaches for SOC estimation

Ensemble learning approaches have garnered significant interest recently because they can achieve high accuracy rates by combining the outputs of multiple models. This research employed three main ensemble techniques, namely Random Forest Regressor, Gradient Boosting Regressor, and Linear Regression (used as the baseline model), to estimate the SOC in Lithium-ion batteries.

2.1. Data collection and preparation

The main data source utilized in this research is the NASA Battery Cycle dataset, which includes crucial information for predicting SOC, such as cycle number, measured voltage, measured current, measured temperature, current, voltage, and time. The primary factor is the high density and presence of several cycles in the records of this dataset, which enables the acquisition of comprehensive information regarding the battery and its performance under various conditions [7,8].

Before training the models, it demonstrated that all data underwent preprocessing to a specific level to guarantee the data's quality and dependability. Specific measures were implemented to address the missing values, such as imputation or deletion of records based on the need and missing data in a model. That is in response to instances that impacted the overall outcomes due to outliers, and they were addressed using statistical tools or specialized knowledge in that specific domain. Standardization is employed to ensure that each feature is mapped into a uniform scale, hence preventing any bias in the model caused by features with large ranges [9].

Feature engineering played a crucial role in enhancing the models' performance by accurately predicting outcomes. Features were employed to condense information about battery cycles by analyzing derived voltages and currents, as well as voltage and current rates within a kilocycle, cycle, or temperature trends over specific time intervals. The incorporation of these designed characteristics and observations allowed for a deeper understanding of the relationship between the input variable and SOC, hence improving the accuracy and robustness of the estimation models. The meticulous data pretreatment and feature engineering work established a solid groundwork for the application of ensemble learning methods. They supplied valuable sources for evaluating different algorithms on SOC estimation in lithium-ion batteries.

2.2. Ensemble learning models

The Random Forest Regressor uses the ensemble learning technique to create many decision trees throughout the training process. Every tree is trained using a randomly selected portion of the data and a randomly selected portion of the features. Averaging the previous research's predictions from individual trees enhances prediction accuracy and mitigates overfitting [2]. Random Forests provide resilience to data noise and demonstrate efficient handling of datasets with high dimensions [10].

Gradient Boosting constructs models in a sequential manner. Each subsequent iteration of the model strives to minimize the inaccuracies present in the previous versions. This strategy improves the formation of more precise associations within the group, which is particularly beneficial in evaluating the variability within the existing dataset. Typically, Gradient Boosting is susceptible to overfitting, although it offers high model accuracy and interpretability [11].

In this research, linear regression is employed as a benchmark or the principal methodology for comparison with other methodologies. While the model's relationship with the input characteristics is linear or additive in terms of the Sum of Coefficients (SOC), it provides clear and easily understandable explanations of the contributions made by each feature. Linear regression is simpler than ensemble approaches but can also offer detailed insights into the distribution and features of the dataset.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p \quad (1)$$

Where $\hat{y}$ is the predicted SOC and β_0 is the intercept term.

2.3. Model training and evaluation

The dataset was subsequently partitioned into the training dataset and the validation dataset to optimize the model's training and evaluation procedures. Cross-validation techniques, such as k-fold cross-validation, are employed to mitigate overfitting and assess the performance of the models on unseen data. The process of hyperparameter tuning involved utilizing methods such as grid search or randomized search to select the optimal set of parameters for each of the approaches employed in ensemble learning.

2.4. Performance evaluation metrics

Various assessment methodologies were used to measure the ensemble learning models' performance in estimating the SOC of lithium-ion batteries. RMSE measures the average difference of errors in SOC values between predicted and observed, assessing the model's precision on the data set.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(y_i - \hat{y})^2} \quad (2)$$

MAPE is the average percentage deviation of SOC values estimated using the predicted and actual values, revealing the relative degree of accuracy in the prediction [12].

$$MAPE = \frac{1}{n}\sum_{i-1}^{n}\left|\frac{y - \hat{y}_i}{y_i}\right| \quad (3)$$

R2 denotes the proportion of SOC variation attributable to the independent variables in the model, and it captures the model's goodness of fit [4,3].

$$R^2 = 1 - \frac{\sum_{i-1}^{n}(y - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2} \quad (4)$$

The explained Variance Score shows how well the model describes the SOC values by recording the percentage of the collective sum of squares that the model accounts for [7].

$$Explained\,Varience = 1 - \frac{Var\,(y - \hat{y})}{Var(y)} \quad (5)$$

Tolerance Levels evaluate the ratio of SOC predictions ranging in a certain tolerance margin (e.g. ± 5%), which brings out the model's accuracy level. Collectively, these metrics help evaluate the performance of ensemble learning models in SOC estimation and the choice and fine-tuning of BMS for specific applications.

2.5. Experimental setup and implementation details

The SOC estimating models were implemented using the Python programming language, which is widely used in machine learning. The Scikit-learn libraries were selected for model construction and evaluation. To attain reliable classification and precise prediction, this research utilized data from NASA's Battery Cycle dataset, which was divided into two sets: the training set (80%) and the testing set (20%). Various techniques, including addressing missing data, normalizing using the Min-Max Scalar, and doing basic feature engineering with a focus on date time for temporal data, were utilized to enhance the performance of the model. The models were refined by hyperparameter tuning utilizing grid search and cross-validation folds to enhance their generalization ability. To assess the effectiveness of the models employed for evaluation, discrepancy metrics such as mean squared error (MSE), R^2, and explained variance score were utilized. Utilizing Matplotlib and Seaborn, the analysis and projections were able to find patterns in SOC values that are significant for managing

battery systems in different applications. **Figure 1** presents a flowchart that provides a concise overview of the research methodology. It illustrates the primary activities and sub-activities involved in the learning and development of the safety culture.

3. Model performance

The results for the best-performing model are as follows. The evaluation metrics for the SOC estimation models in **Table 1** indicate the performance of the ensemble methods used in this research. The RMSE is exceptionally low at 6.61×10^{-5}, suggesting that the models produce minimal errors in predicting SOC values. Similarly, the MAPE is very low at 0.0054, reflecting the models' accuracy regarding the percentage difference between predicted and actual SOC values. The R^2 value is calculated to be 0.9999995, which suggests that most of the variation in SOC can be attributed to the inputs of the models, emphasizing their effectiveness. The Explained Variance Score of 0.9999995 further confirms the models' ability to account for the predicted SOC values' variance effectively.

Table 1. Evaluation Metrics for SOC Estimation Models

Metric	Value
RMSE	6.60527914389
MAPE	0.00542653153
R-squared	0.9999995483553943
Explained Variance Score	0.9999995483555109
Accuracy within 5% Tolerance	100.00%

The accuracy within a 5% tolerance level is 100%, indicating that all SOC predictions fall within a ± 5% range of the actual values. These results collectively suggest that the ensemble methods, particularly the Random Forest Regressor and Gradient Boosting Regressor, significantly outperform traditional methods and simple linear regression in SOC estimation. The low RMSE and MAPE values and high R^2 and explained variance scores demonstrate the models' ability to accurately predict SOC under varying conditions, making them highly suitable for optimizing BMS across diverse applications.

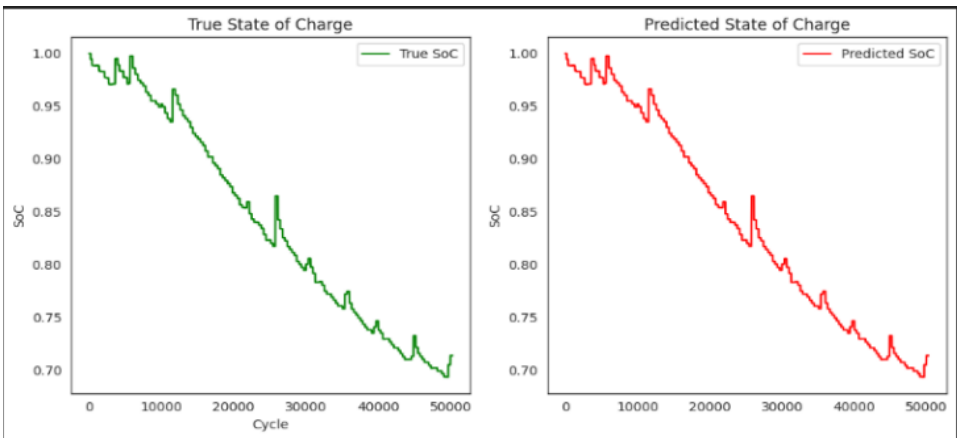

Figure 2. Comparison of true and predicted

Figure 2 displays the original SOC values in green and the anticipated SOC values using the model in red. Specifically, the actual values demonstrate that the anticipated values of the models closely align with these outcomes. The little discrepancies between the predicted and actual SOC levels confirm the effectiveness of ensemble learning methods in accurately determining SOC levels in various situations. These results highlight the potential for using these models in BMS to enhance their efficiency and safety.

3.1. Comparison with traditional methods

Ensemble learning methods offer advantages for prediction and SOC estimate in the following ways. Coulomb counting is considered risky due to the accumulation of errors from sensors on Internet of Things (IoT) devices. Over time, it may become less accurate compared to ensemble methods like Random Forest Regressor and Gradient Boosting Regressor. These methods utilize multiple Random Decision Trees and Iterative Boosting of decision trees, respectively. In addition, Shrivastava and Soon highlighted that machine-learning models are a superior alternative to Kalman filters and electrochemical models [13]. The latter would require precise data on battery parameters and significant processing resources. These models are independent of value assessments, system features, and other aspects. They are capable of performing effectively in varying situations and battery states, hence reducing the need for extensive testing and validation.

Ensemble approaches enhance the accuracy of SOC forecasts by addressing any limitations that individual models may have, as the results are reached separately [14]. These batteries are particularly suitable for real-time systems like electric vehicles, consumer electronics, and renewable energy systems. In these systems, it is crucial to correctly and reliably predict the SOC of the batteries to ensure optimal performance and safe operation. Overall, incorporating ensemble learning methods in SOC estimate is a valuable enhancement compared to traditional methodologies, as it offers numerous advantages, including improved accuracy, speed, and adaptability in addressing lithium-ion phosphate (LIP) related challenges.

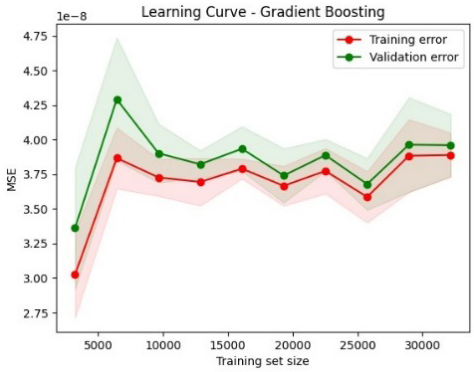

Figure 3. Learning curve for Gradient Boosting model

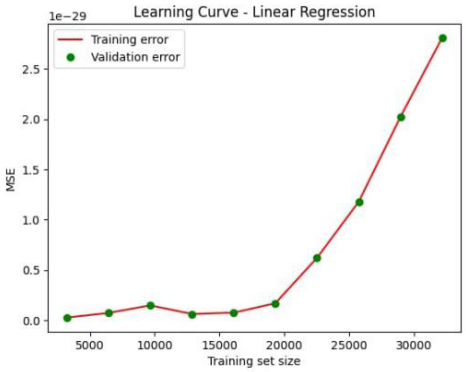

Figure 4. Learning curve for Linear Regression model

Figure 3 depicts the progression of the learning curve for the Gradient Boosting model. As the size of the training data set increases, both the training error and cross-validation fall significantly, indicating efficient learning. Once the number of samples reaches 15,000 to 20,000, the errors stabilize and become consistently low. This indicates a high level of generalization and a favorable balance between bias and variance. This demonstrates the efficacy and precision of the proposed model in BMS for estimating SOC.

Figure 4 demonstrates that as the size of the training set increases, both the training and validation errors initially remain low but then experience a large increase, suggesting the occurrence of underfitting. The rise

in both indicators demonstrates that the model is insufficiently sophisticated to achieve the necessary level of complexity for SOC estimation. Hence, the observed increasing error pattern demonstrates the model's flaws and necessitates the advancement of more sophisticated models, such as ensemble approaches, to achieve more precise SOC estimation in BMS.

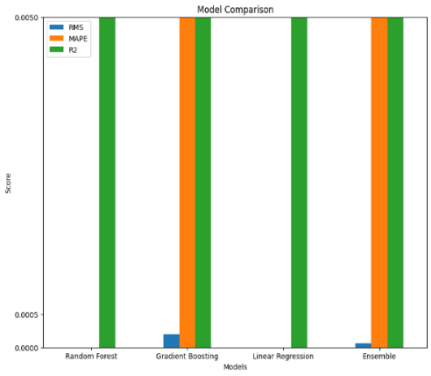

Figure 5. Model comparison for SOC estimation

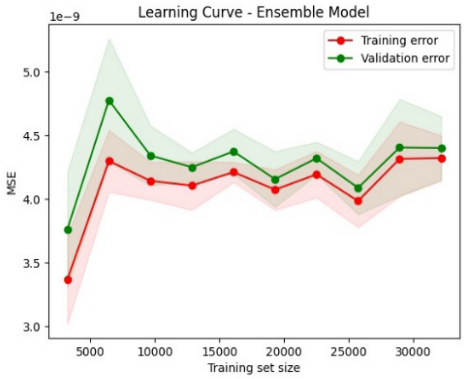

Figure 6. Learning curve for Ensemble model

Figure 5 compares Random Forest, Gradient Boosting, Linear Regression, and an Ensemble model to evaluate RMS, MAPE, and R^2 scores to support the concept of SOC accuracy and the ensemble approach. **Figure 6** illustrates the performance of the Ensemble model in estimating SOC. It displays the MSE for both the training and validation datasets, relative to the size of the training dataset. The increasing convergence of errors as the training set size grows suggests that the model is both robust and accurate in its ability to forecast SOC. This highlights the model's efficacy in battery management applications.

3.2. Practical implications

Improving the accuracy of SOC prediction has crucial applications and practical implications for BMS. Accurate SOC prediction enhances the ability to regulate the charging and discharging of batteries and optimize their total lifespan while minimizing the likelihood of failure. Moreover, in real-world scenarios, the precise forecasting of SOC under various operational circumstances enhances the reliability of energy storage systems, as explained in detail in the methodology.

3.3. Limitations and future work

Although there are overall patterns in the performance of the equity market, the research has certain limitations. The utilized models underwent training and testing using a certain dataset. Thus, it is not necessarily predictable that comparable results are achieved with different datasets and battery kinds. Additional investigation is necessary to see if comparable attributes of battery models are utilized for alternative battery compositions and under varying operational circumstances. Furthermore, including real-time SOC prediction into the BMS and exploring the use of deep learning techniques could enhance accuracy and reduce susceptibility to noise [15,16].

4. Conclusion

The research validates that when comparing the efficacy of ensemble learning techniques, specifically the

Random Forest Regressor and Gradient Boosting Regressor, the estimation of SOC in lithium-ion batteries is more precise when utilizing ensemble learning methods. Therefore, the machine learning models demonstrate superior accuracy in predicting SOC compared to a basic regression method, as seen by the low RMSE and MAPE values and high R^2 and explained variance. Considering this level of performance, they are ideal for direct implementation in various domains, such as electric vehicles and renewable energy systems. According to the assessment, their proposed models have high accuracy and low error rates, making them an effective and reliable option for continuous SOC prediction. By using the unlimited number of compositors in the group, it is feasible to eliminate individual imperfections and get a superior level of predictive precision. Ensemble learning offers this approach as one of its advantages. This work contributes to the existing research on enhancing the effectiveness and security of lithium-ion batteries.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Mehmet K, 2023, SoC Estimation of Lithium-Ion Batteries based on Machine Learning Techniques: A Filtered Approach. Journal of Energy Storage, 72: 108268.

[2] Cao M, Zhang T, Wang Y, et al., 2020, A Deep Learning Method with Ensemble Learning for Capacity Estimation of Lithium-ion Battery. Global Reliability and Prognostics and Health Management (PHM-Shanghai), 2020: 1–6.

[3] Tian H, Li A, Li X, 2021, SOC Estimation of Lithium-Ion Batteries for Electric Vehicles based on Multimode Ensemble SVR. Journal of Power Electronics, 21(9): 1365–1373.

[4] Shen S, Sadoughi M, Li M, et al., 2020, Deep Convolutional Neural Networks with Ensemble Learning and Transfer Learning for Capacity Estimation of Lithium-Ion Batteries. Applied Energy, 260: 114296.

[5] Gou B, Xu Y, Feng X, 2021, An Ensemble Learning-Based Data-Driven Method for Online State-of-Health Estimation of Lithium-Ion Batteries. IEEE Transactions on Transportation Electrification, 7(2): 422–436.

[6] Wang Y, Kou P, Fan J, et al., 2022, A Novel Capacity Estimation Method for Li-Ion Battery Cell by Applying Ensemble Learning to Extremely Sparse Significant Points. IEEE Access, 10: 96427–96441.

[7] Hannan MA, How DNT, Hossain MS, et al., 2020, SOC estimation of li-ion batteries with Learning Rate-Optimized Deep Fully Convolutional Network. IEEE Transactions on Power Electronics, 36: 7349–7353.

[8] Li XJ, Yu D, Byg VS, et al., 2023, The Development of Machine Learning-Based Remaining Useful Life Prediction for Lithium-Ion Batteries. Journal of Energy Chemistry, 82: 103–121.

[9] Lv C, Zhou X, Zhong LX, et al., 2022, Machine Learning: An Advanced Platform for Materials Development and State Prediction in Lithium-Ion Batteries. Advanced Materials, 34: 2101474.

[10] Kunapuli G, 2023, Ensemble Methods for Machine Learning. Simon and Schuster, London.

[11] Sheykhmousa M, Mahdianpari M, Ghanbari H, et al., 2020, Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 6308–6325.

[12] Karch J, 2020, Improving on Adjusted R-Squared. Collabra: Psychology, 6(1): 45.

[13] Chicco D, Warrens MJ, Jurman G, 2021, The coefficient of Determination R-Squared is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. PeerJ Computer Science, 7: e623.

[14] Shrivastava P, Soon TK, Idris MYIB, et al., 2019, Overview of Model-Based Online State-of-Charge Estimation using Kalman Filter Family for Lithium-Ion Batteries. Renewable and Sustainable Energy Reviews, 113: 109233.

[15] How DNT, Hannan MA, Lipu MSH, et al., 2019, State of Charge Estimation for Lithium-Ion Batteries using Model-Based and Data-Driven Methods: A review. IEEE Access, 7: 36116–136136.

[16] Nabipour M, Nayyeri P, Jabani H, et al., 2020, Predicting Stock Market Trends using Machine Learning and Deep Learning Algorithms via Continuous and Binary Data; A Comparative Analysis. IEEE Access, 8: 150199–150212.