2025 11th International Conference on Control, Instrumentation and Automation (ICCIA)

Robust State-of-Health Estimation in Second-Life Li-Ion Batteries Using Ensemble Learning and Noise-Injected Training

Mahdi Alinaghizadeh Ardestani
Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, Technical and Vocational University (TVU)
Tehran, Iran
ardestani@tvu.ac.ir

Nika Rahmani
Artificial Intelligence Research Group, Artificial Intelligence and Control Engineering Research lab (AICER)
Tehran, Iran
nika@aicer.ir

***Abstract*— Accurate State of Health (SoH) estimation is crucial for reusing retired lithium-ion batteries safely. Existing methods lack robustness against real-world sensor noise and disturbances. We propose a noise-augmented training method with feature clipping to improve prediction reliability under measurement imperfections. Our approach enhances model resilience in industrial environments, where noise and calibration drift degrade performance. Evaluations show Random Forest Regression outperforms other models, achieving a 25% lower RMSE (0.85% vs. 1.14%). This demonstrates its effectiveness for practical second-life battery applications, balancing speed and accuracy. The method's noise-resistant design ensures reliable SoH estimation, supporting sustainable battery reuse in energy storage systems.**

Keywords-- Battery health estimation, Electrochemical Impedance Spectroscopy (EIS), Lithium-ion batteries, Machine learning, Random Forest

I. INTRODUCTION

Lithium-ion batteries have become a cornerstone of today's energy landscape, powering everything from electric vehicles (EVs) and portable electronics to large-scale renewable energy storage systems. Their combination of high energy density, long cycle life, and low self-discharge rate has made them the preferred choice for both mobile and stationary applications [1]. With the rapid growth in demand for EVs and clean energy technologies, the global lithium-ion battery market is projected to surpass $180 billion by 2030, driven largely by their adoption in transportation and grid storage [2]. At the same time, the increasing number of EV batteries reaching the end of their primary service life is creating a strong incentive for second-life applications in stationary energy storage, offering both environmental and economic benefits [3]. This shift supports a more sustainable energy ecosystem but also raises new challenges in battery assessment, safety, and lifecycle management.

Beyond transportation and consumer devices, lithium-ion batteries are becoming essential for building reliable and flexible energy systems. They make it easier to integrate renewable sources like wind and solar into the grid, support backup power solutions, and drive innovations such as microgrids and smart homes [4]. Advances in battery chemistry, cooling, and recycling are improving their lifespan and sustainability [5], while their adaptability has expanded their use to fields like aerospace, healthcare, and manufacturing [6]. These trends highlight that the future of clean energy and modern infrastructure will remain closely tied to the continued development and deployment of lithium-ion battery technology.

Reliable estimation of a battery's State of Health (SoH) is critical for both primary and second-life applications, directly influencing safety, performance, and lifecycle management. SoH reflects the degradation level of a battery and is typically defined as the ratio of the current capacity to the nominal capacity. In electric vehicles and grid storage systems, incorrect SoH prediction can lead to premature replacement, underutilization, or, worse, catastrophic failures due to undetected capacity loss or internal damage [7], [8]. Moreover, real-time or near-real-time SoH monitoring is essential for effective battery management systems (BMS), enabling smart decision-making in charge/discharge control, thermal regulation, and remaining useful life estimation [9], [10].

While traditional approaches such as incremental capacity (IC) analysis, differential voltage analysis, and equivalent-circuit modeling remain prevalent [11], they often depend on precise laboratory conditions and may not generalize well to diverse real-world aging patterns. A recent onboard health estimation framework, for example, employs Electrochemical Impedance Spectroscopy (EIS) deconvolved via the Distribution of Relaxation Times (DRT) as input to an LSTM-based model, achieving robust SoH predictions across varying degradation ages and operating conditions with an average

979-8-3315-5950-2/25/$31.00 ©2025 IEEE

RMSPE of approximately 1.69 percent [12]. Such methods provide valuable insights but still face limitations, including high computational cost, sensitivity to measurement noise, and dependence on internal battery parameters that are rarely available in practical systems. In contrast, data-driven approaches based on machine learning and impedance measurements—particularly EIS—have emerged as powerful tools for SoH estimation, offering the potential for non-invasive, accurate, and fast diagnostics under diverse conditions [13]. improving their robustness, generalizability, and seamless integration into real-world battery systems remains an ongoing challenge.

Recent research has increasingly focused on hybrid strategies that combine physics-based insights with data-driven algorithms to overcome the limitations of traditional SoH estimation methods. These methods combine the clarity of electrochemical models with the flexibility of machine learning, allowing accurate predictions even when the data is incomplete or affected by noise [14], [15]. Furthermore, advancements in sensor technology and embedded computing have made it possible to deploy these methods directly within battery management systems, reducing reliance on laboratory-grade equipment [16]. Emerging trends, such as federated learning and cloud-connected diagnostics, also promise to enhance model generalization by pooling data from large, diverse battery fleets without compromising privacy [17]. Together, these innovations are paving the way for SoH estimation techniques that are not only more accurate but also scalable and practical for real-world applications.

In response to the ongoing need for accurate and scalable SoH estimation, this study investigates the effectiveness of various machine learning models in predicting battery health based on EIS features. A publicly available dataset—originally introduced by Marco et al. [18]—was used, containing impedance data collected under varying temperatures and state-of-charge conditions. From this dataset, seven key frequency-domain features were extracted to represent the shape and dynamics of the EIS spectrum. Three machine learning models —XGBoost, Random Forest, and GPR— were trained and evaluated using consistent cross-validation settings. Among all models, the Random Forest Regressor delivered the most accurate predictions, with a notable reduction in RMSE compared to the Gaussian Process Regression (GPR) baseline reported in [19].

Building on this result, we turned our attention to making the model reliable under conditions that closely resemble real-world battery operation. To achieve this, we developed a training approach that combines the predictive power of seven EIS-derived features with targeted noise scenarios—Gaussian, uniform, and calibration bias—introduced during learning.

While many previous EIS-based SoH estimation methods are tested under clean laboratory conditions with minimal measurement noise, real-world applications face electrical interference, environmental fluctuations, and sensor drift that can significantly reduce accuracy. Our noise-augmented training, paired with a simple feature-clipping step, enables the model to remain accurate despite such imperfections. By addressing robustness alongside prediction performance, this approach narrows the gap between controlled lab studies and the noisy, variable conditions of practical second-life battery use, offering a dependable and deployment-ready solution for health monitoring.

The remainder of this paper is organized as follows. Section II presents the modeling framework and methodology adopted in this study. Section III describes the SoH estimation process using three different machine learning approaches. Section IV discusses the simulation results, including performance evaluation under noise-perturbed conditions. Finally, Section V summarizes the conclusions and outlines potential directions for future work.

II. Modeling and Methodology

We used a publicly available dataset originally described by Marco et al. [18]. The dataset contains Electrochemical Impedance Spectroscopy (EIS) measurements collected from second-life lithium-ion cells under varying temperatures and states of charge (SoC). In total, the dataset comprises 375 EIS test cases measured across multiple SoC and temperature combinations. Measurements were taken at SoC levels of 5, 20, 50, 70, and 95%, and across a temperature range of 15, 25, and 35 °C. Seven specific features were extracted from each EIS test in order to predict SoH. These features were selected based on Pearson correlation analysis, which demonstrated a significant numerical relationship between each feature and the battery's state of health (SoH). These features correspond to specific geometrical points on the Nyquist plot that reflect the dynamic behavior of the battery under test. As shown in Fig. 1. the selected points include:

– the **highest frequency point** (F_1), which marks the starting point of the EIS curve;
– the **minimum real part** (F_2), typically corresponding to the end of the semicircle;
– the **lowest frequency point** (F_3), indicating the tail of the impedance spectrum;
– the **zero-crossing point** (F_4), where the imaginary component becomes zero;
– the **imaginary peak** (F_5), representing the maximum capacitive response;
– and two **local minima**, (F_6) and (F_7), found between (F_4, F_5) and (F_5, F_3), which characterize curve inflections in the low- and mid-frequency ranges.

These geometrically derived EIS features offer a compact yet informative representation of the impedance response, capturing its shape without relying on explicit physical models. In addition to these seven spectral features, temperature and state of charge (SoC) were incorporated as supplementary inputs, resulting in a total of nine variables for the machine learning models. This approach enables dimensionality reduction while preserving key diagnostic information, facilitating robust and efficient SoH estimation across varying condition.

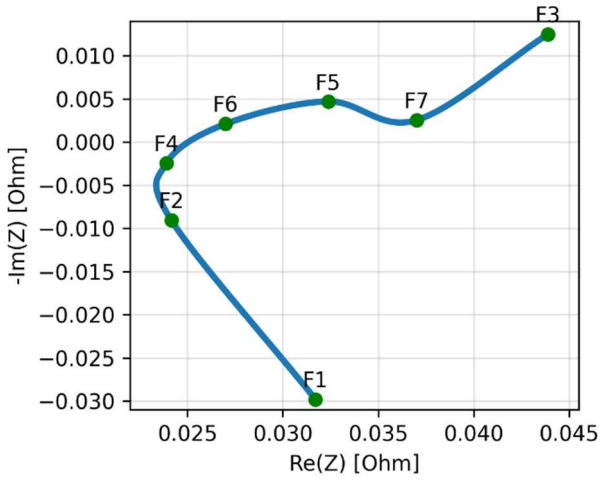


Figure 1. Nyquist Plot with Feature Points F1–F7

III. ROBUST SOH ESTIMATION

Three machine learning models were applied in this work to develop and evaluate the proposed state-of-health estimation:

A. GPR

In the study by Marco et al. [19], Gaussian Process Regression (GPR) was adopted as the primary modeling approach for battery State of Health (SoH) estimation using features extracted from EIS measurements. GPR is a non-parametric, kernel-based method that models the relationship between inputs and outputs by defining a prior over functions and updating it based on observed data. This probabilistic nature allows GPR to not only make point predictions but also quantify uncertainty, which is particularly valuable in applications where measurement noise and operational variability are significant. In Marco's work, the training process involved a combination of spectral feature engineering and hyperparameter optimization to achieve a robust mapping between the impedance-derived inputs and the battery health metric. The model's kernel hyperparameters—such as length-scale, variance, and noise level—were optimized through a cross-validation procedure to minimize prediction error.

B. XGBoost

XGBoost (Extreme Gradient Boosting) is a machine learning algorithm that builds many small decision trees and combines their results to make more accurate predictions. Each new tree is created to correct the mistakes made by the previous ones. The method includes built-in features like regularization to control model complexity, efficient parallel processing, and the ability to handle missing data. These capabilities make it flexible for working with datasets that have nonlinear relationships or moderate levels of noise.

In this work, we configured XGBoost by adjusting a few key parameters. We set the number of trees, the maximum depth for each tree, and the learning rate so the model would improve steadily without overfitting. Regularization parameters were tuned to prevent the model from memorizing noise, and we used random subsampling of both data and features for each tree to make the model more robust. Training was monitored with a validation set, and stopped early when performance stopped improving. These settings were refined through several trial-and-error runs until we found a balance between accuracy and training speed.

C. Random Forest

Random Forest (RF) is an ensemble learning method that averages the predictions of many decision trees trained on bootstrap samples of the data. At each split, trees consider only a random subset of features, which decorrelates the trees and reduces variance. This design makes RF effective at capturing nonlinear relationships while remaining relatively stable in the presence of noisy or partially redundant inputs—properties that align well with impedance-derived feature spaces.

In this study, we flattened and transformed the dataset introduced by Marco et al. [18] into a structured tabular format. Each sample consists of seven impedance-derived features (real and imaginary components at characteristic frequencies), along with the state of charge (SoC), temperature, and the corresponding State of Health (SoH). In total, the dataset comprises 375 samples. Prior to model training, raw measurements were smoothed using a centered rolling mean (window = 21) to reduce high-frequency fluctuations, and any remaining missing values were imputed using backward and forward filling. The processed dataset was then randomly divided into training and testing subsets using an 80/20 split, resulting in 300 training samples and 75 test samples.

To find a balanced configuration, we tested multiple variations of the model's settings—such as the number of trees and the depth of each tree—and selected the combination that provided the most consistent performance across different parts of the dataset. We also measured how quickly the model could produce predictions, ensuring it would be practical for real-time applications.

After training, we assessed the model using standard accuracy metrics and visual inspections. The results showed that its predictions followed the actual SoH values very closely, with small and randomly distributed errors. We also looked at which features the model considered most important, providing insights into the aspects of EIS data that have the greatest influence on battery health estimation. Finally, the trained model and its preprocessing steps were saved, making it ready for deployment or further testing in real-world battery management systems.

Each model was trained using 10-fold cross-validation for consistent and unbiased comparison. The dataset was randomly partitioned into ten equally sized subsets. In each iteration, one subset was held out as the validation set, while the remaining nine subsets were used for training.

IV. EXPERIMENTAL RESULTS

Model performance was evaluated using four standard regression metrics: Root Mean Square Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). RMSE was primarily used as the benchmark metric for ranking models, given its sensitivity to larger errors.

The implementation was carried out primarily in Python using the scikit-learn and XGBoost libraries. All experiments were executed on a standard personal computer without the use of GPU acceleration.

A. Quantitative Comparison

Table I presents a comparative overview of the evaluated models in terms of RMSE, MSE, MAE and R^2. The Random Forest model demonstrated the best overall performance, achieving an RMSE of **0.85%**, significantly outperforming the baseline Gaussian Process Regression (GPR), which recorded an RMSE of **1.14%**.

Table I. Comaprison of ML Models for SoH Prediction

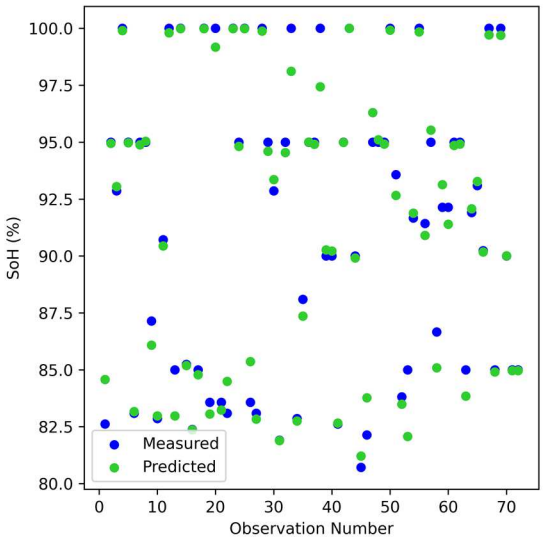
Model	RMSE (%)	R^2 Score	MSE (%)	MAE (%)
Random Forest	0.8441	0.9819	0.7125	0.5117
GPR (Baseline)	1.1397	0.9750	1.6407	0.8573
XGBoost	2.8164	0.8116	7.9322	2.1068

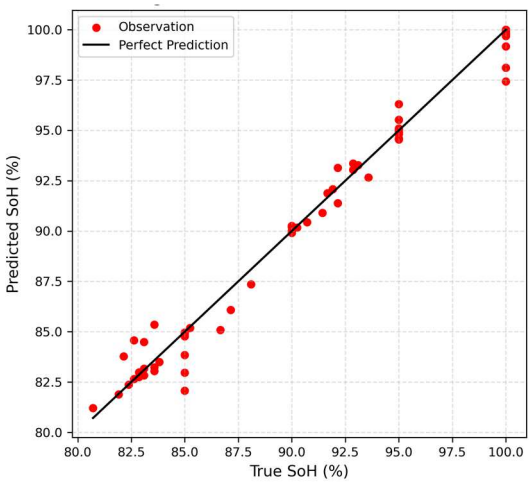
B. Execution Time Analysis

Training the Random Forest model with the selected hyperparameters took approximately **27 seconds**, which is acceptable given the dataset size and cross-validation settings. This was measured on a workstation with Intel 8-core CPU (Intel64 Family 6 Model 142), 7.8 GB RAM, Windows 11, running Python 3.13.1 with scikit-learn 1.6.1. Once trained, the model produced predictions almost instantly, with an average inference time of around **7.5 milliseconds** enabling near real-time prediction capability. These results suggest that Random Forest achieves high predictive performance and can be used effectively even on systems with limited computing power.

C. Visual Evaluation

To visually evaluate the Random Forest model's predictive behavior, three plots are provided. The first plot displays the measured and predicted SoH values across all test observations, showing that the predicted values generally follow the trend of the actual data. The second plot shows predicted values against the true values, along with a diagonal line representing ideal prediction. Most points are located near this line, which reflects a reasonable level of accuracy without major deviations. Additionally, an error-bar plot of the residuals is presented to illustrate the distribution of prediction errors. The residuals remain centered around zero, indicating that the model does not exhibit systematic bias. The error bars represent the standard deviation of residuals within different predicted SoH ranges, showing how prediction errors are dispersed around zero across the operating range.





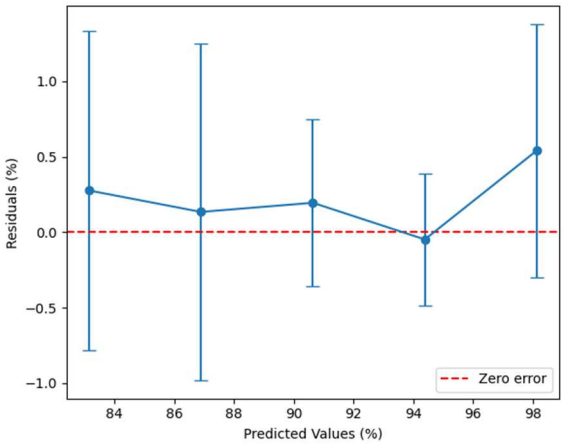


Figure 2. Model evaluation results using the Random Forest model: a) Predicted and measured SoH values across test samples;

b) Scatter plot of predicted versus true SoH with the diagonal line;
c) Error-bar plot of residuals showing the distribution of prediction errors around zero. Vertical axis expressed in %.

D. Noise Robustness Evaluation

To assess real-world applicability, we tested the Random Forest model under different types of artificial noise designed to mimic common sensor and measurement errors in battery management systems (BMS). We applied three kinds of noise to the EIS-based features, with each scaled to match the natural range of that feature in the training data:

a) Gaussian noise – to imitate random fluctuations and electrical interference.

b) Uniform noise – to represent broad, evenly spread measurement errors.

c) Bias shift – a small, consistent offset applied to all measurements, like a sensor calibration drift.

Unlike many studies where noise is only added during testing, here we also introduced these noise patterns during training. This "noise-augmented" approach, combined with a simple clipping step to remove extreme outliers, helped the model learn to handle imperfect data and stay accurate even under challenging conditions. Figure. 3. illustrates the model's performance (RMSE) under each noise scenario. Despite the severity of some perturbations, the error remained below 1.4% in all cases, with the lowest sensitivity observed for bias shift and Gaussian noise, and slightly higher errors under uniform noise. These results show that training the model with added noise makes it much more reliable than models trained only on clean data. This approach is well-suited for real-world battery systems, where sensor readings are often imperfect.

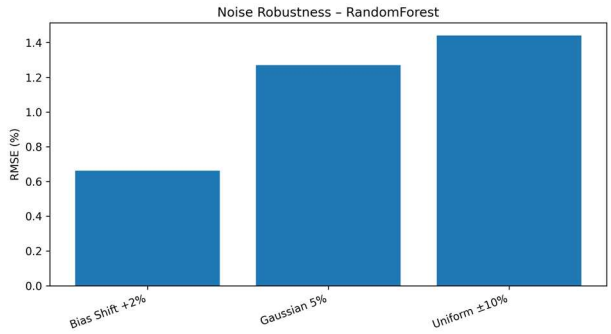


Figure 3. Robustness of the proposed Random Forest model under different noise scenarios applied to EIS-derived features: Gaussian noise, uniform noise, and bias shift

E. Performance Analysis

To further examine the prediction quality of the Random Forest model, a numerical comparison between the actual and predicted SoH values is presented in Table II. The predicted values closely follow the true SoH measurements across the test set, with minimal deviation. This direct side-by-side comparison highlights the model's accuracy and its ability to generalize well without overfitting. The small residuals observed across most samples reinforce the consistency of the model's performance under standard test conditions.

Table II. Actual and Predicted SoH Values

Actual	Predicted	Residual
90	90.01953	-0.01953
82.61905	82.59437	0.02468
95	95.01775	-0.01775
100	100	0
90	89.90917	0.090833
80.71429	81.36577	-0.65148
82.14286	83.96569	-1.82283

Random Forest outperformed GPR for several reasons. First, it naturally handles noisy or incomplete data more effectively by combining the outputs of multiple decision trees and randomly selecting subsets of features. This makes it well suited for EIS measurements taken under non-ideal conditions. Second, it does not depend on choosing a specific mathematical function to describe the data, which in GPR can lead to reduced accuracy if the wrong choice is made. Third, Random Forest can capture complex, nonlinear relationships between the EIS features (F1–F7) and variables like temperature and state of charge (SoC) without requiring manual feature engineering. In our experiments, this led to lower RMSE and MAE and higher R^2 compared to the GPR baseline, in line with the benefits of our noise-augmented training strategy.

V. CONCLUSION

This study evaluated multiple machine learning models for estimating the State of Health (SoH) of second-life lithium-ion batteries using features derived from electrochemical impedance spectroscopy (EIS). Among the tested models—including GPR and XGBoost—Random Forest consistently outperformed all others, achieving the lowest RMSE and MAE, the highest R^2, and showing no signs of overfitting. Visual and residual analyses confirmed its ability to closely track actual SoH values with minimal error. Beyond its strong prediction accuracy, Random Forest also proved efficient and stable, delivering results almost instantly and maintaining performance under varying input conditions. These qualities make it particularly suitable for integration into battery management systems, where both precision and rapid response are critical.

Overall, the combination of EIS-based features with an ensemble learning approach like Random Forest provides a practical, robust, and scalable solution for real-world second-life battery health monitoring. Future work will focus on extending this approach to larger, more diverse datasets and exploring hybrid or physics-informed models to further improve generalization and resilience under real-world operating conditions.

REFERENCES

[1] L. Yao, S. Xu, A. Tang, F. Zhou, J. Hou, Y. Xiao, and Z. Fu, "A review of lithium-ion battery state of health estimation and prediction methods," *World Electric Vehicle Journal*, vol. 12, no. 3, p. 113, 2021.

[2] IDTechEx, *Lithium-ion Batteries for Electric Vehicles 2021–2031*. Cambridge, UK: IDTechEx, Tech. Rep., 2021.

[3] M. H. S. M. Haram, J. W. Lee, G. Ramasamy, E. E. Ngu, S. P. Thiagarajah, and Y. H. Lee, "Feasibility of utilising second life EV batteries: Applications, lifespan, economics, environmental impact, assessment, and challenges," *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4517–4536, 2021.

[4] J.-M. Tarascon and M. Armand, "Issues and challenges facing rechargeable lithium batteries," *Nature*, vol. 414, no. 6861, pp. 359–367, 2001.

[5] B. Dunn, H. Kamath, and J.-M. Tarascon, "Electrical energy storage for the grid: a battery of choices," *Science*, vol. 334, no. 6058, pp. 928–935, 2011.

[6] G. E. Blomgren, "The development and future of lithium ion batteries," *Journal of The Electrochemical Society*, vol. 164, no. 1, p. A5019, 2016.

[7] M. Berecibar, M. Garmendia, I. Gandiaga, J. Crego, and I. Villarreal, "State of health estimation algorithm of LiFePO4 battery packs based on differential voltage curves for battery management system application," *Energy*, vol. 103, pp. 784–796, 2016.

[8] C. Pastor-Fernández, K. Uddin, G. H. Chouchelamane, W. D. Widanage, and J. Marco, "A comparison between electrochemical impedance spectroscopy and incremental capacity-differential voltage as Li-ion diagnostic techniques to identify and quantify the effects of degradation modes within battery management systems," *Journal of Power Sources*, vol. 360, pp. 301–318, 2017.

[9] Y. Xing, E. W. M. Ma, K.-L. Tsui, and M. Pecht, "An ensemble model for predicting the remaining useful performance of lithium-ion batteries," *Microelectronics Reliability*, vol. 53, no. 6, pp. 811–820, 2013.

[10] Y. Zhang, Q. Tang, Y. Zhang, J. Wang, U. Stimming, and A. A. Lee, "Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning," *Nature Communications*, vol. 11, no. 1, art. 1706, 2020.

[11] Y. Li, M. Abdel-Monem, R. Gopalakrishnan, M. Berecibar, E. Nanini-Maury, and N. Omar, *et al.*, "A quick on-line state of health estimation method for Li-ion battery with incremental capacity curves processed by Gaussian filter," *Journal of Power Sources*, vol. 373, pp. 40–53, 2018.

[12] J. Peng, Y. Gao, and J. Jichang, "State of health estimation for lithium-ion batteries using a hybrid EIS based and DRT analysis with a multi scale kernel extreme learning machine," *World Electric Vehicle Journal*, vol. 16, no. 4, art. 224, 2025.

[13] W. Li, J. Chen, K. Quade, D. Luder, J. Gong, and D. U. Sauer, "Battery degradation diagnosis with field data, impedance-based modeling and artificial intelligence," *Energy Storage Materials*, vol. 53, pp. 391–403, 2022.

[14] G. R. Sylvestrin, J. N. Maciel, M. L. M. Amorim, J. P. Carmo, J. A. Afonso, and S. F. Lopes, *et al.*, "State of the art in electric batteries' state-of-health (SoH) estimation with machine learning: A review," *Energies*, vol. 18, no. 3, p. 746, 2025.

[15] M. A. Khan, R. D. P. Pineda, Y. Li, and S. Onori, "Onboard battery state-of-health estimation using electrochemical impedance spectroscopy and deep learning," *Proc. IEEE Conf. on Decision and Control (CDC)*, Milan, Italy, Dec. 2024, pp. 1–8.

[16] Z. Zhang, H. Min, H. Guo, Y. Yu, W. Sun, and J. Jiang, *et al.*, "State of health estimation method for lithium-ion batteries using incremental capacity and long short-term memory network," *Journal of Energy Storage*, vol. 64, p. 107063, 2023.

[17] S. Yang, Z. Zhang, R. Cao, M. Wang, H. Cheng, and L. Zhang, *et al.*, "Implementation for a cloud battery management system based on the CHAIN framework," *Energy and AI*, vol. 5, p. 100088, 2021.

[18] M. Rashid, M. Faraji-Niri, J. Sansom, M. Sheikh, D. Widanage, and J. Marco, "Dataset for rapid state of health estimation of lithium batteries using EIS and machine learning: Training and validation," Data in Brief, vol. 46, art. 108844, Aug. 2023.

[19] M. Faraji-Niri, M. Rashid, J. Sansom, M. Sheikh, D. Widanage, and J. Marco, "Accelerated state of health estimation of second life lithium-ion batteries via electrochemical impedance spectroscopy tests and machine learning techniques," *Journal of Energy Storage*, vol. 58, p. 106295, 2023.