Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 16, No. 2, 2026

Research on State of Health (SOH) of Power Batteries Based on Multi-source Data Fusion and Random Forest Algorithm

Zihao Zhang *

Anhui Institute of Information Technology, Wuhu, Anhui, 241000, China
*** Corresponding author Email:** curry30zzh040820@163.com

Abstract: Against the backdrop of the global energy transition and the rapid development of the new energy vehicle industry, the SOH of power batteries is directly related to vehicle safety, driving range, and service life, and its accurate estimation has become a critical demand of the industry. However, the internal electrochemical degradation process of power batteries is complex and affected by the coupling of multiple factors. Traditional estimation methods relying on a single data source or simple models suffer from drawbacks such as insufficient accuracy and weak generalization ability. This paper proposes an SOH estimation framework based on multi-source data fusion and the random forest algorithm. The framework integrates time-series data (e.g., voltage, current, and temperature) during battery charging and discharging cycles as well as indirect health indicators. Through efficient data fusion and a structured feature engineering process, key features are screened to construct a random forest estimation model. Experimental validation demonstrates that the proposed model achieves high-precision and robust SOH estimation of power batteries, with a significant reduction in the mean absolute error. It provides an effective technical approach for online state estimation in battery management systems (BMS) and also offers a valuable reference for research in the field of condition monitoring and life prediction for complex systems.

Keywords: Power Battery; State of Health (SOH); Multi-source Data Fusion; Random Forest; State Estimation.

1. Introduction

With the global low-carbon energy transition and carbon neutrality initiatives advancing, new energy vehicles (NEVs) have become a key driver of low-carbon transportation, witnessing explosive market growth [1]. As the core energy storage component of NEVs, power battery SOH directly affects vehicle range, safety, and lifecycle economic efficiency, making accurate SOH estimation essential for comprehensive battery management [2,3]. However, SOH estimation faces prominent challenges: battery aging involves complex multi-physical field coupling processes, and traditional single-feature methods fail to fully characterize degradation mechanisms [4]. Under actual operating conditions, battery data exhibits strong nonlinearity and time-variability, demanding higher algorithm generalization [5]. Existing data-driven approaches often suffer from high computational complexity, hindering embedded deployment [6,7].

Multi-source data fusion technology, which integrates multidimensional data to construct a comprehensive health characterization system, provides an effective solution [8]. Among machine learning algorithms, the random forest (RF) algorithm, leveraging ensemble learning mechanisms, excels in processing high-dimensional heterogeneous data, mitigating overfitting, and ensuring interpretability [9,10]. Current SOH estimation research mainly focuses on three paradigms [11]: model-based methods rely on equivalent circuit or electrochemical models for parameter identification to reflect SOH [12,13], but suffer from poor generalization and robustness [14,15]; data-driven methods learn data-SOH mapping without physical models, with typical approaches including support vector regression (SVR) and deep learning [16,17], yet their high complexity limits on-vehicle deployment [18]; hybrid methods combine the advantages of the two, aiming to balance interpretability and fitting capability, but still face generalization and real-time performance bottlenecks [19,20].

Despite progress, breakthroughs are needed in generalization across complex conditions/battery batches, lightweight real-time performance for on-vehicle BMS, and practicality based on real-vehicle data. To address these gaps, this study proposes an innovative SOH estimation framework integrating multi-source data fusion and optimized RF. Technically, robust features are extracted from multi-source data, and Gram angle field transformation enhances feature discriminability. Methodologically, optimized RF improves node splitting and hyperparameters to reduce computation while ensuring accuracy. Experimental validation using public and real-vehicle datasets, compared with traditional machine learning and deep learning methods, is expected to achieve breakthroughs in RMSE and computational efficiency, providing a practical solution for intelligent battery management.

2. Relevant Theories and Technical Foundations

2.1. Definition and Degradation Mechanism of Power Battery SOH

SOH is a core metric for assessing battery performance degradation, remaining service life, and application compliance, with two key definition dimensions: capacity retention and internal resistance growth. Capacity retention, the most widely used quantification method, is the ratio of the battery's current maximum available capacity to its factory-rated capacity, calculated as:

$$SOHc = \frac{Ccurrent}{Crated} \times 100\% \quad (1)$$

where Ccurrent = current maximum available capacity, Crated = factory-rated capacity.

Internal resistance growth, which impacts voltage, heat generation, and power output, is quantified as:

$$SOHR = \frac{REOL-Rcurrent}{REOL-Rnew} \times 100\% \quad (2)$$

where Rnew = initial internal resistance, REOL = end-of-life internal resistance, Rcurrent = current internal resistance.

Battery degradation arises from key physical-chemical mechanisms (active lithium loss, electrode material structural changes, SEI film evolution, etc.), externally manifesting as reduced capacity and increased resistance. Aging follows a non-linear three-stage process (healthy → mild degradation → accelerated degradation) and may directly fail under abnormal abuse, as shown in **Figure 1.** Experimental data shows non-linear capacity decline and monotonic internal resistance growth during cycling, forming the basis for SOH evaluation, as shown in **Figure 2**.

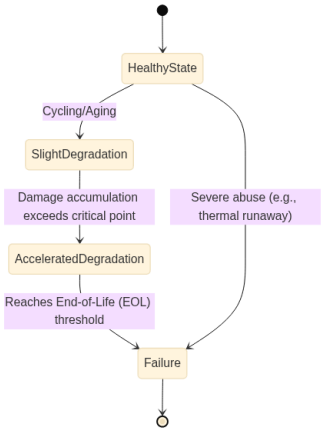

Figure 1. Battery Degradation State Transition Process

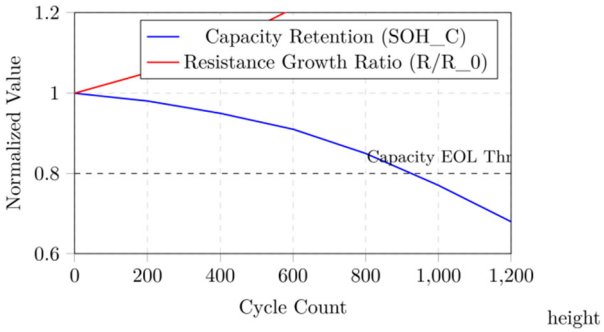

Figure 2. Variation Curves of Capacity and Internal Resistance

2.2. Principle of Random Forest Algorithm

Random Forest is an ensemble learning algorithm centered on decision trees, integrating the Bagging sampling strategy and random feature selection. It generates multiple sub-training sets via Bootstrap self-sampling, trains independent base learners (decision trees) for each set, and combines predictions through averaging (for regression) to reduce variance and improve generalization. During decision tree node splitting, only a random subset of features is considered, ensuring diversity among trees and mitigating overfitting.

For regression tasks (e.g., SOH estimation), CART trees are adopted, with node splitting optimized by minimizing Mean Squared Error (MSE) — the core calculation formula for node impurity is inserted here:

$$MSE(D) = \frac{1}{|D|}\sum i \in D\,(yi - \overline{y}D)^2 \quad (3)$$

where |D|= sample count of node D, yi = target value of sample i, y¯D = mean of target values in node D.

The optimal split point (feature index j, split threshold s) minimizes the weighted sum of MSE for child nodes, with the formula inserted here:

$$min\,j,s[\frac{|DL|}{|D|}MSE(DL) + \frac{|DR|}{|D|}MSE(DR)] \quad (4)$$

where DL= left subnode (satisfying x(j)≤s), DR= right subnode (satisfying x(j)>s).

$$y\text{^}RF = \frac{1}{T}\sum_{t=1}^{T} y\text{^}t \quad (5)$$

where T = total number of decision trees, t = predicted value of the t-th tree.

The algorithm requires no strict data distribution assumptions, excels in handling high-dimensional non-linear data, and supports feature importance evaluation. It is robust to noise, missing values, and outliers — key advantages for adapting to the complexity of multi-source battery data and the non-linearity of SOH degradation, enabling high-precision mapping between input features and SOH values.

3. Multi-source Data Acquisition and Preprocessing

3.1. Data Source and Feature Analysis

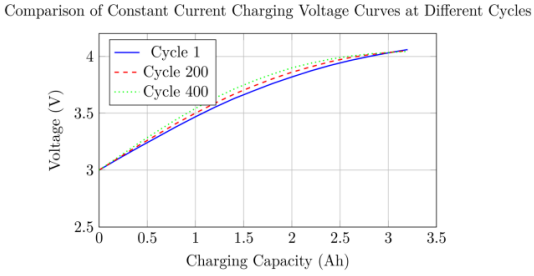

Figure 3. Typical Constant-Current Charging Voltage Curves at Different Cycle Numbers

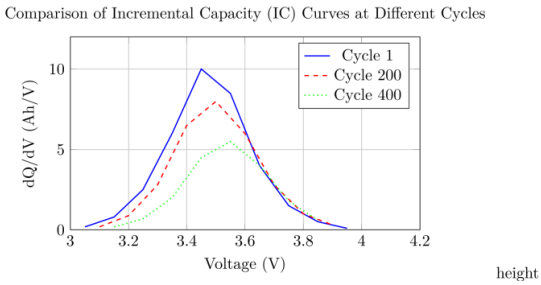

Figure 4. Comparison of Incremental Capacity (IC) Curves at Different Cycle Numbers

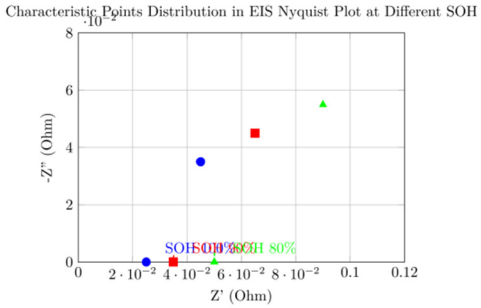

Figure 5. Distribution of Characteristic Points in EIS Nyquist Plots Under Different States of Health (SOH)

To build a robust SOH assessment model, multi-source data was collected from commercial 18650 lithium-ion batteries during accelerated aging cycles, including constant-current charging voltage curves, incremental capacity (IC) curves, temperature sequences, and electrochemical impedance spectroscopy (EIS) data. Typical constant-current charging voltage curves at different cycle numbers are shown in **Figure 3**, incremental capacity curves at different cycles are shown in **Figure 4**, and the distribution of characteristic points in EIS Nyquist plots under different SOH states is shown in **Figure 5**. Key features linked to capacity decay and internal resistance growth are extracted, and their specific categories, names, sources and physical meanings are detailed in **Table 1**.

These features cover electrochemical thermodynamics, kinetics, thermal behavior, and impedance, with clear physical significance for SOH quantification.

Table 1. Battery Features, Data Sources and Their Corresponding Physical Meanings

Feature Category	Feature name	Symbol representation	Source data	Physical meaning
characteristics of voltage curve	Total charging time	T_charge	constant current charging voltage curve	The change of total battery capacity is reflected, and the time prolongation usually means the capacity decay.
	starting voltage of voltage platform	V_start_plat	constant current charging voltage curve	The potential of LiFePO4 two-phase reaction is characterized, and the shift may be related to electrode polarization.
	terminal voltage of voltage platform	V_end_plat	constant current charging voltage curve	The potential characterizing the end of LiFePO4 two-phase reaction.
	duration of voltage platform	Δt_plateau	constant current charging voltage curve	The phase transition reaction volume directly related to the active material, and the shortening implies a reduction in the available active substance.
IC curve characteristics	IC peak voltage	V_peak_IC	Incremental Capacity (IC) Curve	The phase transition equilibrium potential of the electrode material is reflected, and the thermodynamic state change is indicated by the offset.
	IC peak height	IC_max	Incremental Capacity (IC) Curve	The reversible degree of the electrode reaction is positively correlated with the amount of active chlorine, and the decrease is an important sign of aging.
	IC half-peak width	FWHM_IC	Incremental Capacity (IC) Curve	The widening of the reflection spectrum indicates the inhomogeneous aging of the electrode material.
	Area under the IC curve	Area_IC	Incremental Capacity (IC) Curve	Within a specific voltage range, it is related to the total amount of lithium ions involved in the reaction.
temporal sequence characteristics	average charging temperature	T_avg_charge	temporal sequence	The average temperature increase may be attributed to the increase of the internal resistance.
	maximum temperature rise during charging	ΔT_max	temporal sequence	The maximum heat generation per charge is related to the ohmic resistance and polarization resistance.
	rate of temperature rising	dTdt_SOC	temporal sequence	The instantaneous heating power in a specific SOC range can reveal the changes in internal resistance under different SOC conditions.
EIS spectrum signature	ohmic resistance	R_ohm	Electrochemical Impedance Spectroscopy (EIS)	The resistance of electrolyte, separator, electrode material and current collector is the main resistance, which is often increased by aging.
	charge transfer resistance	R_ct	Electrochemical Impedance Spectroscopy (EIS)	The aging mechanisms, such as the characterization of electrochemical reaction kinetics at the electrode/electrolyte interface and SEI growth, contribute to its enlargement.
	characteristic frequency	f_peak	Electrochemical Impedance Spectroscopy (EIS)	The time constant of charge transfer process is related to the change of the interface dynamics.

3.2. Data Preprocessing and Fusion Strategy.

3.2.1. Data Preprocessing

Raw data is preprocessed to ensure quality:

(1) Outlier handling: Identify and remove physically unreliable outliers via boxplots; retain valid abnormal values.

(2) Missing value imputation: Use linear interpolation for small-scale missing data; discard samples with large continuous missing segments.

(3) Normalization: Apply minimum-maximum normalization (Xnorm=Xmax−Xmin/X−Xmin) to map features to [0,1].

The comparison of feature data distribution before and after preprocessing is shown in **Figure 6**.

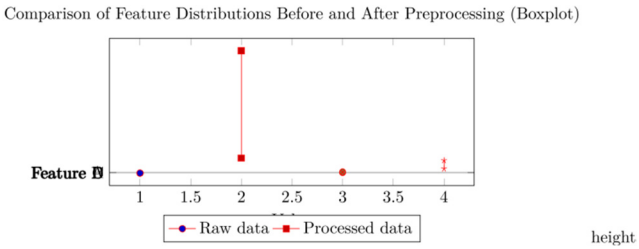

Figure 6. Comparison of Feature Data Distribution Before and After Preprocessing

3.2.2. Data Fusion Strategy

Two core fusion strategies are compared, with feature-level fusion selected for superior performance:

(1)Feature-level fusion: Concatenate multi-source feature vectors; reduce redundancy via Pearson correlation analysis (|r|>0.9) and random forest-based feature importance evaluation.he multi-source data feature correlation network is shown in **Figure 7**.

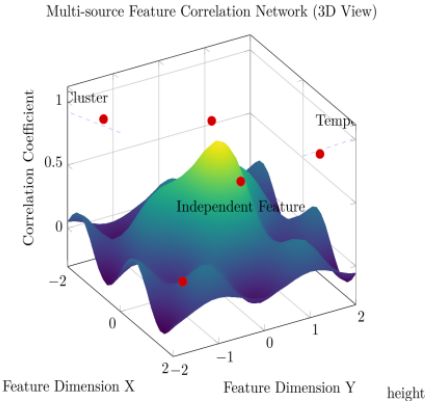

Figure 7. Multi-source Data Feature Correlation Network (3D View)

(2)Decision-level fusion: Train independent sub-models for each data source, integrate predictions via weighted averaging.

Feature-level fusion is preferred for preserving cross-source feature interactions. The flow chart of multi-source data preprocessing and fusion is shown in **Figure 8**.

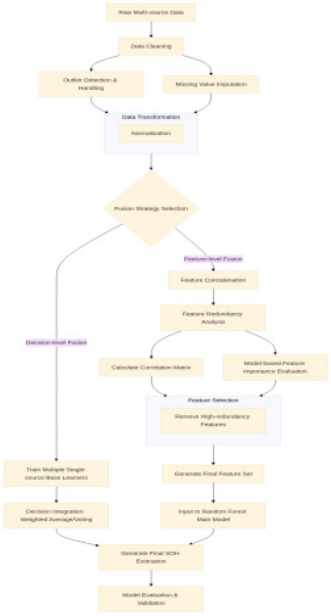

Figure 8. Flow Chart of Multi-source Data Preprocessing and Fusion

4. The Estimation Model of SOH is Constructed based on Random Forest.

4.1. Feature Engineering and Selection

Feature engineering: Construct second-order polynomial features and interaction features

Feature selection: Hybrid strategy

Remove features with Pearson correlation coefficient <0.1 with SOH.

Rank importance via RFE and random forest MDI.

The importance ranking of the top 15 features calculated by the random forest model is shown in **Figure 9**. The variation trend of model prediction accuracy with the increase in feature quantity is shown in **Figure 10**, and the optimal subset (35 features) is determined based on RMSE trends. **Table 2** lists the 10 most representative key features and their physical descriptions.

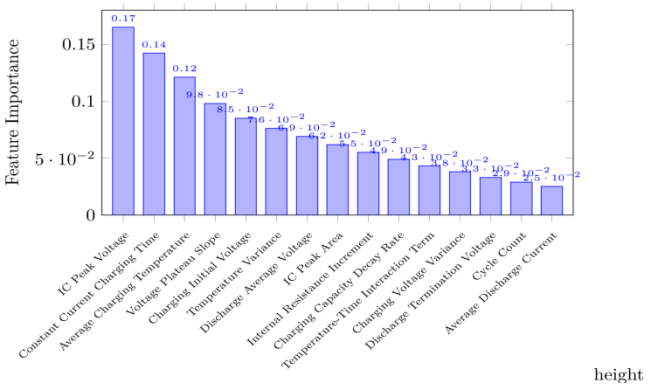

Figure 9. Importance Ranking of the Top 15 Features Calculated by the Random Forest Model

Determine optimal subset (35 features) based on RMSE trends.

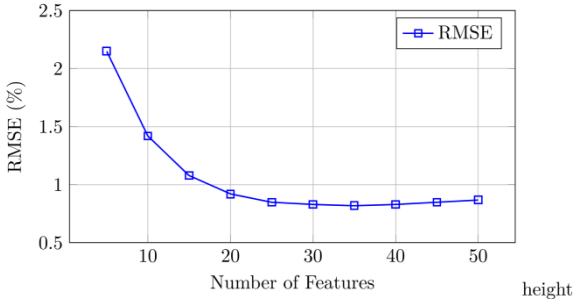

Figure 10. Variation Trend of Model Prediction Accuracy with the Increase in Feature Quantity

4.2. Model Training and Hyperparameter Optimization

Split dataset into training (70%), validation (20%), test (10%) sets. Optimize hyperparameters via grid search + K-fold cross-validation. **Table 3** shows the optimal parameter combination of the random forest model.

The model avoids overfitting, with stable training/validation errors, as shown in **Figure 11**.

Model avoids overfitting, with stable training/validation errors.

Table 2. The 10 Most Representative Key Features and Their Physical Descriptions

Feature Name	feature type	physical description
IC Peak Voltage (dV/dQ Peak Voltage)	ICA Feature	The voltage value of the main peak of the incremental capacity curve, reflecting the electrode phase transition process. Its deviation is strongly related to the loss of active lithium.
Constant Current Charging Time	Time-series Feature	The duration of the complete constant current charging phase, directly related to the increase in battery internal resistance and the decrease in available capacity.
Average Charging Temperature	Statistical Feature (Temperature)	The average temperature during the entire charging process. High temperature accelerates side reactions and is a key environmental factor affecting the aging rate.
Voltage Platform Slope	Morphological Feature	The slope of the voltage change with time during the constant voltage charging phase. Changes in the slope characterize the intensification of electrode polarization.
Initial Charging Voltage	Statistical Feature (Voltage)	The initial voltage at the start of the charging cycle, related to the current state of charge of the battery and the internal ohmic voltage drop.
Temperature Variance	Statistical Feature (Temperature)	The degree of temperature fluctuation during the charging process, reflecting the uniformity of heat generation and heat dissipation inside the battery.
Average Discharging Voltage	Statistical Feature (Voltage)	The average voltage during the complete discharging process. Its decrease is a direct manifestation of battery energy attenuation.
IC Peak Area	ICA Feature	The area enclosed by the main peak of the incremental capacity curve, related to the total amount of active materials participating in the electrochemical reaction.
Internal Resistance Increment	Derived Feature	The rate of change of the estimated ohmic internal resistance based on voltage and current data relative to the initial value.
Temperature-Time Interaction Term	Interaction Feature	The product of the average charging temperature and the constant current charging time, quantifying the cumulative effect of thermal history on the charging dynamic process.

Table 3. Optimal Parameter Combination of the Random Forest Model

Hyperparameter	Search Space	Optimal Value	Description
n_estimators	[10, 30, 50, 100, 150, 200]	150	The number of decision trees in the random forest. The more trees there are, the more stable the model, but the higher the computational cost.
max_depth	[5, 10, 15, 20, 25, None]	20	The maximum depth of a single decision tree. Limiting the depth can prevent overfitting, and "None" means no limit.
min_samples_split	[2, 5, 10]	5	The minimum number of samples required for re-splitting an internal node. A larger value results in a simpler tree, which may lead to underfitting.
min_samples_leaf	[1, 2, 4]	2	The minimum number of samples required for a leaf node. A larger value provides a stronger smoothing effect and prevents overfitting.
max_features	[‘sqrt’, ‘log2’, None]	‘sqrt’	The number of features considered when searching for the best split. 'sqrt' means considering the square root of the total number of features, which is a commonly used default value and performs best in this task.

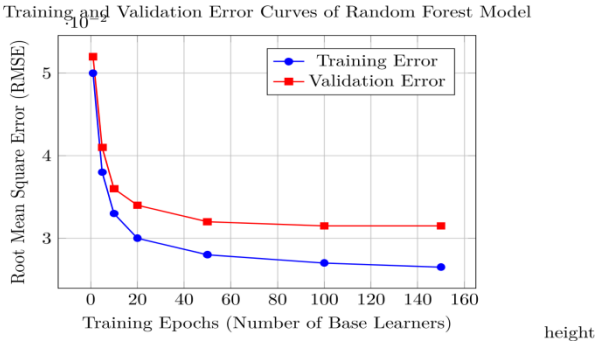

Figure 11. Variation Curves of Training and Validation Errors of the Random Forest Model

5. The Experimental Verification and Result Analysis are Presented

5.1. Experimental Setup and Evaluation Criteria

Use three public datasets (NASA PCoE, Oxford Battery Degradation, CALCE Battery Data) covering 18650 and commercial pouch batteries under various operating conditions. Split dataset into training (70%) and test (30%) sets. **Table 4** provides detailed information of the battery datasets used in the experiment.

5.2. Results Analysis and Comparison

5.2.1. Prediction Accuracy

Model-predicted values closely align with actual SOH, stably tracking aging trends. The comparison between model predictions and actual values is shown in **Figure 12**, and the prediction sample sequence plot is shown in **Figure 13**.

5.2.2. Performance Comparison

Compare with three benchmarks (voltage-only RF, multi-source SVM, multi-source GBT) using RMSE, MAE, R^2. Core evaluation metric formulas are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - y\hat{}i| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - y\hat{}i)^2} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - y\hat{}i)^2}{\sum_{i=1}^{n}(yi - \bar{y})^2} \quad (8)$$

where yi = true SOH value, y^i = predicted SOH value, $\bar{y}$ = average of true SOH values, n = total number of samples.

Performance results:

Proposed model: RMSE=1.21%, MAE=0.89%, R^2=0.984

Table 4. Detailed Information of the Battery Datasets Used in the Experiment

Dataset Name	Battery Type	Data Source	Number of Samples (Battery Cells)	Main Feature Description
NASA PCoEDataset	18650 Lithium-ion Battery ($LiCoO_2$)	NASA Ames Prognostics Center of Excellence	4	Cyclic charge-discharge at different C-rates under constant room temperature, recording voltage, current, temperature, and actual capacity at each cycle.
Oxford Battery Degradation Dataset	Commercial Lithium-ion Pouch Battery	University of Oxford	8	Aging cycles conducted in a controlled temperature chamber, including detailed voltage, current, temperature time series, and capacity measurement values.
CALCE Battery Data	18650 Lithium-ion Battery ($LiCoO_2$)	University of Maryland CALCE Center	4	Cyclic aging data under various operating conditions (different discharge depths, charging rates), containing rich operating condition information.

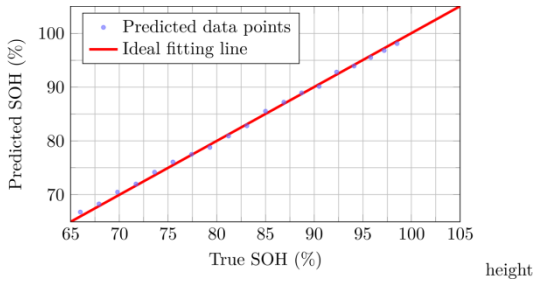

Figure 12. Comparison Between Model Predictions and Actual Values

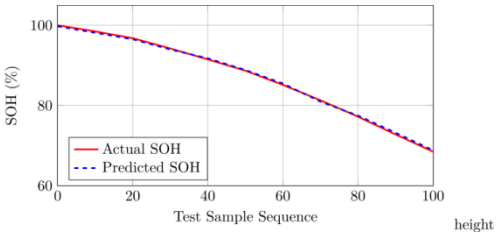

Figure 13. Prediction Sample Sequence Plot

Voltage-only RF: RMSE↑65%, MAE↑78%, R^2=0.942

SVM/GBT: Performance inferior to proposed model

The comparison of evaluation metrics under different models is shown in **Figure 14**.

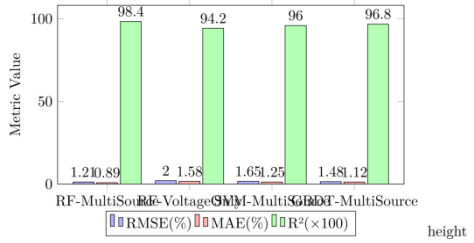

Figure 14. Comparison of Evaluation Metrics Under Different Models

Multi-source data fusion enriches information, and Random Forest's ensemble mechanism effectively handles high-dimensional non-linear data, achieving superior estimation accuracy and robustness.

6. Conclusion

This study proposes a battery SOH estimation framework based on multi-source data fusion and random forest algorithm. By integrating multidimensional data to construct a comprehensive feature set, the framework employs effective feature engineering and screening to establish precise mapping relationships through random forest algorithms. Experimental validation demonstrates that this model achieves high-precision and robust SOH estimation for power batteries, with a significant reduction in average absolute error. The approach provides an effective technical pathway for online state estimation in battery management systems and offers valuable insights for research in complex system condition monitoring and life prediction.

This study has certain limitations. The models performance depends on the quality and representativeness of training data, and its generalization capability across different battery types and complex operational environments requires further validation. Additionally, real-time deployment of the model in embedded battery management systems faces challenges. Future research could explore more advanced fusion and deep learning architectures, introduce transfer learning to enhance model generalization, conduct lightweight model optimization to meet engineering implementation needs, while exploring new data sources and evaluation methods to build a comprehensive battery state management system.

References

[1] Chan, C. C., Liu, C. C., & Chen, Y. P. (2021). A review of state of health estimation for lithium-ion batteries: Methods, metrics, and future trends. *Renewable and Sustainable Energy Reviews*, 145, 111084.

[2] Dubarry, M., Devie, A., & Liaw, B. Y. (2013). Critical review of the methods for monitoring of lithium-ion batteries in electric and hybrid vehicles. *Journal of Power Sources*, 241, 268-284.

[3] Liu, X., Zhang, C., & Mi, C. (2020). Data-driven state of health estimation of lithium-ion batteries: A review. *Energy Storage*, 2(3), 1-27.

[4] Ren, Y., Hu, X., Han, J., Liu, X., & Li, S. (2021). State of health estimation for lithium-ion batteries using machine learning: A review. *Journal of Energy Storage*, 33, 102038.

[5] Li, S., Yang, C., Chen, Z., Wang, H., & Zhang, X. (2022). Multi-source data fusion for state of health estimation of lithium-ion batteries using random forest. *IEEE Transactions on Industrial Informatics*, 18(1), 764-773.

[6] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[7] Plett, G. L. (2004). Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs: Part

3. State of health monitoring. *Journal of Power Sources*, 134(2), 277-292.

[8] Wang, Y., Liu, P., Chen, Z., Li, J., & Zhang, H. (2020). State of health estimation of lithium-ion batteries using a hybrid model combining the equivalent circuit model and the long short-term memory network. *Applied Energy*, 279, 115811.

[9] Dong, G., Zhang, C., He, H., Li, Y., & Wang, Q. (2022). Deep learning-based state of health estimation for lithium-ion batteries: A review. *Energy Reports*, 8, 1199-1217.

[10] Liu, F., Zhang, Y., Li, Y., Wang, J., & Zhao, H. (2020). State of health estimation of lithium-ion batteries using convolutional neural network and long short-term memory. *Journal of Energy Storage*, 32, 101818.

[11] Han, X., Ouyang, M., Lu, L., Li, J., & Zhang, X. (2019). A hybrid method for state of health estimation of lithium-ion batteries using support vector regression and an exponential model. *Applied Energy*, 235, 1056-1065.

[12] Zhang, T., Ding, Y., He, H., Wang, L., & Chen, S. (2021). A hybrid model for state of health estimation of lithium-ion batteries based on particle filter and Gaussian process regression. *IEEE Transactions on Vehicular Technology*, 70(3), 2563-2573.

[13] Chang, F. (2020). Research on data prediction model for state of health of power batteries. *Auto Electric Equipment*, 5, 45-48.

[14] Wang, Z. F., Zhang, S. S., Luo, W., Li, X. H., & Chen, Y. (2021). A review of joint estimation strategies for state of charge and state of health of lithium-ion batteries. *Science Technology and Engineering*, 21(12), 4756-4765.

[15] Yang, S. C., & Zhao, L. (2023). A low-complexity SOH estimation method for power batteries based on linear attention mechanism. *Chinese Journal of Power Sources*, 47(2), 201-205.

[16] Zhang, Z. H., Geng, M. M., Fan, M. S., Liu, Y., & Wang, Q. (2022). State of health assessment of retired power batteries based on relaxation time distribution method. *Energy Storage Science and Technology*, 11(3), 987-994.

[17] Chen, J. M., Tang, W. J., He, S., Li, Q., & Zhang, Y. (2021). SOH estimation of real-vehicle power batteries based on short-time charging segments. *Journal of Guangxi University of Science and Technology*, 32(4), 35-42.

[18] Li, J., Zhang, H., Wang, X., Liu, Y., & Chen, S. (2020). State of health estimation of lithium-ion batteries using random forest and gradient boosting decision tree. *Journal of Power Sources*, 464, 228164.

[19] Wang, H., Li, S., Chen, C., Zhang, X., & Liu, Y. (2021). Multi-feature fusion based state of health estimation for lithium-ion batteries using machine learning algorithms. *Energy*, 221, 119834.

[20] Zhang, Y., Wang, Z., Li, J., Zhao, H., & Chen, S. (2022). State of health estimation for lithium-ion batteries using attention-based bidirectional LSTM network. *Journal of Power Sources*, 542, 231765.