Health monitoring systems for EV battery

Sasmita C S[1] , *Anbuselvi* Mathivanan[2], *Saravanan* Palaniswamy[3] , *Selvam* M[4]

[1]Department of Electronics and Communication Systems, Sri Krishna Arts and Science College, Coimbatore, India
[2]Associate Professor, Department of ECE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
[3]Associate Professor, Department of EEE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
[4] Assistant Professor, Department of Electronics and Communication Systems, Sri Krishna Arts and Science College, Coimbatore, India

Abstract. Electric vehicle (EV) battery health monitoring is crucial to guaranteeing dependability, performance, and safety. A data-driven State of Charge (SOC) prediction framework utilising machine learning models assessed across three distinct battery chemistries—Lithium-ion, Lithium Polymer, and Lead-acid in this paper. Different operating conditions were captured using publicly accessible datasets from Mendeley Data, the CALCE Battery Research Group, and open-source GitHub repositories. Due to its robustness and low computational complexity, a Random Forest Regressor (RFR) was employed as the main SOC estimation model. Its performance was compared with that of a Recurrent Conditional Variational Autoencoder (RC-VAE) to analyse modelling limitations and cross-chemistry generalisation. The Random Forest model is evaluated experimentally using Mean Absolute Error, Root Mean Squared Error, and the coefficient of determination (R^2). The results show that the Random Forest model offers more consistent and dependable SOC predictions, whereas the RC-VAE performs worse under specific datasets and scaling conditions. Additionally, as a proof-of-concept, a voice-activated, lightweight chatbot interface was incorporated to enable users to ask questions about SOC information and get basic charging-related advice through natural language interaction. The suggested method demonstrates how well cross-chemistry SOC estimation can be combined with user-friendly interfaces for useful EV battery monitoring applications.

Keywords: Electric vehicle batteries, State of Charge, Machine learning, Random Forest regressor, Battery monitoring

1. Introduction

Electric vehicles (EVs) are increasingly being adopted as a sustainable alternative to internal combustion engine vehicles, driven by the need to reduce carbon emissions and dependence on fossil fuels. However, the performance and reliability of EVs are heavily dependent on the

condition of the battery, which is the most critical and expensive component. Monitoring the health of EV batteries in real time is essential to ensure operational safety, prolong battery lifespan, and optimise energy efficiency.

One important metric for evaluating battery performance is State of Charge (SOC). Range estimation and charging management are made possible by accurate SOC prediction. Conventional model-based SOC estimation methods, like equivalent circuit models (ECMs), frequently deteriorate under changing load and temperature conditions and necessitate exact parameter identification. As a result, the flexibility and resilience of data-driven machine learning techniques have drawn interest.

In this study, a battery health monitoring system is developed using machine learning models trained on multiple public datasets. The system supports three battery chemistries—Lithium-ion, Lithium Polymer, and Lead-acid—and employs a Random Forest Regressor (RFR) as the primary SOC estimation model due to its stability and low computational complexity. A Recurrent Conditional Variational Autoencoder (RC-VAE) is additionally evaluated as a comparative deep learning model to analyse cross-chemistry generalisation and modelling limitations.

Furthermore, the system is integrated with a lightweight natural language chatbot interface as a proof-of-concept, enabling users to query SOC and basic charging-related information through text-based interaction. The proposed work emphasises system-level integration, cross-chemistry applicability, and user-centric design rather than algorithmic novelty, providing a practical and scalable solution for intelligent EV battery health monitoring.

2. Literature Review

Battery health monitoring in EVs is critical for accurate SOC and SOH estimation, ensuring performance, safety, and longevity. Traditional methods like Coulomb counting, OCV, and Kalman filtering are sensitive to sensor noise and require precise battery models, limiting real-world applicability.

Data-driven approaches using ML algorithms such as SVM, ANN, and Random Forest Regressors (RFR) have shown improved SOC and SOH estimation. For instance, hybrid deep learning models combining DCRNN with SVM feature selection achieved low RMSE on CALCE datasets [1], while CNN-based 3D histogram features with transfer learning enhanced SOH prediction across domains [2].

Generative models have also been explored for data augmentation and health estimation. Conditional GAN-LSTM networks trained on NASA and CALCE data provided accurate predictions from noisy, limited datasets [3], and Recurrent Conditional Variational Autoencoders enabled realistic data generation using the Severson dataset [5].

Adaptive BMS frameworks integrating edge computing, cloud platforms, and digital twins improve efficiency and lifespan but pose scalability challenges [4].

This paper presents a lightweight, chemistry-agnostic SOC prediction framework using RFR and RC-VAE trained on CALCE, Mendeley, and GitHub datasets, with a chatbot interface for intuitive, natural language battery health queries.

3. Methodology

This study presents a hybrid data-driven SOC estimation framework for electric vehicle batteries using Li-ion, LiPo, and Lead-acid datasets. Random Forest Regressor and RC-VAE models are employed to capture both static and temporal battery characteristics. A lightweight

text- and voice-enabled interface is integrated to provide intuitive SOC access and basic charging guidance for practical EV battery monitoring applications.

3.1. Dataset Description

3.1.1. CALCE CS2_8 Li-ion Battery

The CALCE CS2_8 dataset is a publicly available lithium-ion battery dataset from the Centre for Advanced Life Cycle Engineering (CALCE), University of Maryland. It represents a prismatic Li-ion cell with 1100 mAh capacity and a $LiCoO_2$ cathode, widely used for battery diagnostics and performance studies.

Experiments were conducted under controlled laboratory conditions using a CADEX testing system. The battery was charged using a CC–CV protocol at 0.5C to 4.2 V, followed by constant-voltage charging until the current dropped below 0.05 A, and discharged to 2.7 V.

The dataset contains multiple charge–discharge cycles stored as individual .txt files. All discharge cycles were combined for this study, yielding 359,794 time-indexed samples with voltage, current, temperature, time, and capacity. Only discharge data were used for SOC estimation, covering a voltage range of 2.7–4.2 V under near-ambient conditions.

Its high resolution, controlled setup, and extensive prior use make CALCE CS2_8 a reliable benchmark for SOC prediction models.[7]

3.1.2. Mendeley Lithium Polymer (LiPo) Battery

The Lithium Polymer (LiPo) battery dataset was obtained from Mendeley Data (*"LiPo Battery LP-503562-IS-3 EIS, Capacity, ECM Data," Version 1, 2023*), representing a BAK Technology LP-503562-IS-3 LiPo battery. It includes multiple charge–discharge cycles measured under controlled laboratory conditions.

Charging followed a CC–CV protocol at 1 A up to 4.2 V, then constant-voltage charging until the current dropped to 0.2 A. Discharge cycles were at 1 A to 2.75 V, with additional 3 A stress tests for high-load behaviour.

The dataset contains raw cycle files with voltage, current, time, charge, and capacity, along with model-fitted ECM files for ageing analysis. This study used only raw discharge data, calculating SOC from normalised extracted capacity. The subset includes 24 cycles and 1,098 samples across 2.75–4.2 V, providing a robust benchmark for SOC prediction under varying operating conditions.[8]

3.1.3. GitHub Repository Lead-Acid Battery

The Universal 100 Ah lead-acid battery dataset, from a Texas Instruments fuel-gauge experiment, corresponds to a sealed UB121000 battery and captures electrical and thermal behavior during a controlled discharge under near-ambient conditions.

It contains 16,432 time-indexed samples at ~5-second intervals, including voltage, current, and temperature. SOC was computed via coulomb counting and normalized for model training. The data represent a single discharge cycle, suitable for discharge-based SOC estimation. Despite lacking multiple charge cycles, the dataset provides enough samples to evaluate Random Forest and RC-VAE models. Covering 22.7–22.9 V, it serves as a reliable benchmark for lead-acid SOC modelling.[6]

3.2. SOC Prediction Models

3.2.1. Random Forest Regressor (RFR)

The Random Forest Regressor (RFR) is employed as a robust SOC estimation model across multiple battery chemistries (Li-ion, LiPo, Lead-Acid). The RFR combines predictions from multiple decision trees to reduce variance and improve generalisation:

$$y_{SOC=}\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{ft}(\boldsymbol{x}) \quad (1)$$

Here, T denotes the number of trees, $f_t(x)$ the prediction of the tth tree, and x the input features, including voltage, current, capacity, ECM parameters, and derived lag/slope features. All features were normalised to [0,1], leakage-prone variables were excluded, and a 70:30 train–test split with cross-validation was used. The RFR employed 100 trees, a maximum depth of 15, and a minimum of 2 samples per leaf for balanced performance.

3.2.2. Recurrent Conditional Variational Autoencoder (RC-VAE)

The RC-VAE is employed as a comparative model to capture the temporal and nonlinear dynamics of battery behaviour across chemistries. The model encodes input sequences into a latent representation and decodes them to predict SOC:

Encoder:

$$q\phi(z \mid x1{:}T) = N(\mu\phi(x1{:}T), \sigma\phi2(x1{:}T)) \quad (2)$$

Decoder:

$$p\theta(\widehat{SOCi}{:}T \mid z) \quad (3)$$

Training objective:

$$L_{RC-V_AE} = \frac{1}{N}\sum_{i=1}^{n}(\widehat{SOCi} - SOCi)2 + \beta DKL(q\phi(z \mid xi) \parallel p(z)) \quad (4)$$

where the first term represents reconstruction loss, and the second term regularises the latent space via KL divergence.

Preprocessing corrections, including SOC normalization to [0,1] and standardisation of ECM parameters, significantly improved stability and prediction accuracy. The RC-VAE, while still underperforming RFR in direct SOC prediction, effectively learns meaningful battery health representations and can be used for cross-chemistry generalisation studies.

3.2.3. Dataset Sizes and Balance Across Chemistries

The study evaluated SOC prediction using datasets from three battery chemistries. While Li-ion and LiPo datasets included multiple cycles, the lead-acid dataset had only a single discharge cycle, limiting comprehensive evaluation. However, consistent preprocessing, including feature normalization, leakage prevention, and uniform SOC computation, ensured fair cross-chemistry comparison.

Table 1. *Dataset Sizes and Balance Across Chemistries*.

Chemistry	Cycles used	Total samples	Notes
Li-ion	35	359,794	CALCE dataset
LiPo	24	1098	Mendeley dataset
Lead-Acid	1	16,432	GitHub dataset

3.3. Data Preprocessing

3.3.1.Data Quality and Cleaning

The first step in preprocessing began with inspecting all datasets for missing, corrupted, or inconsistent measurements. Anomalies in voltage, current, and capacity—such as spikes, negative values, or empty entries—were corrected via interpolation or removed, ensuring high-quality data and preventing error propagation into the SOC prediction models for stable, reliable learning.

3.3.2. Feature Normalization

All input features were normalised to [0,1] to remove scale differences among voltage, current, capacity, and ECM parameters. This ensures numerical stability, prevents large-magnitude features from dominating learning, and maintains consistent performance across battery chemistries, improving the predictive accuracy of RFR and RC-VAE models.

3.3.3. Exclusion of Cumulative Features

Cumulative features like total capacity were excluded to prevent data leakage, ensuring the model predicts SOC using only real-time information. This avoids artificially high accuracy and ensures realistic, reliable battery monitoring.

3.3.4. Derived Features and Temporal Structuring

Additional features were derived from raw data, including voltage and current slopes, 1-step lag values, and rolling-window sequences for temporal modelling. These capture battery dynamics, enabling RC-VAE to learn sequential dependencies and RFR to leverage richer inputs, improving SOC prediction accuracy.

3.3.5. SOC Computation

SOC values, which serve as the target for model training, were computed from discharge curves using the formula:

$$SOC(t)=\frac{\boldsymbol{Qextracted(t)}}{\boldsymbol{Qfull}}\times 100 \quad (5)$$

This provides a consistent and physically meaningful reference across all battery chemistries. By standardising SOC calculation, the models can be trained and evaluated fairly, even when datasets differ in scale, cycle length, or sampling frequency.

3.3.6. Train-Test Splitting and Cross-Validation

To prevent overfitting, datasets were split 70:30 for training and testing, with cross-validation applied to validate generalization. This approach addresses dataset size imbalances (e.g., Li-ion vs. Lead-acid) and ensures models are evaluated on unseen data representative of real battery operation.

4. Results and Discussion

4.1. Evaluation Metrics

SOC prediction performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2), providing complementary measures of accuracy, error magnitude, and model fit.

1. Mean Absolute Error (MAE):

MAE measures the average absolute difference between the predicted SOC ($\hat{y}_i$) and the actual SOC (y_i) across all samples, providing an interpretable estimate of average prediction error:

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

2. Root Mean Squared Error (RMSE):

RMSE penalises larger errors more heavily than MAE and reflects the standard deviation of prediction errors. It is computed as:

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

3. Coefficient of Determination (R^2):

R^2 quantifies how well the model predictions fit the observed data, with a value of 1 indicating perfect prediction. It is defined as:

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

Where $\bar{y}$is the mean of the actual SOC values. An R^2 value close to 1 indicates that most of the variance in SOC is captured by the model, while negative values indicate poor predictive performance.

These three metrics were applied consistently to both RFR and RC-VAE models across all battery chemistries, enabling fair and comparable evaluation of SOC prediction performance.

4.2. Quantitative Performance

The quantitative performance of the Random Forest Regressor (RFR) and Recurrent Conditional Variational Autoencoder (RC-VAE) was evaluated across LiPo, Lead-Acid, and Li-ion battery chemistries using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The comparative results for each chemistry are summarised in Tables 2–4.

Table 2. *LiPo Battery (Mendeley Dataset) Performance.*

Model	MAE	RMSE	R^2
RFR	0.0171	0.0965	0.8984
RC-VAE	0.2359	0.2750	0.1753

In Table 2, RFR achieves low errors and high R^2, demonstrating strong agreement with actual SOC even with only 24 cycles, highlighting its ability to learn nonlinear feature–SOC relationships. RC-VAE showed higher errors and low R^2, indicating that limited temporal data hindered stable latent learning.

Table 3. *Lead-Acid Battery (GitHub Dataset) Performance.*

Model	MAE	RMSE	R^2
RFR	0.0142	0.3614	0.9988

RC-VAE	0.663	4.9039	0.7706

Table 3 shows Lead-Acid battery results. RFR achieved near-perfect predictions despite only a single discharge cycle, thanks to the battery's simple discharge behaviour and strong voltage–current–SOC correlation. RC-VAE had higher errors but a reasonably high R^2, capturing overall SOC trends while struggling with precise point-wise estimation due to limited temporal data.

Table 4. *Li-ion Battery (CALCE Dataset) Performance.*

Model	MAE	RMSE	R^2
RFR	4.9673	11.1334	0.8367
RC-VAE	0.2117	0.3329	0.6108

Table 4 presents Li-ion battery results from the CALCE dataset. RFR achieved an R^2 of 0.8367 but higher MAE and RMSE, reflecting sensitivity to inter-cycle variability. RC-VAE showed lower MAE and RMSE by leveraging temporal dependencies, yet its lower R^2 indicates limited ability to capture long-term SOC variance across multiple cycles.

4.3. Visual Validation

Visual validation was performed to qualitatively assess the agreement between actual and predicted SOC values from RFR and RC-VAE models. Predicted-versus-actual plots and time-series comparisons are commonly used in SOC studies to confirm model reliability beyond numerical metrics.

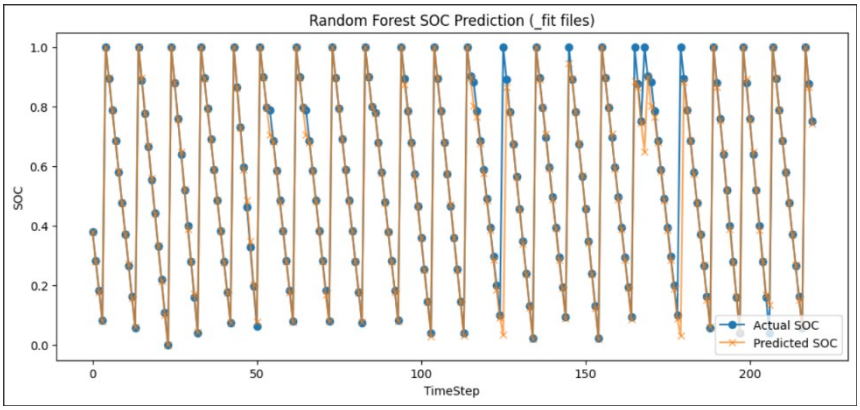

Figure 1. Random Forest SOC Prediction for Mendeley Dataset

For the Mendeley LiPo dataset, Figure 1 shows the Random Forest SOC predictions closely following the actual SOC across cycles. Only minor deviations occur at isolated points, demonstrating stable learning and strong generalisation. The low errors and high R^2 further confirm the model's accuracy and its suitability for this dataset.

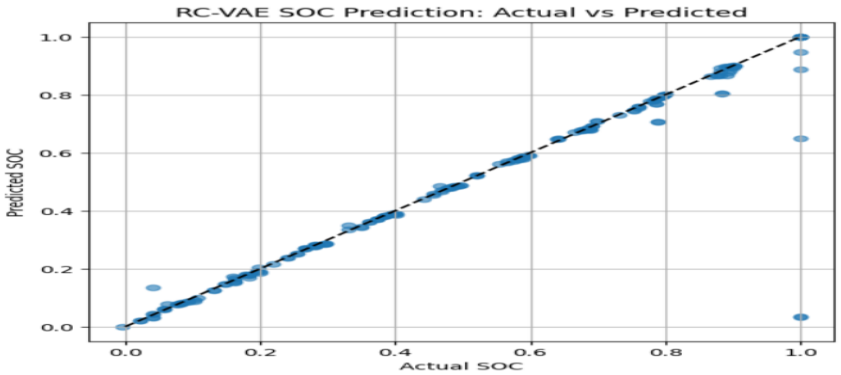

Figure 2. RC-VAE SOC Prediction for Mendeley Dataset

Figure 2 shows RC-VAE SOC predictions exhibiting lag and smoothing, especially during dynamic regions, reflecting high errors and low R^2. This indicates limited learning of SOC dynamics, a common issue with autoencoder-based models under rapidly changing conditions.

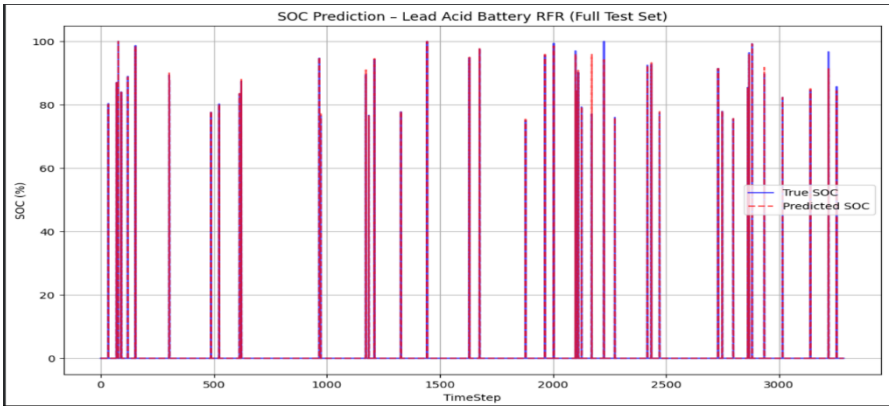

Figure 3. RFR SOC Prediction for Lead Acid Dataset

Figure 3 shows Random Forest SOC predictions for the Lead-Acid battery closely matching actual SOC throughout the test sequence, including abrupt changes. The alignment confirms very low MAE and RMSE and a near-unity R^2. This highlights the robustness of ensemble tree methods for nonlinear, noisy battery data.

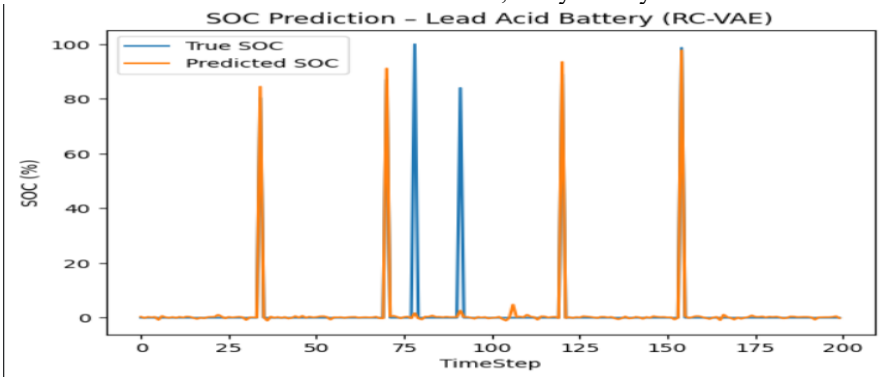

Figure 4. RCVAE SOC Prediction for Lead Acid Dataset

Figure 4 shows RC-VAE SOC predictions for the Lead-Acid battery, displaying fluctuations and spikes that deviate from the true SOC, especially during rapid transitions. This visual mismatch aligns with high MAE and RMSE, indicating poor temporal learning and limited generalisation for this battery chemistry.

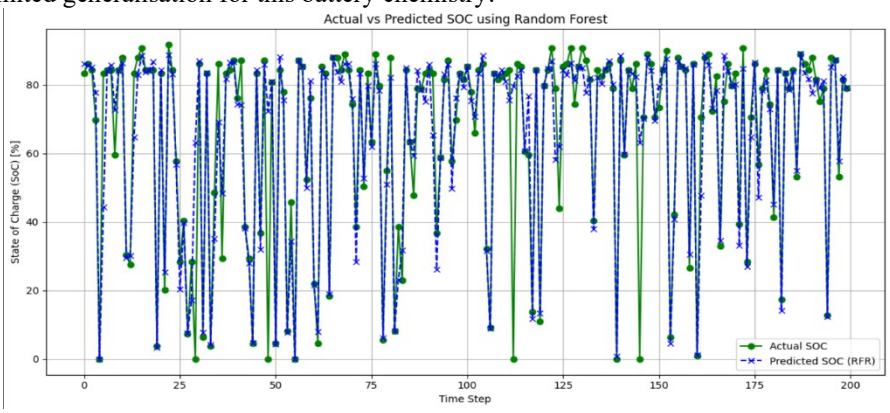

Figure 5. RFR SOC Prediction for CALCE Dataset

Figure 14 shows Random Forest SOC predictions for the CALCE dataset closely matching actual SOC across varying conditions. Minor discrepancies appear during sharp transitions, but overall trends are well preserved. This confirms reliable performance and supports Random Forest as a strong baseline for SOC estimation across battery chemistries.

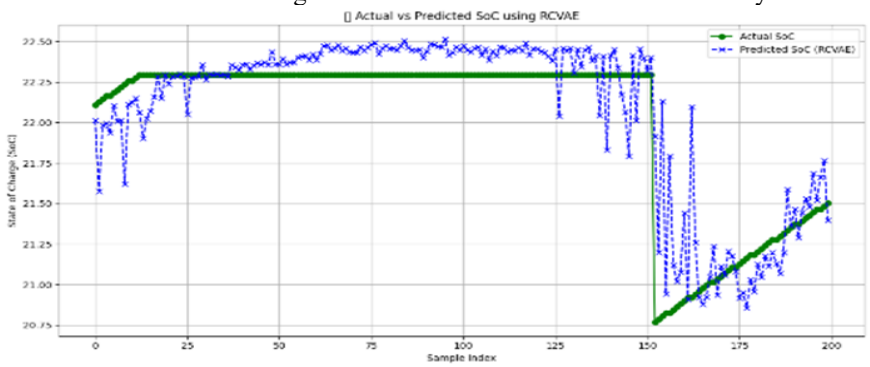

Figure 6. RCVAE SOC Prediction for CALCE Dataset

Figure 13 presents the RC-VAE results for the CALCE Li-ion dataset, where moderate alignment between actual and predicted SOC is observed. While the overall trend is captured, deviations become more evident at higher SOC ranges. This visual observation is consistent with the moderate R^2 score and error values, indicating partial but incomplete modelling of Li-ion SOC behaviour.

5. Conclusion

The conclusions presented in this study provide a comprehensive summary of the findings, highlighting the key advantages of the Random Forest Regressor (RFR) for SOC prediction across different battery chemistries. RFR consistently demonstrated superior predictive accuracy, low error metrics, and robust generalisation across Li-ion, LiPo, and lead-acid batteries. This highlights the suitability of ensemble-based, deterministic models for structured

datasets, especially when the dataset includes limited discharge cycles or exhibits relatively linear behaviour, as in the case of lead-acid batteries.

However, the initial experiments revealed extremely large errors for lead-acid batteries when using RC-VAE, primarily due to inconsistencies in SOC scaling and feature normalization. These discrepancies were addressed by normalising SOC targets to the range [0,1] and standardising input features using z-score normalization, which stabilised model training and produced meaningful predictions. After this correction, all performance metrics were re-evaluated, confirming the reliability of the reported results.

The validated findings demonstrate that while RC-VAE provides probabilistic modelling advantages for complex, multi-cycle chemistries like Li-ion and LiPo, RFR remains the most reliable choice for datasets with limited cycles or linear characteristics. Overall, the conclusions not only summarise the technical achievements and comparative evaluation of SOC prediction models but also highlight the generalisation, accuracy, and practical applicability of the system, providing a strong foundation for future research and real-world deployment in battery health monitoring.

6. Acknowledgement

The authors would like to express their sincere gratitude to the Department of Electronics and Communication Engineering at Sivasubramaniya Nadar College of Engineering for providing the necessary facilities and academic support throughout the project. The authors also acknowledge the availability of open-source datasets from CALCE, Mendeley Data, and GitHub, which were instrumental in training and evaluating the machine learning models used in this study.

7. Reference

1. L. Zhang, J. Liu, and S. Wang, “State of health prediction in electric vehicle batteries using a deep learning model,” World Electric Vehicle Journal, vol. 15, no. 385, pp. 1–23, 2024. https://doi.org/10.3390/wevj15090385
2. M. Chen, Y. Li, and T. Zhao, “A general framework for lithium-ion battery state of health estimation: From laboratory tests to machine learning with transferability across domains,” Applied Energy, vol. 381, art. no. 125086, 2025.
3. A. Kumar, S. Singh, and H. Lee, “Generative adversarial network for state of health estimation of lithium-ion batteries,” in Proc. IEEE Int. Conf. Prognostics and Health Management (ICPHM), 2023, pp. 1–6.
4. N. Gupta and P. Sinha, “Adaptive battery management systems for the new generation of electric vehicles,” in Artificial Intelligence for Smart Battery Management Systems, Springer, 2023, ch. 4, pp. 67–84.
5. Y. Wang, M. Fang, and X. Zhou, “Generating comprehensive lithium battery charging data with generative AI,” Applied Energy, vol. 377, art. no. 124604, 2025. https://doi.org/10.1016/j.apenergy.2024.124604
6. Universal 100 Ah battery data (Lead-Acid). Available: http://www.mrsolar.com/content/pdf/Universal/UB121000.pdf
7. CALCE CS2_8 lithium-ion battery dataset. Available: https://calce.umd.edu/batteries/data/
8. LiPo Battery LP-503562-IS-3 dataset (Mendeley Data). Available: https://data.mendeley.com/datasets/stcppt2r68/1