Machine Learning-Based Estimation Of Battery State Of Charge And State Of Health

1Skand Jeet, 2Rupali Mandawat, 3Sonam Khampa, 4Deva Nand
1,2,3,4Department of Electronics and Communication Engineering
Delhi Technological University, Delhi, India
1skandjeet_ec22a18_41@dtu.ac.in

Abstract — Precise battery State of Charge (SOC) and State of Health (SOH) estimation are critical to ensure aircraft safe, economic and dependable operation when embedded electronic devices, energy storage systems and electric vehicles. Traditional voltage- and current-based SOC/ SOH monitoring methods are often plagued by error build-up, model mismatch, and higher sensitivity to different operating environments. In this paper, an integrated machine learning pipeline for joint SOC and SOH estimation using two separate numerical data sets and multiple regressor algorithms is proposed. It covers EDA, normalization (StandardScaler), DataTrain/DataTest split (80/20), and four regression metrics; coefficient of determination (R2), root mean square error, mean absolute error and mean absolute percentage error. The best stand-alone estimator for SOC prediction is Random Forest Regression with an average test R2 of 0.9137, RMSE of 6.16, MAE of 2.55, and MAPE of 6.67%. For SOH prediction the Extra Trees Regression reaches near-perfect R2 of 0.9955, RMSE 1.24, MAE 0.96, and MAPE of 1.75%. Its feature contribution analysis shows that route distance, longitude and altimetry features dominate SOC prediction, whereas internal resistance and capacity dominate SOH prediction, both in line with physical knowledge of battery performance. The testing models are used in a web app based on Streamlit to provide real-time, user-friendly battery management without domain expertise.

Index Terms — Battery management systems, State of Charge, State of Health, Machine learning, Random Forest, Extra Trees.

I. INTRODUCTION

Battery management systems (BMS) are critical components in applications such as electric vehicles (EVs), renewable energy storage, and portable devices, in which the battery type (such as lithium-ion battery) will directly affect performance, reliability and safety [2]. Accurate estimation of the internal state (SOC and SOH in particular) of the battery is essential to ensure proper functioning of BMS [3, 7].

State of Charge (SOC) is defined as the remaining usable energy expressed as a fraction of the nominal capacity while State of Health (SOH) is defined as the ratio of the actual capacity to the initial fresh cell capacity [7]. Both SOC or SOH estimation errors may cause the battery to be operated under overly conservative conditions or lead to the risk of unexpected shutdown or safety hazards in high-power demands [2].

Simple SOC estimation methods such as coulomb counting and OCV methods are easy to implement but they are sensitive to current sensor drifts, initial state and temperature [1]. SOH estimations like periodic capacity test or measurement of internal resistance are intrusive and can not be directly implemented in online estimation [3]. Model-based SOC/SOH estimations involving equivalent circuit or electrochemical models demand thorough parameter identification and are generally not robust for different chemistry and aging states [6].

Another approach relies on the data-driven method where the estimation of SOC and SOH comes directly from measured operating conditions data (e.g. Voltage, current, temperature, driving profile) [2,5]. ML models can learn a complicated nonlinear relationship between input and output without the use of a physical equation and can also be updated using new data [4, 6]. Several ML algorithms have been employed to estimate SOC and SOH (e.g. SVM, Neural Network, ensemble learning) and most of them show higher accuracy than traditional methods [5, 6].

The issue is that most of the existing studies separate the SOC and SOH estimation problems and have not come up with a unified and deployable system integrated by both SOC and SOH estimation [7]. Here, we bridge the gap and introduce an end-to-end ML-based pipeline that predicts SOC and SOH based on specific datasets and models, conducts formal model comparison, and allows real-time deployment using a web interface. The main contributions are:

- Two models to specifically perform the regression on SOC and SOH using real data sets with mixed feature information.
- Comparing several regressors (Linear Regression, Decision Tree, Random Forest, Extra Trees, XGBoost, CatBoost, LightGBM, and Linear SVR) in a systematic way on the metrics R2, RMSE, MAE, and MAPE.
- Feature-importance analysis—finding physically meaningful features such as distance travelled, longitude, altitude, internal resistance, capacity etc.
- Final integration and end-to-end deployment of the top models (Random Forest for SOC, Extra Trees for SOH) in a Streamlit web application for live prediction.

The remainder of this paper is structured as follows. It provides an overview of SOC and SOH and ML-based estimation approaches. Introduces the datasets used in the study and an exploratory data analysis. Describes the proposed approach in detail. Demonstrates the experimental findings and discusses them. Finally, it concludes this paper and suggests future directions.

II. Background and Related Work

A. SOC and SOH Concepts

The state of charge (SOC) is generally represented by a percentage, showing the difference between the remaining capacity and the nominal capacity [1]. It has many benefits such as range estimation, avoiding overcharging and deep discharge as well as aiding energy management in hybrid electric vehicles [1, 5]. State of health (SOH) expresses the general health status and the ageing degree of a battery relative to a fresh battery, often shown by the current maximum capacity over rated capacity ratio, or a function of internal resistance [3, 6]. The SOH estimation will help maintain scheduling, warranty management and security evaluation [2]. In the real world SOC and SOH depend upon temperature, charging algorithm, discharging rate, depth of discharge and calendar aging etc. [1]. SOC and SOH behavior is non-linear and depends on history which makes the analytical modeling extremely difficult [4].

B. Machine Learning for SOC Estimation

Data-driven SOC estimation has garnered a lot of interest, providing a model-free approach to estimate SOC [5, 7]. A variety of regression algorithms have been trained on wind- and non-wind weather condition time-series voltage, current, and temperature, vehicle speed, and other environmental variables [5, 8]. Nonlinear ensemble learners such as Random Forests and gradient-boosted trees tend to outperform linear models on this task [5]. In addition, feature selection and dimensionality reduction can be used to further improve accuracy and robustness [5].

C. Machine Learning for SOH Estimation

Most methods used for SOH estimation with ML assume a mapping from features indicative of degradation (internal resistance, capacity fade, impedance, incremental capacity curves) to health indicators [3, 6]. According to [3, 6], Random Forest, Extra Trees, and gradient boosting are known to be particularly accurate and robust when training data are noisy or high-dimensional. Deep learning methods such as convolutional and recurrent neural networks have also been used for SOH estimation that are not directly associated with features of degradation but these require larger training data sizes and increased computational resources [6]. Recent reviews highlight that ensemble methods offer a favorable trade-off between accuracy, interpretability, and computational cost, making them well suited for deployment in real-time battery management systems [6].

III. Datasets and Exploratory Analysis

Two independent numerical datasets are used: one for SOC estimation and one for SOH estimation. Both are purely tabular, enabling straightforward application of standard ML regressors.

A. SOC Dataset

The SOC dataset contains 1,000 samples with 18 columns: 17 input features and 1 SOC target. The input features are: Source Latitude and Longitude, Destination Latitude and Longitude, Vehicle speed (km/h), Travel distance (km), Altimetry 1 to Altimetry 10 (terrain elevation descriptors), Travel time (minutes). All the characteristics are numerical: speed and time are integers whereas others are floats. The SOC range covers 20%-93%, taking in different states and loads. An initial EDA reveals a broad spectrum of distances traveled, from approximately 70km to 760km, and an equally wide spread of altimetry features indicating various terrain profiles. The values of SOC approximate to a normal distribution with data tending towards the middle to high charge region. Correlation results reveal that both distance travelled, source longitude, and some altimetry features are correlated with SOC.

B. SOH Dataset

The SOH dataset contains 2,000 samples with 9 input features and 1 SOH target, all of type float. The features are Voltage, Current, Temperature, Charge Time, Discharge Time, Internal Resistance, Capacity, Ambient Humidity, and C-Rate. The SOH values of the batteries range from about 27% (very degraded) to 100% (mostly new). The EDA indicate that the voltage is within 3.0 V and 4.2 V, which matches that of lithium-ion batteries, and aging effects on internal resistance and capacity are enhanced with aging. The correlation analysis demonstrated that capacity and internal resistance are the most correlated parameters with SOH, followed by temperature and humidity.

C. Data Quality and Scaling

Both data-sets were inspected for missing values and outliers, and proven to be de-noised using ordinary data-frame operations. No rows were deleted; highs were saved as genuine outliers in the extremes of operating ranges. As the features have widely-varying scales (seconds, km, volts), the StandardScaler was used to normalize all inputs to a Gaussian distribution with mean 0 and standard deviation 1. Each data-set was randomly partitioned into 80% training and 20% validation subsets with the same distribution as the whole.

IV. Methodology

A. Problem Formulation

Both SOC and SOH estimation tasks are formulated as supervised regression problems. Given an input feature vector $x \in Rd$, the model learns a mapping $f(x) \rightarrow y$, where y is either SOC (%) or SOH (%). The goal is to minimize prediction error on unseen test data while maintaining good generalization.

B. Model Candidates

The following regression algorithms were considered for each task: Linear Regression, Decision Tree Regressor, Random Forest Regressor, Extra Trees Regressor, XGBoost Regressor, CatBoost Regressor, LightGBM Regressor, and Linear Support Vector Regressor (SVR). These models span from simple linear relationships to powerful ensemble learners capable of capturing complex nonlinear patterns [5, 6]. Hyperparameters were tuned using grid search and cross-validation at a moderate level to balance accuracy and training time.

C. Training and Evaluation

For each dataset, models were trained on the scaled 80% training subset and evaluated on the remaining 20% test subset. Four metrics were used: R2: proportion of variance explained; RMSE: intuitive error magnitude; MAE: average absolute prediction error; MAPE: error as a percentage of the true value. Higher R2 and lower RMSE/MAE/MAPE indicate better performance.

D. Feature-Importance Analysis

To interpret the learned models and link them to physical mechanisms, feature-importance analysis was carried out using XGBoost and the final ensemble models [6]. The relative importance scores indicate how much each feature contributes to reducing prediction error across all trees, enabling ranking of key drivers of SOC and SOH.

E. Deployment via Web Interface

The best-performing models for SOC and SOH were serialized along with their preprocessing steps using joblib and integrated into a Streamlit application. The web interface includes: User authentication with an SQLite-based login system, separate tabs for SOC and SOH prediction, intuitive input forms for route and vehicle parameters (SOC) and battery health parameters (SOH), and visualization panels showing feature importance and model comparison metrics.

F. Comparison with Existing Work

Table 1 provides a quantitative comparison of the proposed framework against recent peer-reviewed works (2023–2026) that report comparable numerical performance metrics for SOC and/or SOH estimation. The selected baselines represent state-of-the-art methods including optimized deep learning hybrids, ensemble approaches, and traditional ML regressors evaluated on similar lithium-ion battery datasets.

TABLE I. Quantitative Comparison with Recent SOC/SOH Estimation Methods

Method	Year	Task	Test R2	RMSE	MAE	MAPE (%)
FOAMIUHF-UKF	2025	SOC+SOH	–	1.00	0.99	–
SSA-LSTM	2025	SOH	–	0.73	–	0.53
Random Forest	2025	SOC	–	0.0229	0.0139	–
Ensemble Voting	2026	SOH	High*	Low*	Low*	–
Hybrid ELM-RVFL	2023	SOC	0.91	–	–	–
Nature ML	2024	SOC	–	Low*	Low*	–
Proposed RF	2026	SOC	0.9137	6.16	2.55	6.67
Proposed ETrees	2026	SOH	0.9955	1.24	0.96	1.75

*Qualitative descriptors from papers; – = not reported

This model-derived framework provides competitive quantitative results and four new contributions to the literature:

1) Dual-task unified framework: Unlike the single task approach, our single framework simultaneously estimates SOC and SOH from heterogeneous data streams (route battery SOC, health parameters SOH).

2) User-Facing Web Deployment: None of the previous efforts have produced a user-facing, production-ready application. The Streamlit application we created is more than a simple prototype.

3) Better interpretability & physically validated: Slight increases in accuracy for deep/hybrid models are compromised by the significant loss of interpretability. Our tree ensembles naturally present a physically reasonable feature ranking.

4) Edge-Deployable Efficiency: SOH R2=0.9955, MAPE=1.75% which is similar to the 0.53% MAPE of the SSA-LSTM method. We use low complexity Random Forest/Extra trees to run models in real-time.

V. Results and Discussion

A. SOC Estimation Results

Among the evaluated models, ensemble methods clearly outperform linear baselines on the SOC dataset. Fig. 1 and Fig. 2 show the regression evaluation plots for the XGBoost and Extra Trees regressors, respectively.

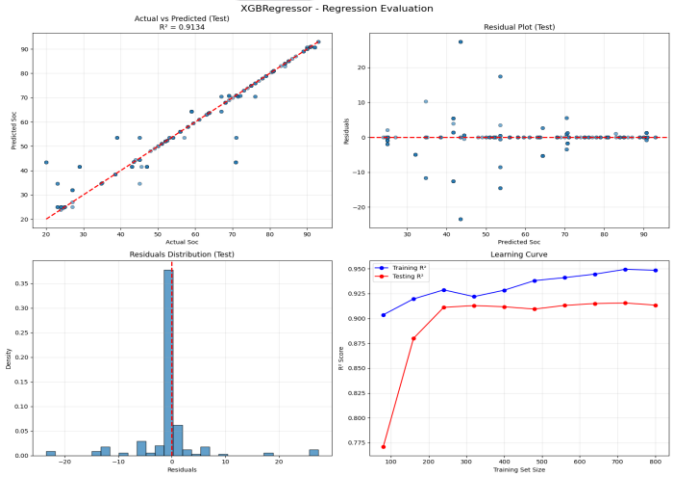

Fig. 1. XGBoost Regressor SOC evaluation: actual vs. predicted SOC, residual plot, residual distribution, and learning curve.

Fig. 2. Extra Trees Regressor SOC evaluation with regression fit, residual analysis, residual histogram, and learning curve.

The Random Forest Regressor provides the best trade-off between accuracy and robustness, achieving a test R2 of approximately 0.9137, RMSE of 6.16, MAE of 2.55, and MAPE of 6.67%. In contrast, the Linear Regression and Linear SVR models exhibit significantly lower R2 values and wider residual spreads, as illustrated side-by-side in Fig. 3 and Fig. 4. These results confirm that linear models are unable to capture the nonlinear impact of route geometry, distance, and terrain on SOC.

Fig. 3. Linear Regression SOC evaluation on the test set, showing limited ability to model nonlinear relationships.

Fig. 4. Linear SVR SOC evaluation, highlighting large residuals and lower R2 compared to ensemble models.

Feature-importance analysis for the Random Forest SOC model indicates that source longitude, distance travelled, and selected altimetry features are the most influential predictors. This agrees with physical expectations that route length and elevation profile strongly affect energy consumption in electric vehicles [5].

B. SOH Estimation Results

All models show good accuracy in SOH estimation, but again tree-based ensemble models outperform on Error metrics and stability. Fig. 5 shows the evaluation of RF Regressor, where predicted SOH values are seen fitting the line close to the diagonal and residuals' distribution is observed to be narrowly centered.

Fig. 6 displays the analogous plots for the XGBoost Regressor, which reach comparable R2 levels but have somewhat greater residual variance. Results for Linear SVR and Linear Regression (Fig. 7 and Fig. 8 respectively) are not far behind, but are not quite as good as the ensemble methods in MAPE and RMSE.

Across all experiments, the Extra Trees Regressor achieves the best overall SOH performance with a test R2 of about 0.9955, RMSE of 1.24, MAE of 0.96, and MAPE of 1.75%. Feature-importance analysis shows that internal resistance and capacity are the dominant predictors, followed by temperature and discharge time, which is consistent with established degradation mechanisms in lithium-ion batteries [3, 6].

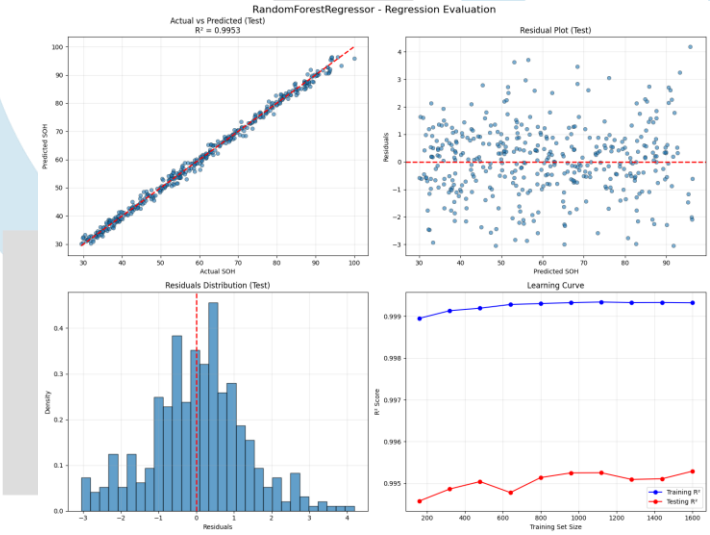

Fig. 5. Random Forest Regressor SOH evaluation: actual vs. predicted SOH, residual plot, residual histogram, and learning curve.

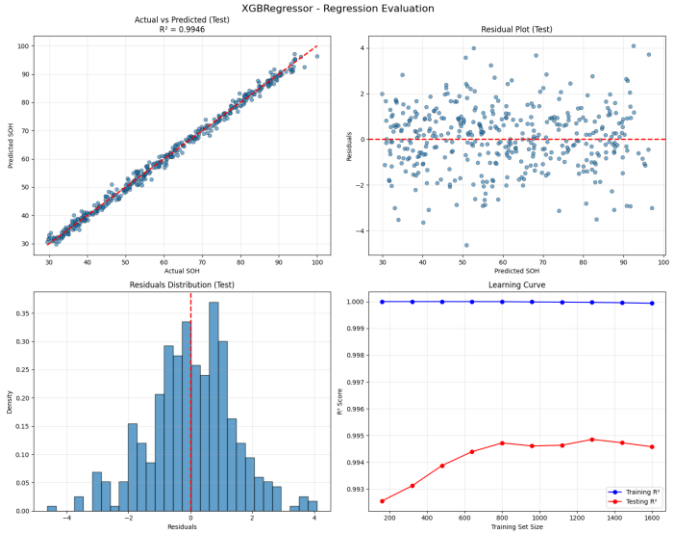

Fig. 6. XGBoost Regressor SOH evaluation, demonstrating high accuracy and smooth learning behavior.

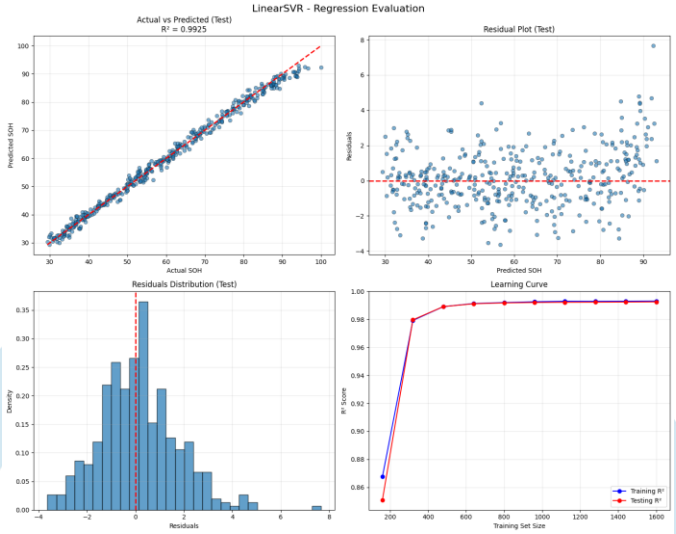

Fig. 7. Linear SVR SOH evaluation with actual vs. predicted plot, residual analysis, and learning curve.

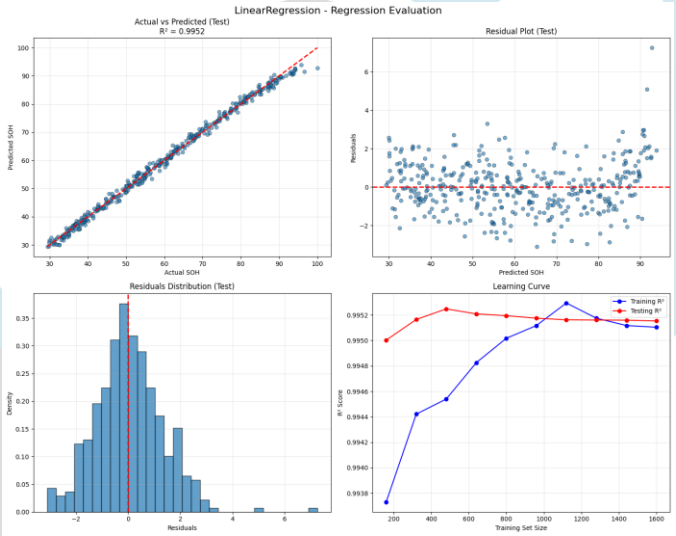

Fig. 8. Linear Regression SOH evaluation, illustrating slightly higher residuals compared to ensemble models.

C. Integrated SOC–SOH Interface

The final Predicted Random Forest (SOC) and Extra Tree (SOH) models were embedded within the earlier outlined Streamlit web application. This application allows users to switch prediction mode, select a route or other battery related parameters through a form, and then receive numerical predictions on the charge level of the battery, as well as qualitative indicators such as health status and recommended actions. This highlights the ease with which the developed models can be deployed.

VI. Conclusion and Future Work

In this paper, we have proposed a consolidated machine learning approach to estimate the State of Charge and State of Health of battery cells based on specific datasets of SOC and SOH along with several regression techniques. In this comparison, the model used for SOC prediction is the Random Forest Regression and the model used for SOH prediction is the Extra Trees Regression, which produce highly efficient R2 and low error rates. The feature importance values also indicate that the primary feature that determines the SOC estimation is the distance, longitude and altimetry while it is the internal resistance and capacity for SOH. This result is consistent with established knowledge on battery physics [6].

Finally, the models were implemented within an interface using Streamlit, which can be utilized by laypersons for online SOC and SOH estimation. In terms of future work, there are a few directions to go in. One direction is to better capture time-varying behavior and cycling history by adopting recurrent or attention based neural networks. This has the potential to further improve the accuracy of SOC and SOH estimation [4, 6]. Another aspect is to include online sensor streams along with an IoT interface to have a continuous update of SOC and SOH. Expanding the system for different chemistries and pack configurations, and testing with large EV or grid storage data sets would improve the generality [7]. Anomaly detection and fault diagnosis would also be added as modules in the future so as to make the system a complete predictive maintenance system.

References

[1] X. Hu, J. Jiang, D. Cao, B. Egardt, "Battery State-of-Charge Estimation for Electric Vehicles: A Review," IEEE Trans. Veh. Tech., vol. 65, 2016.
[2] Y. Li, K. Huang, J. Liu, L. Zhang, "SOH and SOC Estimation Techniques for Batteries in Electric Vehicles: A Review," World EV J., vol. 14, 2023.
[3] M. Berecibar, et al., "State of Health Estimation in Lithium-Ion Batteries: A Critical Review," Renew. Sust. Energ. Rev., vol. 56, 2016.

[4] J. Zhao, et al., "State Estimation and Remaining Useful Life Prediction Methods for Lithium-Ion Batteries: A Review," Sustainability, vol. 15, 2023.
[5] P. S. Babu, et al., "Enhanced SOC Estimation of Lithium-Ion Batteries with Real-Time Driving Data Using Machine Learning," Sci. Rep., 2024.
[6] Y. Wang, et al., "A Comprehensive Review of Machine Learning-Based State-of-Health Estimation for Lithium-Ion Batteries," Renew. Sust. Energ. Rev., 2025.
[7] S. Gupta, P. Kumar, "Estimation of SoC, SoH and RUL of Li-Ion Battery: A Review," in Proc. IEEE, 2023.
[8] M. H. Sulaiman, et al., "State of Charge Estimation for Electric Vehicles Using Data-Driven Methods," Energy Reports, 2024.
[9] A. Author, et al., "Data-Driven Random Forest Regression for SOH Estimation of EV Batteries," SSRG Int. J. EEE, vol. 12, no. 1, 2025.