Battery Cell Performance Analysis and SOH/SOP Prediction

Abstract

This report presents a full analysis of multi-OEM lithium-ion cell test data provided as part of the Lohum Summer Internship problem statement. The objective was to (i) identify the meaning of Feature 10, (ii) conduct intra- and inter-OEM performance comparison, and (iii) build machine-learning models to estimate State of Health (SOH) and State of Power (SOP). The analysis follows the tasks described in the problem statement and executes a complete workflow involving data loading, preprocessing, feature extraction, performance computation, visualizations, and Random Forest–based regression models with hyperparameter tuning. Results demonstrate significant OEM-level differences and strong SOH/SOP prediction performance.

1 Introduction

Second-life battery markets face a major challenge due to lack of standardization in end-of-first-life cell feedstock. Cells from different OEMs vary in chemistry, form factor, and degradation patterns, complicating their classification and selection for secondary applications such as stationary storage. The problem statement provides electrical test data from three OEMs and outlines specific objectives: identifying Feature 10, conducting OEM-wise comparison, and building predictive models.

The analysis presented in this report is derived from the detailed notebook implementation.

2 Dataset Description

The dataset consists of electrical test data from three OEMs, each containing ten text files with measurements from 256 cells. Each file contains columns representing time, voltage, current, discharge capacity, power, and an unknown Feature 10. The column definitions, as specified in the problem statement, are:

- id1 (non-informative)
- cell number
- step number
- id2 (non-informative)

- data count
- time
- time (minutes)
- voltage (mV)
- current (mA)
- discharge capacity (mAh)
- Feature 10
- power (W)

3 Methodology

The full workflow implemented in the analysis notebook included:

3.1 Data Loading

Each manufacturer was associated with specific filename patterns. Files were read, assigned column labels, and combined into consolidated dataframes.

3.2 Identifying Discharge Steps

Discharge steps were detected by inspecting current signs. Any step with negative current was marked as a discharge phase.

- Manufacturer A: step 3
- Manufacturer B: step 3
- Manufacturer C: step 5

3.3 Cell Performance Metrics

For every discharge step, the following metrics were computed:

- Maximum discharge capacity (mAh)
- Total energy delivered (mWh)
- Maximum power (W)
- Initial voltage, mean voltage, mean current, mean power
- Test duration (minutes)

3.4 Health Metric Estimation

State of Health (SOH) and State of Power (SOP) were defined using the 95th percentile as the reference value:

$$SOH = \frac{\text{capacity}}{\text{95th percentile capacity}}, \quad SOP = \frac{\text{energy}}{\text{95th percentile energy}}$$

3.5 Machine Learning Models

Random Forest Regressors were used for predicting SOH and SOP.

$$X = [\text{duration}, V_{\text{init}}, \overline{V}, \overline{I}, \overline{P}]$$

GridSearchCV was used with the search space described in the analysis notebook. Five-fold cross-validation was used to compute: RMSE, MAE, MAPE, and R^2.

4 Results and Discussion

4.1 OEM-Level Data Summary

The total number of test records loaded from each OEM was:

- Manufacturer A: 381,364 records
- Manufacturer B: 527,113 records
- Manufacturer C: 583,593 records

4.2 Cell Performance Summary

OEM	Avg Capacity (mAh)	Std Dev (mAh)	Cells Analyzed
A	4994.5	1426.7	256
B	4733.9	83.6	256
C	4577.3	84.0	256

Table 1: Cell Performance Metrics Across OEMs

Manufacturer A exhibited the largest spread, indicating heterogeneous feedstock quality, whereas B and C showed tightly clustered capacities.

4.3 SOH/SOP Reference Values

- Manufacturer A: Ref capacity 10047.8 mAh, SOH mean 0.497
- Manufacturer B: Ref capacity 4826.0 mAh, SOH mean 0.981
- Manufacturer C: Ref capacity 4714.0 mAh, SOH mean 0.971

This shows Manufacturer A's dataset contains more degraded or diverse cells.

4.4 Model Selection for ML

Manufacturer A was chosen for modeling as it displayed:

- Highest average capacity (4994.5 mAh)
- Largest variability—ideal for learning generalizable patterns

4.5 SOH Prediction Results

- RMSE: 0.0227
- MAE: 0.00929
- MAPE: 1.54%
- R^2: 0.9751

4.6 SOP Prediction Results

- RMSE: 0.0132
- MAE: 0.00961
- MAPE: 1.00%
- R^2: 0.5914

SOH prediction achieved excellent accuracy, while SOP prediction was moderately accurate due to lower feature-to-target correlation.

5 Conclusion

The analysis successfully fulfilled all objectives from the problem statement. The following conclusions were drawn:

- Significant OEM-level differences exist in performance, variability, and degradation.
- Manufacturer A exhibits highest variation and highest mean capacity.
- SOH and SOP can be predicted from simple electrical features with low error rates.
- The Random Forest model showed excellent performance for SOH (MAPE 1.54%) and good performance for SOP (MAPE 1.00%).

This framework can be extended for large-scale second-life battery screening and automated quality classification.