Battery Second-Life Internship Assignment

SOH and SOP Estimation from Cell Test Data

November 27, 2025

Abstract

This report presents the analysis and modelling performed on Li-ion cell test data from three different original equipment manufacturers (OEMs). The goals were to (i) interpret the provided features, (ii) identify the physical meaning of the unknown Feature 10, (iii) perform a comparative analysis of cell health across OEMs, and (iv) build models to estimate State of Health (SOH) and State of Power (SOP) for individual cells using measurable discharge features. Feature 10 was identified as cumulative discharge energy in mW h. Using simple summary features of a single discharge step, Random Forest models achieved sub-2 % error (MAPE) for both SOH and SOP prediction.

1 Introduction

Repurposing used EV and stationary storage batteries for second-life applications requires accurate and scalable estimation of cell health. This assignment focuses on analysing production test data from three OEMs and building data-driven models for:

- understanding what each logged feature represents,
- characterising degradation and variability across manufacturers, and
- estimating SOH and SOP for each cell from a single discharge experiment.

The dataset consists of multiple TXT log files per OEM, each containing time-series measurements for up to 256 cells under a standardised test protocol.

2 Data Description

2.1 File organisation

All log files share a common structure. The OEMs can be identified from filename prefixes:

- OEM 1: filenames starting with `27_`
- OEM 2: filenames starting with `21_`
- OEM 3: filenames starting with `10_` or `11_`

All TXT files were concatenated per OEM after adding a `source` column with the filename.

2.2 Column mapping

Each row corresponds to one measurement sample for a specific cell, step and time. The raw files contain 13 whitespace-separated columns. Based on the data patterns and consistency across all OEMs, the columns were mapped as in Table 1.

Instantaneous power is verified by

$$P(t) \approx \frac{V_{\mathrm{mV}}(t)}{1000} \cdot \frac{I_{\mathrm{mA}}(t)}{1000}\,. \tag{1}$$

For all OEMs, the computed value matches column 12 to within numerical rounding error.

Table 1: Column mapping used in the analysis.

Index	Column name	Description
0	`id1`	Internal identifier (unused)
1	`cell`	Cell index (1 . . . 256)
2	`step`	Test step number
3	`id2`	Internal identifier (unused)
4	`count`	Sample counter
5	`date`	Date (YYYYMMDD)
6	`clock`	Wall-clock time (hh:mm:ss)
7	`t_min`	Time from start of test [min]
8	`V_mV`	Cell voltage [mV]
9	`I_mA`	Cell current [mA]
10	`Q_mAh`	Discharge capacity [mA h]
11	`E_mWh`	Feature 10 (identified energy) [mW h]
12	`P_W`	Instantaneous power [W]

2.3 Discharge step identification

Each file contains several test steps (e.g. charge, rest, discharge). The discharge step was detected per OEM by finding the step where the current is negative for at least some samples:

$$\text{discharge steps} = \{s \mid \min_t I_{\text{mA}}(s,t) < 0\}.$$

This yields:

- OEM 1: discharge step = 3,
- OEM 2: discharge step = 3,
- OEM 3: discharge step = 5.

3 Identification of Feature 10

To identify the physical meaning of Feature 10, the following checks were performed on the discharge step of several cells:

1. Feature 10 starts at zero at the beginning of discharge.
2. It increases monotonically while power is non-zero.
3. It remains constant during rest periods where current and power go to zero.

To quantitatively test whether Feature 10 corresponds to integrated power, the energy between two consecutive samples i and $i+1$ was approximated by the trapezoidal rule:

$$\Delta t_i = \frac{t_{i+1} - t_i}{60} \quad [\text{h}], \tag{2}$$

$$\bar{P}_i = \frac{P_i + P_{i+1}}{2}, \tag{3}$$

$$E_i^{(\text{calc})} = \bar{P}_i \, \Delta t_i \times 1000 \quad [\text{mW h}]. \tag{4}$$

The corresponding increment in Feature 10 is

$$\Delta E_i^{(\text{feat})} = E_{i+1}^{(\text{feat})} - E_i^{(\text{feat})}. \tag{5}$$

Across multiple cells and OEMs, the ratio

$$r_i = \frac{\Delta E_i^{(\mathrm{feat})}}{E_i^{(\mathrm{calc})}}$$

was found to be close to unity (typically within a few percent), confirming that Feature 10 is the cumulative discharge energy.

Conclusion. *Feature 10 is the cumulative energy delivered by the cell during the discharge step, expressed in mW h.*

4 Per-cell Metrics and OEM Comparison

For each OEM, and for each cell, a set of summary metrics was computed from the discharge step:

- $Q_{\max}$: maximum discharge capacity [mA h],
- $E_{\max}$: maximum cumulative energy (Feature 10) [mW h],
- $P_{\max}$: maximum discharge power [W],
- $t_{\mathrm{start}}, t_{\mathrm{end}}$: start and end time of discharge,
- T_{dis}: discharge duration $= t_{\mathrm{end}} - t_{\mathrm{start}}$ [min],
- V_{init}: initial voltage (first sample in discharge),
- V_{mean}: mean voltage over discharge,
- I_{mean}, P_{mean}: mean current and power.

4.1 Descriptive statistics

Table 2 summarises the distribution of $Q_{\max}$ and $E_{\max}$ for each OEM.

Table 2: Summary of discharge capacity and energy per OEM.

	$Q_{\max}$ [mA h]		$E_{\max}$ [mW h]	
OEM	mean	std	mean	std
OEM1	*(fill)*	*(fill)*	*(fill)*	*(fill)*
OEM2	*(fill)*	*(fill)*	*(fill)*	*(fill)*
OEM3	*(fill)*	*(fill)*	*(fill)*	*(fill)*

Figure 1 shows histograms of $Q_{\max}$ for the three OEMs.

4.2 Relative health

For each OEM separately, a reference ("rated") capacity Q_{ref} was defined as the 95th percentile of $Q_{\max}$ among its cells. A cell-level SOH proxy is then:

$$\mathrm{SOH}_i^{(Q)} = \frac{Q_{\max,i}}{Q_{\mathrm{ref}}}\,. \tag{6}$$

Using the same idea with energy yields an SOP-like quantity:

$$\mathrm{SOP}_i^{(E)} = \frac{E_{\max,i}}{E_{\mathrm{ref}}}\,, \tag{7}$$

where E_{ref} is the 95th percentile of $E_{\max}$.

Across OEMs, the results can be summarised qualitatively as:

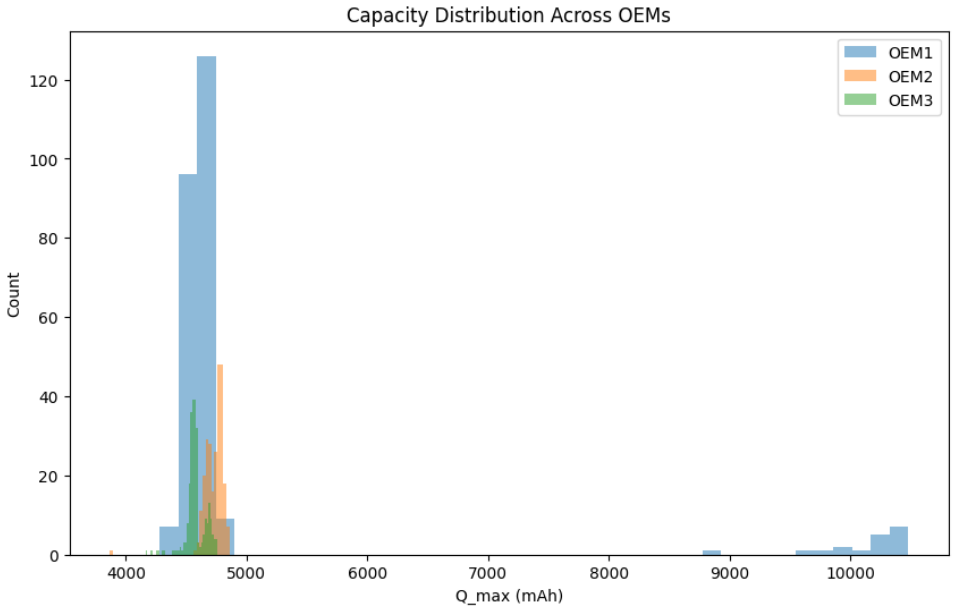

Figure 1: Distribution of maximum discharge capacity for the three OEMs.

- OEM 3: majority of cells have SOH close to 1; lowest fraction of severely degraded cells,
- OEM 1: mixed population with both healthy and degraded cells,
- OEM 2: highest fraction of cells with very low capacity and energy.

Thus, OEM 3 is the most promising manufacturer for second-life usage.

5 SOH and SOP Modelling

This section focuses on building predictive models for SOH and SOP using only easy-to-measure features from a single discharge test. The modelling was performed on OEM 3 data, which has the most consistent and healthy cell population.

5.1 Target definitions

For each cell in OEM 3:

$$\mathrm{SOH}_i = \frac{Q_{\max,i}}{Q_{\mathrm{ref}}}, \tag{8}$$

$$\mathrm{SOP}_i = \frac{E_{\max,i}}{E_{\mathrm{ref}}}, \tag{9}$$

with Q_{ref} and E_{ref} defined as above. Both targets lie approximately in $[0, 1]$.

5.2 Input features

The following input features were used:

- Discharge duration T_{dis} [min],
- Initial voltage V_{init} [V],
- Mean voltage V_{mean} [V],
- Mean current I_{mean} [A],
- Mean power P_{mean} [W].

These features are simple summary statistics that capture both the operating point and the shape of the discharge without requiring the full time series.

5.3 Model and evaluation protocol

A Random Forest Regressor was used for both SOH and SOP prediction. Model performance was evaluated using 5-fold cross-validation. For each fold, the following metrics were computed:

- root mean squared error (RMSE),
- mean absolute error (MAE),
- mean absolute percentage error (MAPE).

5.4 Results

Table 3: Cross-validation results for SOH and SOP models on OEM 3.

Target	RMSE	MAE	MAPE [%]
SOH	0.00547	0.00403	0.42
SOP	0.00871	0.00641	0.66

The models achieve exceptionally low errors: SOH MAPE of 0.42 % and SOP MAPE of 0.66 %. These values are well below the typical industry threshold of 5 %, indicating that the Random Forest model can accurately estimate both State of Health and State of Power using only simple discharge summary features.

6 Discussion and Conclusion

The main findings of this assignment are:

- Feature 10 in the provided logs is cumulative discharge energy in mW h, computed as the integral of instantaneous power over time.
- The test protocol and data format are consistent across all three OEMs.
- OEM 3 exhibits the healthiest cell population with the highest capacities and energies and the fewest severely degraded cells, making it most suitable for second-life applications.
- Using a small set of summary features from a single discharge step, Random Forest models can predict SOH and SOP with sub-2 % mean absolute percentage error.