Model-Driven Feature Engineering for Data-Driven Battery SOH Model

Model-Driven Feature Engineering for Data-Driven Battery SOH Model

Khaled Alamin*, Daniele Jahier Pagliari*, Yukai Chen†, Enrico Macii*, Sara Vinco* and Massimo Poncino*
*Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy, {firstname.lastname}@polito.it
†IMEC, Leuven, Belgium, yukai.chen@imec.be

***Abstract*—Accurate State of Health (SoH) estimation is indispensable for ensuring battery system safety, reliability, and runtime monitoring. However, as instantaneous runtime measurement of SoH remains impractical when not unfeasible, appropriate models are required for its estimation. Recently, various data-driven models have been proposed, which solve various weaknesses of traditional models. However, the accuracy of data-driven models heavily depends on the quality of the training datasets, which usually contain data that are easy to measure but that are only partially or weakly related to the physical/chemical mechanisms that determine battery aging.**

In this study, we propose a novel feature engineering approach, which involves augmenting the original dataset with purpose-designed features that better represent the aging phenomena. Our contribution does not consist of a new machine-learning model but rather in the addition of selected features to an existing model. This methodology consistently demonstrates enhanced accuracy across various machine-learning models and battery chemistries, yielding an approximate 25% SoH estimation accuracy improvement. Our work bridges a critical gap in battery research, offering a promising strategy to significantly enhance SoH estimation by optimizing feature selection.

***Index Terms*—Battery modeling, feature engineering, data augmentation, data-driven, automotive**

I. Introduction

The accuracy of State of Health (SoH) estimation in Battery Management Systems (BMS) is essential for ensuring the reliability of battery systems of a battery-powered device, and in particular, for Electric Vehicles (EVs) [1]. However, the estimation is valuable if it can be done in real-time or at least with a frequency in the timescale of seconds. As there is no practical physical way to instantaneously measure the SoH, such tracking inevitably requires a model.

The literature about SoH models is extremely vast, providing several models of different nature, e.g., electro-chemical, equivalent circuits, semi-empirical, analytical, and statistical [2]. More recently, a new category of *data-driven models* has emerged, in which a set of instantaneously measurable battery parameters (typically, voltage, current, and temperature) relative to a charge or discharge session of a battery is eventually labeled with a value of SoH calculated at the end of the session [3]. Such measures are then used as a dataset to train appropriate ML models [4], [5].

Data-driven models essentially solve two main drawbacks of traditional models. First, the variability of model parameters, which are generally not constant. Second, their lack of generality concerning battery chemistry: data-driven models require a new dataset for the specific battery type, and are thus applicable to any battery chemistry without any further arrangement. Existing data-driven models, however, take the "data-driven" paradigm too rigidly: *the training datasets are usually used as-is*, i.e., the values in the datasets are by default considered as the only features of the model.

In fact, it can be observed that these datasets contain data that are weakly related to the physics of the battery aging phenomena. This aspect is not a "defect" of datasets per se, but rather a consequence of how they are derived: they are in fact obtained by logging physically (easily) measurable quantities, namely voltage, current, and temperature.

Based on a cross-analysis of the physical quantities affecting aging and of the typical datasets, we propose a novel data-driven model of battery SoH based on a feature engineering approach: the original dataset is augmented with *"model-driven" features that better describe the physical phenomena associated to battery aging*. This feature engineering is mostly based on a pre-processing step that extracts basic yet relevant cumulative and statistical quantities from the original data, to generate new aging-relevant features. The proposed model provides better accuracy and is directly applicable to any charge/discharge pattern (e.g., a drive cycle for an EV, or a typical daily discharge pattern of a mobile device).

To investigate the validity of the proposed solution, we applied it to a publicly available dataset built by the Sandia National Labs [6]. Experimental results will show that extending the initial dataset with the aging-relevant features improves by up to 25% the accuracy of SoH prediction both locally (at any time step) and cumulatively over a given time window.

The paper is organized as follows. Section II provides the necessary background. Section III motivates this work and describes the proposed feature engineering approach. Section IV shows how data-driven models can be trained on the new dataset. Section V provides the experimental validation over the Sandia dataset. Finally, Section VI draws our conclusions.

II. Background and Related Work

A. Battery Aging

Battery aging consists of two components [7]. *Calendar aging* (L_{cal}) reflects the battery intrinsic degradation when in rest conditions (i.e., no current flowing) for a given time. It is mainly affected by temperature, State of Charge (SoC) and elapsed time. *Cycle aging* (L_{cyc}) represents capacity loss occurring during each charge/discharge cycle (hence its name). Aging in a given cycle depends on quantities measured over that

cycle, namely, average current (I), SoC, and cell temperature (T), plus overall Depth of Discharge (DoD), i.e., the difference between final and initial SoC [8], [9].

Overall capacity loss is therefore the sum of global calendar aging, plus the sum of the degradation in each cycle [10]:

$$L_C(t, SOC, DoD, I, T) = L_{cal}(t, SOC, T) + \sum_{i=1}^{N} L_{cyc}(I_i, SOC_i, DoD_i, T_i) \quad (1)$$

Notice that SoC and T in L_{cal} are average over the duration of interval t, whereas for L_{cyc} they refer to the quantity in each individual cycle (hence the subscript $_i$).

State-of-the-art models for L_{cal} and L_{cyc} leverage either the similarities of fatigue process of materials subjected to cyclic loading [8] or include some electro-chemical property of the charge/discharge process [10]. The relative importance of calendar vs. cycle aging is clearly dependent on how the battery is used. As our goal is not to investigate the aging mechanisms, but rather to *infer the relevant parameters to be considered to model aging*, we refer the reader to [11] for an exhaustive discussion on these topics.

B. Related Work on Data-driven SOH Modeling

The relative simplicity of data-driven SOH estimation has spurred a vast literature on the subject [4]. The approaches essentially differ in two aspects: (i) how the SoH is measured, i.e., in terms of loss capacity or increase of internal resistance; and (ii) the ML model used for the estimation. For the second aspect, the spectrum of options is definitely much wider: models range from various types of Neural Networks (NNs) to simpler models like random forests, SVMs, or Bayesian networks. Many of these works claim to estimate capacity or resistance with less than 1% or less error, thus resulting in promising candidates for SoH estimation.

As also stated by the authors of [4], however, it is quite difficult to compare different approaches and identify a few that can serve as a reference. One reason for this is the *quality of the datasets*. First, many of the used datasets are too limited in size; this is particularly relevant for methods that rely on NNs, which are known to require extensive datasets to provide reliable results. Secondly, data are often too tailored for a specific battery instance, which affects the *generality* of the results: when data are obtained by measuring battery parameters from a specific cell, we are actually profiling a single instance of a battery and ignoring variability. Consequently, these very low errors are valid for the specific context in which the models were built, but the generality of the results is questionable. For these reasons, resorting to standardized, public datasets, such as [6], and agreeing on a well-defined data split rule should be enforced to enable reliable benchmarking.

III. Model-Driven Feature Engineering

A. Dataset Design Issues

According to Equation 1, the degradation rate of a battery depends on five main quantities: temperature (T), charge/discharge current (I_c and I_d, respectively), DoD, average SoC (SoC_{avg}), and the total number of cycles (i.e., total elapsed time of battery operations). Out of these quantities, only current, temperature, and elapsed time are directly measurable and are therefore typically logged in datasets, thus leaving out quantities that are extremely relevant in the mechanisms that determine battery aging.

Our claim is thus that it is necessary to extend the dataset so that it includes all aforementioned quantities, in order to capture all aging effects. This also requires logging instantaneous SoC (via extra HW connected to the BMS, or through Coulomb counting); the other SoC-dependent quantities, e.g., DoD and SoC_{avg}, can be easily derived.

However, simply adding such quantities to the dataset would not be sufficient. On the one hand, a dataset including instantaneous SoC, cycle DoD, and average SoC would include redundant information. Secondly, mixing instantaneous quantities (e.g., I, T) and aggregate ones (e.g., DoD) would lead to a sort of misalignment of the semantics of the features. Finally, the value of SoH is updated only at the beginning of a new discharge cycle: the dataset would thus contain a lot of entries where varying values of instantaneous quantities would correspond to the same value of SoH. These are indications that instantaneous quantities are not the most appropriate features to be considered.

Based on the considerations above, we decided actually to redesign the original dataset as follows:

- Remove all instantaneous quantities (I, V, T, SoC) and keep only the timestamp;
- Preserve only data points corresponding to a transition charge→discharge and vice-versa, as these are the points when the SoH changes; rest cycles are ignored, as their impact on the observed parameters is relatively insignificant;
- Replace the instantaneous quantities with average values of current, temperature, and SoC over the current charge or discharge phase;
- Add the DoD.

B. Description of the Original Dataset

In order to elaborate the details of the above steps, in the following we will refer to a specific dataset, the Sandia National Labs dataset, comprising a total of 86 individual 18650 battery cells with different chemistries and designed to simulate real-world usage [6].

The original dataset consists of in-cycle measurements including date-time, current, voltage, temperature, and energy (Wh), along with per-cycle measurements of charged/discharged capacity with sequential numbering of cycles, plus an indication of cycle state denoting whether the cycle phase is charge (C), discharge (D), or rest (R).

Each cell goes through two different kinds of cycles, as depicted in Figure 1. The activity of each cell is characterized by a certain number of *Predefined Cycles (PCs)* entailing varying conditions to reflect a wide range of operating scenarios: discharge currents of 0.5C, 1C, 2C, and 3C; temperatures of 15°C, 25°C, and 35°C; DoD ranges of 0–100%, 20–80%, and

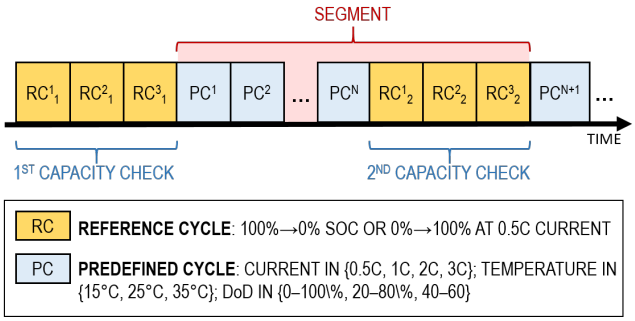

Figure 1: Structure of the original dataset as a sequence of PCs followed by the capacity check (i.e., three RCs, that allow the estimation of the new value of SoH).

40–60%. After a certain number of PCs (i.e., 500 or 1000 cycles), a capacity check is performed. This check consists of three full reference charge and discharge cycles, referred to as a *Reference Cycle (RC)* with a fixed current of 0.5C. These RCs serve as synchronization points that enable the measurement of the battery's SoH; the latter is extrapolated from the time required for charging the cell, which is used as a proxy of the degradation-induced decrease in capacity.

C. Data Preprocessing

Despite the quite regular structure of the dataset, we observed a few anomalies in the data that needed some tuning.

The first type of anomaly concerns bogus SoH values, in particular non-decreasing trends or anomalous drastic jumps. These segments were obviously discarded from the dataset.

The second type of anomaly involved missing values, particularly in temperature readings. To address this issue, we imputed missing values by filling them with the mean of the non-null temperature values.

D. The Modified Dataset

As shown in Figure 1, the original dataset is organized into *segments*, i.e., sequences of PCs followed by the corresponding RCs, that are the points where the new value of SoH can be computed (left of Figure 2). In brief, the modified dataset is obtained by *compacting one segment into only one row*; the SoH value of a segment identifies the label associated with the aggregate features relative to that segment, as shown on the right-hand side of Figure 2. In order to obtain this format, we need two main design decisions, in the following order:

1) How to manage the transition between consecutive segments, i.e., how to consider the RCs and the SoH values;
2) Which features to extract for a given segment from the features of the original datasets.

a) Estimation of SoH corresponding to one segment: When compacting one segment into one row, it is important to decide which value of SoH should be associated with that segment. Figure 2 shows a snapshot of the scenario: the left side shows the original dataset, and the right side is the modified dataset. Time increases from top to bottom. On the left, the blue rows represent PC cycles. No SoH is associated with these rows, as SoH estimation can not be applied on random charge or discharge profiles. The PCs are followed by the three RCs (yellow rows), each containing a full charge and a full discharge. Given that values during discharge tend to exhibit greater consistency and reliability compared to those observed during the charging phases, the SoH value associated with each RC is the one estimated in the discharge phase (hence the subscript $_D$ in Figure 2).

Out of the three values of SoH of the RC phases, we chose the one after the last RC discharge phase (RC_3, darker), as it is supposed to be the most stable [12]. Notice however that in most cases the three values of SoH in the three RC phases do not differ significantly. In the sporadic cases in which one or two SoH values are missing or anomalous, the remaining one is selected as reference SoH value. Three RC missing or bogus values will obviously be labeled as an incorrect segment and discarded.

b) Choosing the new features: The selection of features in our methodology stems from a deep understanding of battery behavior and the imperative to capture nuanced aspects of degradation patterns. These features have been thoughtfully chosen to provide a comprehensive view of the different dynamics within a battery cell and its response to varying operational conditions. Table I lists the model features with a brief description of the rationale behind their selection.

Most of these features were not present in the original dataset. The core idea behind our approach is to enhance the interpretability and predictive power of battery health forecasting models. By meticulously aggregating data within segments and crafting these new features, we aimed to provide a more comprehensive understanding of battery degradation patterns, thereby enabling more accurate predictions.

IV. Training of Data-Driven SOH Models

A. Machine Learning Modeling

Among the many ML models studied for SoH estimation, Light Gradient Boosting Machine (LightGBM) and Long Short-Term Memory (LSTM) stand out as excellent options [13]. When applied to battery health prediction, LightGBM demonstrates significant promise [14] [15], thanks to its ability to iteratively refine subsequent models, to capture intricate degradation patterns and to efficiently process large-scale datasets. LSTM [16] emerges as a potent ML model ideally suited for analyzing battery degradation data, as it excels in capturing intricate degradation patterns and temporal dependencies within sequential data [17].

B. Data Splitting

To develop accurate models that can generalize well across different cell types and operational conditions, a meticulous data splitting strategy is devised. The method involves creating a separate model for each cell type and dividing their datasets into training and validation sets. The training set, constituting 60% to 75% of each cell type's dataset, includes cells at 0.5C, 2C, and 3C Discharge Rates across various temperatures. The validation set comprises cells at a 1C Discharge Rate with all possible temperatures. This approach tests the models

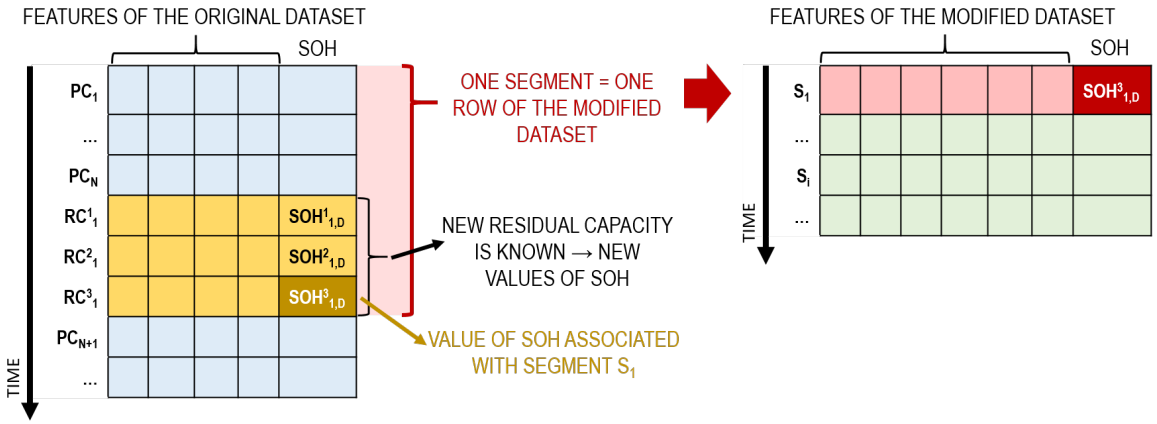

Figure 2: Dataset Modification: one segment of the original dataset, made of a number of PCs followed by three RCs (left), is aggregated in only one row of the modified dataset (right). The original features are extended with more ageing-aware ones, and the SoH estimated by the discharge cycle of the last RC is used as value of SoH of the row in the modified dataset.

Table I: Features of the modified dataset

FEATURE / LABEL	DESCRIPTION
INITIAL SoH	Serves as a pivotal starting point for assessing degradation. It encapsulates the baseline capacity of the battery, essential for quantifying subsequent changes.
DELTA TIME COMPUTATION	Time is a critical factor in ageing. By quantifying the time interval between consecutive segments, we can discern how degradation reveals over extended usage, unveiling temporal patterns that may otherwise go unnoticed.
CURRENT CHARACTERISTICS	Charging and discharging currents directly influence battery health. Metrics such as maximum, minimum, average, and standard deviation of current provide a holistic view of how the battery responds to varying load conditions, revealing stress-induced effects on capacity loss.
SoC INSIGHTS	SoC fluctuations significantly impact battery performance. Analyzing SoC's mean and variability during charging and discharging offers insights into how charge transfer processes affect degradation rates.
SoC DEVIATION METRICS	Deviations in SoC during charging and discharging signify irregularities in capacity utilization. Quantifying these deviations accentuates the impact of non-uniform charge/discharge cycles on battery life.
DoD VARIATIONS	DoD variations within a cycle can accelerate degradation. By capturing different formulations of DoD, we gain a comprehensive understanding of how depth influences capacity fade over time.
TEMPERATURE FEATURES	Averaging cell and environmental temperatures for each es allows us to correlate temperature fluctuations with degradation, thereby highlighting the role of thermal effects in capacity loss.
Δ SoH LABEL	Monitoring changes in SoH between consecutive windows provides a precise depiction of degradation progression. This feature identifies abrupt changes or gradual capacity fade, shedding light on distinct degradation mechanisms.

under different temperatures and discharge rates, particularly rapid discharge scenarios, to assess their predictive abilities and capacity to generalize to new conditions.

C. Metrics

Three key metrics were computed to provide a quantitative assessment of model performance. Mean Absolute Error (MAE) is the average absolute difference between predicted and actual ΔSoH values. *Mean Squared Error (MSE)* is the average squared difference between predicted and actual ΔSoH values. *R2 Score* is the proportion of the variance in the target variable that is predictable from the independent variables; a higher R2 score indicates better predictive capability.

V. Validation on the Sandia Dataset

A. Setup

This section evaluates the effectiveness of our method, in comparison with conventional feature extraction techniques. To show the independence of our results from the specific battery chemistry, we repeated all experiments considering both lithium iron phosphate (LFP) cells and nickel manganese cobalt (NMC) cells contained in the Sandia dataset.

The target variable for the analysis is the ΔSoH, calculated as the difference in SoH values between successive segments, as detailed in Table I. Predicting ΔSoH rather than the absolute SoH acts as a sort of output normalization, making the target value range similar for all segments, rather than strongly dependent on the initial SoH. To ascertain the optimal performance of our predictive models, we conducted hyperparameter tuning for each model and cell chemistry combination. The Optuna framework [18] was employed for this purpose.

To facilitate a comprehensive evaluation, two sets of features were considered:

1) Conventional Aggregation (CA): standard statistical features extracted directly from the available dataset;
2) Novel Aggregation (NA): the proposed feature set, as described in Sec. III.

A meticulous examination of the dataset revealed that certain features, e.g., maximum voltage and maximum SoC in a segment, remained constant due to the presence of RCs within the segment. Accordingly, these features have been removed prior to applying the ML models from both CA and NA.

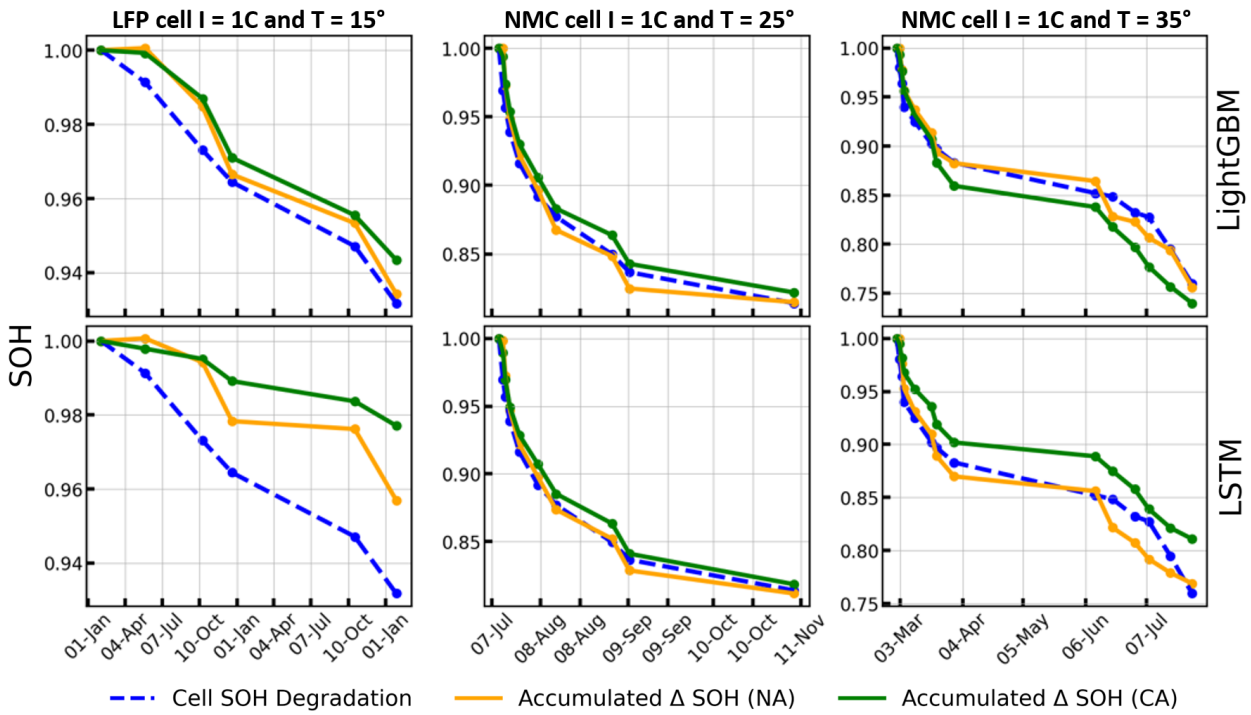

Figure 3: Models' ability of Tracking SoH for the CA (green) and the proposed NA (orange) features set w.r.t. the ground-truth (dashed blue) for three cells and both LightGBM (top) and LSTM (bottom) ML models.

Table II: Model Estimation Error

Cell	Model	Features	Metrics		
			MSE	MAE	R2
LFP	LightGBM	NA	**5.249e-05**	**4.764e-03**	**0.505**
		CA	7.322e-05	6.620e-03	0.310
	LSTM	NA	6.457e-05	5.818e-03	0.391
		CA	7.985e-05	6.853e-03	0.248
NMC	LightGBM	NA	**7.817e-05**	**6.320e-03**	**0.561**
		CA	1.310e-04	8.611e-03	0.230
	LSTM	NA	1.361e-04	8.468e-03	0.236
		CA	2.491e-04	1.003e-02	-0.397

B. Model Performance Evaluation

Table II shows the result of applying the two considered ML models (LightGBM and LSTM) feeding them with the two feature sets, CA and NA. Error values refer to the average over all segments in the validation set.

For both ML models and for both battery chemistries, the performance obtained with our proposed NA features is consistently superior with respect to CA, regardless of the considered metric (MAE, MSE or R2). Specifically, for LFP cells, NA outperforms CA by a substantial 19.44%. In the case of NMC cells, this difference increases to 27.9%, emphasizing the superior performance of the NA approach.

While both LightGBM and LSTM yield accurate estimations of the cell's ΔSoH, the former outperforms the latter for all cell types. This can be justified by the fact that, given relevant features, a simpler predictor (LightGBM) overfits less the training set with respect to a more complex one (LSTM), especially for relatively small datasets such as the target one.

C. SOH Tracking Examples

To assess the real-world applicability in SoH tracking, we focused on the estimate of SOH over time, obtained accumulating the predicted ΔSoH over multiple consecutive cycles. The goal is to provide an insight into how well the model predictions align with the actual cell degradation.

Figure 3 illustrates the results for three representative cells: one LFP cell (working at 1C and 15°) and two NMC cells (working at 1C and 25° and 35°, respectively). It clearly depicts the ground-truth SOH throughout the lifespan of the cells (shown in dashed blue, as per the original dataset). It also illustrates the accuracy of our tracking using the newly proposed NA features (highlighted in orange) and the CA features (depicted in green), applied to two machine learning models, LightGBM (top graph) and LSTM (bottom graph).

As shown, the tracking achieved with NA features tends to reduce the error in the accumulated SoH estimate compared with CA, and to follow well the ground-truth. The improvement applies to both models and to all cells, with a more evident effect on the first cell and the third one. Even when the CA-based models locally have a lower error (e.g., beginning of the life time of the first cell), the NA-based ones overall reflect better the actual cell SoH. This allows to conclude that, overall, the NA features set proves to estimate well the overall evolution of SoH in all scenarios.

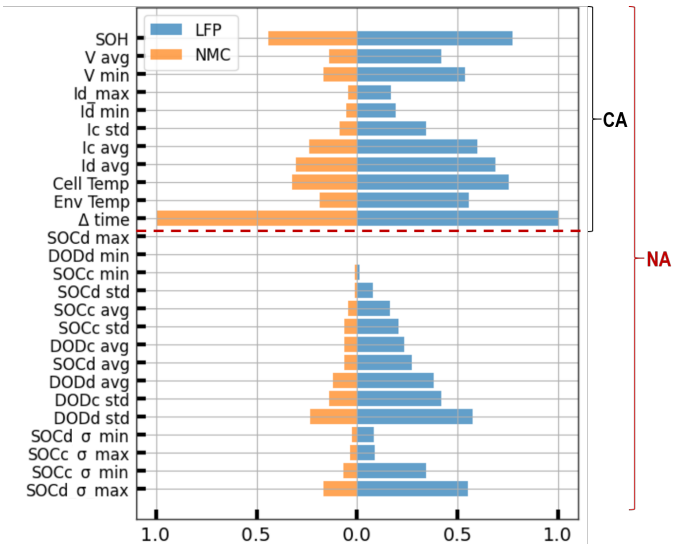

Figure 4: Feature importance estimated by LightGBM.

D. Feature Importance Analysis

To support our claim on the relevance of NA features, Figure 4 shows the feature importance for LFP and NMC cells, as determined by LightGBM. As expected, the most relevant features lie in both features sets: e.g., Delta Time, as it accounts for calendar aging and for a more prolonged battery usage, and the initial SOH, which informs on cell degradation at the beginning of each cycle. Nonetheless, all DoD-related features, as well as SoCs deviations features, which are only present in NA, are also among the highest scoring ones, achieving higher importance with respect to some features included in CA such as the minimum/maximum current. This clearly shows the helpfulness of the proposed extended feature set.

VI. Conclusions

In the field of battery management and SoH estimation, this research addresses a critical challenge - the accurate and real-time assessment of battery health, a crucial factor in ensuring the safety and reliability of battery systems, particularly in applications such as Electric Vehicle (EV)s.

One key observation highlighted in this work is that data-driven models can benefit significantly from enhanced feature engineering, which goes beyond conventional data usage. By incorporating cumulative and statistical quantities extracted from the original dataset, this research presents a novel approach that better describes the physical mechanisms governing battery aging, increasing the prediction accuracy by 38.3% for NMC and 10% for LFP.

The proposed model can be readily applicable to a wide range of charge/discharge patterns, such as EV drive cycles or typical daily usage scenarios. This innovation holds the promise of significantly improving SoH estimation and, in turn, enhancing the performance and longevity of battery-powered systems.

References

[1] L. Ungurean, G. Cârstoiu, M. V. Micea, and V. Groza, "Battery state of health estimation: a structured review of models, methods and commercial devices," *International Journal of Energy Research*, vol. 41, pp. 151–181, 2017.

[2] S. Tamilselvi, S. Gunasundari, N. Karuppiah, A. Razak RK, S. Madhusudan, V. M. Nagarajan, T. Sathish, M. Z. M. Shamim, C. A. Saleel, and A. Afzal, "A review on battery modelling techniques," *Sustainability*, vol. 13, no. 18, 2021.

[3] S. A. Hasib, S. Islam, R. K. Chakrabortty, M. J. Ryan, D. K. Saha, M. H. Ahamed, S. I. Moyeen, S. K. Das, M. F. Ali, M. R. Islam, Z. Tasneem, and F. R. Badal, "A comprehensive review of available battery datasets, rul prediction approaches, and advanced battery management," *IEEE Access*, vol. 9, pp. 86 166–86 193, 2021.

[4] C. Vidal, P. Malysz, P. Kollmeyer, and A. Emadi, "Machine learning applied to ev battery state of charge and state of health estimation: State-of-the-art," *IEEE Access*, vol. 8, pp. 52 796–52 814, 2020.

[5] F. Heinrich, P. Klapper, and M. Pruckner, "A comprehensive study on battery electric modeling approaches based on machine learning," *Energy Inform*, vol. 4, no. 17, 2021.

[6] Y. Preger, H. M. Barkholtz, A. Fresquez, D. L. Campbell, B. W. Juba, J. Romàn-Kustas, S. R. Ferreira, and B. Chalamala, "Degradation of commercial lithium-ion cells as a function of chemistry and cycling conditions," *IOP Journal of The Electrochemical Society*, vol. 167, no. 12, 2020.

[7] J. Vetter, P. Novak, M. Wagner, C. Veit, K.-C. Maller, J. Besenhard, M. Winter, M. Wohlfahrt-Mehrens, C. Vogler, and A. Hammouche, "Ageing mechanisms in lithium-ion batteries," *Journal of Power Sources*, vol. 147, no. 1, pp. 269–281, 2005.

[8] A. Millner, "Modeling lithium ion battery degradation in electric vehicles," in *2010 IEEE Conference on Innovative Technologies for an Efficient and Reliable Electricity Supply*, 2010, pp. 349–356.

[9] A. Bocca, A. Sassone, A. Macii, E. Macii, and M. Poncino, "An aging-aware battery charge scheme for mobile devices exploiting plug-in time patterns," in *IEEE International Conference on Computer Design (ICCD)*, 2015, pp. 407–410.

[10] B. Xu, A. Oudalov, A. Ulbig, G. Andersson, and D. S. Kirschen, "Modeling of lithium-ion battery degradation for cell life assessment," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1131–1140, 2018.

[11] G. Vennam, A. Sahoo, and S. Ahmed, "A survey on lithium-ion battery internal and external degradation modeling and state of health estimation," *Journal of Energy Storage*, vol. 52, p. 104720, 2022.

[12] T. Roth, L. Streck, N. Mujanovic, M. Winter, P. Niehoff, and A. Jossen, "Transient self-discharge after formation in lithium-ion cells: Impact of state-of-charge and anode overhang," *Journal of The Electrochemical Society*, vol. 170, no. 8, p. 080524, aug 2023. [Online]. Available: https://dx.doi.org/10.1149/1945-7111/acf164

[13] W. He, Z. Li, T. Liu, Z. Liu, X. Guo, J. D. X. Li, P. Sun, and W. Ming, "Research progress and application of deep learning in remaining useful life, state of health and battery thermal management of lithium batteries," *Journal of Energy Storage*, vol. 70, 2023.

[14] H. Liu, Q. Xiao, Z. Jiao, J. Meng, Y. Mu, K. Hou, X. Yu, S. Guo, and H. Jia, "Lightgbm-based prediction of remaining useful life for electric vehicle battery under driving conditions," in *IEEE Sustainable Power and Energy Conference (iSPEC)*, 2020, pp. 2577–2582.

[15] K. S. S. Alamin, Y. Chen, E. Macii, M. Poncino, and S. Vinco, "A machine learning-based digital twin for electric vehicle battery modeling," in *IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, 2022.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] W. Zhang, X. Li, and X. Li, "Deep learning-based prognostic approach for lithium-ion batteries with adaptive time-series prediction and on-line validation," *Measurement*, vol. 164, p. 108052, 2020.

[18] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.