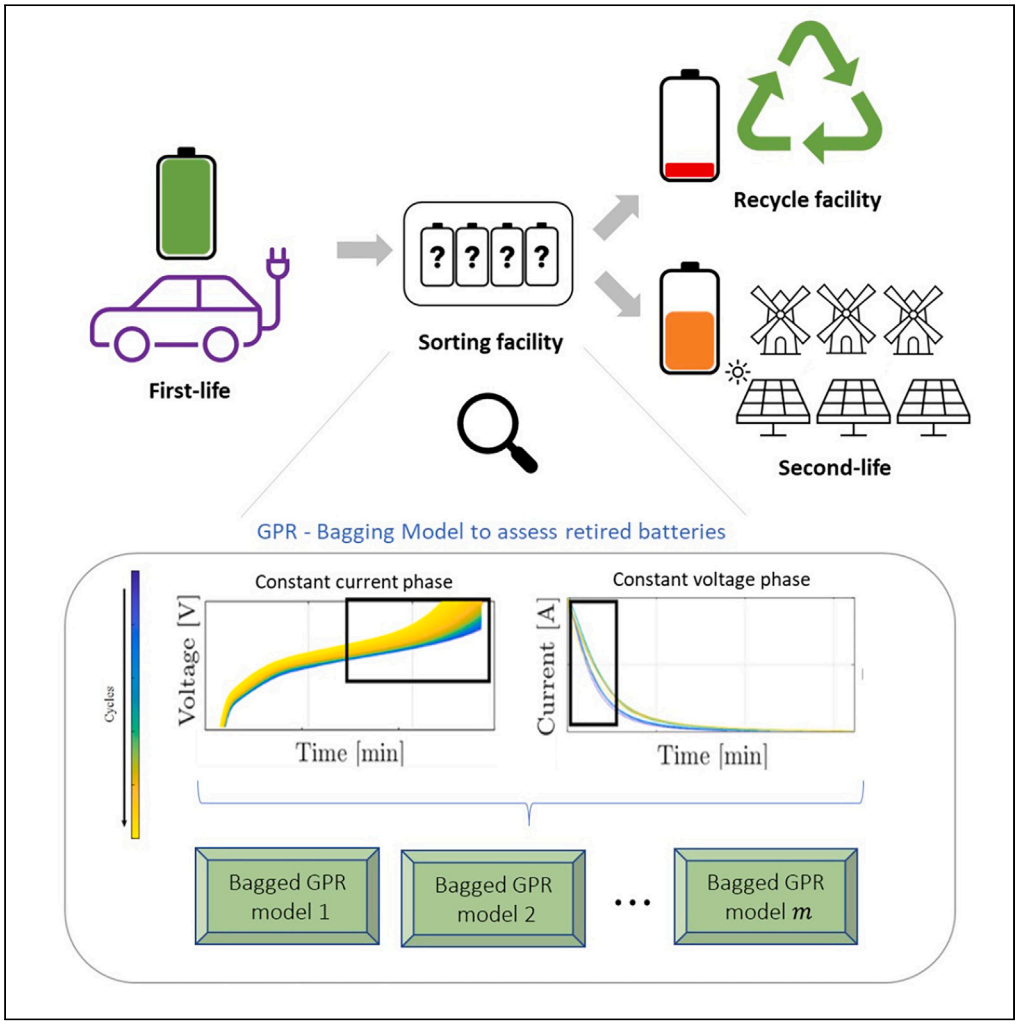
iScience

CellPress
OPEN ACCESS

Article

Evaluating the feasibility of batteries for second-life applications using machine learning



Aki Takahashi,
Anirudh Allam,
Simona Onori

sonori@stanford.edu

Highlights

Accurate and fast method to assess health of retired EV batteries

Proposed ensemble based GPR with bagging model has lower computational burden

Model trained and validated on different battery chemistries and operating conditions

Mean of RMSEs is observed to be less than 1.48%

Takahashi et al., iScience *26*, 106547
April 21, 2023
https://doi.org/10.1016/j.isci.2023.106547

iScience

CellPress
OPEN ACCESS

Article

Evaluating the feasibility of batteries for second-life applications using machine learning

Aki Takahashi,[1] Anirudh Allam,[1] and Simona Onori[1,2,*]

SUMMARY

This article presents a combination of machine learning techniques to enable prompt evaluation of retired electric vehicle batteries as to either retain those batteries for a second-life application and extend their operation beyond the original and first intent or send them to recycling facilities. The proposed algorithm generates features from available battery current and voltage measurements with simple statistics, selects and ranks the features using correlation analysis, and employs Gaussian process regression enhanced with bagging. This approach is validated over publicly available aging datasets of more than 200 with slow and fast charging cells, with different cathode chemistries, and for diverse operating conditions. Promising results are observed based on multiple training-test partitions, wherein the mean of Root Mean Squared Percent Error and Mean Percent Error performance errors are found to be less than 1.48% and 1.29%, respectively, in the worst-case scenarios.

INTRODUCTION

Lithium-ion battery technology is used in a wide variety of industries because of its superior energy and power density among available electrochemical devices, as well as their cycling capability. The loss in performance observed over time is because of multiple degradation phenomena, including loss of active material, corrosion, passivation, lithium plating, and solid electrolyte interphase layer growth.[1]

On retirement from their first life application, batteries are sent to warehouses where they are piled up and stored waiting to be screened. Their health history is unknown and therefore it is critical to be able to assess the level of deterioration to decide whether the battery can be safely utilized in later applications such as backup power, residential storage, EV charging, and utility scale storage[2] to capture the maximum value, both economic and environmental, or sent to recycling facilities. According to a McKinsey report, the supply of second-life Electric Vehicle (EV) batteries could surpass 200 GW-hours per year by 2030, with the potential to meet half of the forecast global demand for utility-scale energy storage in that year.[3]

Battery state of health (SOH) is a metric for diagnosing battery degradation during testing and operation. Although many unique invasive measurements to diagnose the health of the cell are possible during the testing phase in a laboratory setting, instead during real-time operation, only non-invasive measurements in the form of temperature, voltage and current are accessible for diagnostic purposes, which are typically done via model-based or data-driven approaches. Recent research work has shown promising results in assessing battery health, in terms of impedance, capacity, or power, via data-driven approaches.[4–6] For example, neural network models independently find relationships between degradation indicators and battery health,[7,8] in support vector regression and random forest algorithms, the differences between features are defined to learn the relationship between features and the response.[9,10] Gaussian process regression (GPR), a nonparametric approach to regression, has become an increasingly popular method because of its interpretable outputs and the available prediction uncertainty.[4,11,12] There are different approaches to GPR, namely, where training is done on some early cycles and predictions are made thereafter,[13,14] or where models are trained on some cells and then predictions are made on different test cells.[12,15] The former approach assumes that the battery undergoes similar use cases in both early and later cycles, making it an inappropriate approach for repurposing battery applications. On the other hand, with the latter approach the prediction algorithm can be extrapolated to cells with similar operating conditions. Currently, there is lack of a universal, chemistry agnostic approach to battery

[1]Department of Energy Science and Engineering, Stanford University, Stanford, CA 94305, USA

[2]Lead contact

*Correspondence: sonori@stanford.edu

https://doi.org/10.1016/j.isci.2023.106547

Table 1. Summary of different datasets used in this work and their experimental conditions

Dataset	Cathode	Form factor	Nominal capacity [Ah]	Temperature [°C]	Aging protocol	Charging C-rate [1/h]	Voltage range [V]	Number of cells
Birkl.[16]	NMC	Pouch	0.74	40	Artemis driving schedule	1	2.7–4.2	8
Bole et al.[17]	LCO	18650	2.1	Room	Statistically random discharge	1	3.2–4.2	20
Severson et al.[5]	LFP	18650	1.1	30	2-step CC-CV charge, CC discharge	1–6	2–3.6	124
Attia et al.[20]	LFP	18650	1.1	30	4-step CC-CV charge, CC discharge	4–8	2–3.6	45

health assessment and life prediction, because of the nonlinear nature of battery degradation dynamics.[5,16,17] In addition, scalability of data-driven models must be addressed to prepare for the influx of battery data.[18]

In this article, a data-driven approach which combines ensemble methods with GPR is proposed to quickly estimate overall battery capacity. The model uses features from voltage and current information in a limited window of the charge profile. Time is not used explicitly, which, while valuable[13], may not be practical when the battery undergoes incomplete charging or discharging processes. In our approach, we use bagging for the ensemble learning. This work highlights the importance of selecting statistical features over a time horizon that embeds degradation information to develop robust data-driven models that accurately and quickly assess battery health.

Contributions of this work are as follows: (1) An ensemble based GPR model is proposed to estimate overall battery health – in terms of capacity loss - quickly and efficiently, (2) features from voltage and current measurements are proposed for fast screening, (3) the validation of the model is tested on multiple datasets of lithium-ion batteries, each with different experimentation and aging processes.

RESULTS

Battery health indicators (HI)

Health indicators to describe the battery's lifetime performance include, SOH_C and SOH_E, defined as a ratio of the battery's actual capacity and actual energy (at any given point in time) relative to the nominal capacity (Q_{nom}) and energy (E_{nom}), respectively, measured at the beginning of life, as follows:

$$SOH_C(t) = \frac{Q(t)}{Q_{nom}} \cdot 100\% \quad \text{(Equation 1)}$$

$$SOH_E(t) = \frac{E(t)}{E_{nom}} \cdot 100\%, \quad \text{(Equation 2)}$$

where $Q(t)$ and $E(t)$ are the capacity and energy at time t. Although a deterioration of 20% in SOH_C is the industry standard for battery end of life, SOH_E is supplemented with changes in the battery voltage. As the battery is used, internal resistance increases because of the aging mechanisms, resulting in a shift of the capacity-voltage curve over time.[19] Thus, SOH_E deterioration can also be used to identify the increase in resistance of the cell.

Data processing

Datasets from publicly available repositories, summarized in Table 1, were used in this work (see Experimental Procedures) on proper processing. Despite the different operating conditions, chemistry, and aging trajectories of the batteries as noticed in Figures 1A–1C, a strong linear correlation among SOH_C and SOH_E is observed in Figures 1D–1F for NMC, LCO, and LFP, respectively. Therefore, the analysis that follows is conducted using the capacity-based HI, SOH_C, as aging metric, denoted as SOH henceforth.

Feature generation is carried out on the charge cycles because constant current and/or constant voltage (CC-CV) data is generally standardized as opposed to the discharge cycle, which is heavily determined by the intended application and is user-dependent. It is observed that the CC and CV regimes are key

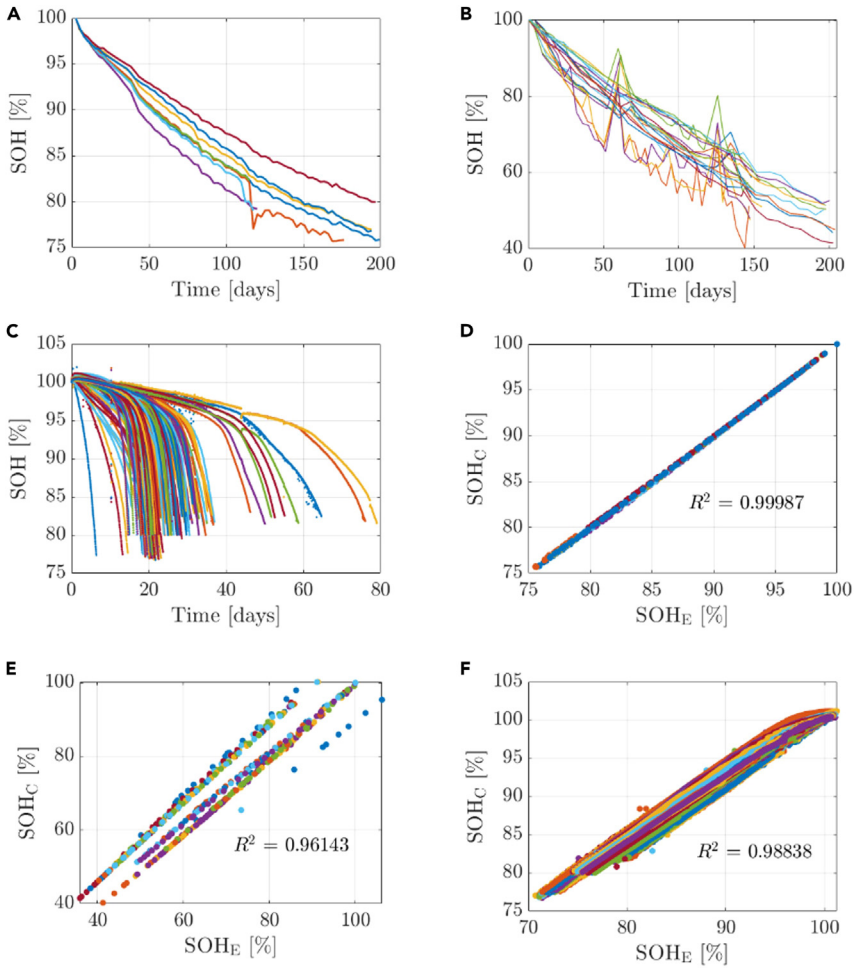


Figure 1. SOH and SOE for aged cells of different chemistry

(A–C)**SOH** versus time for different cell chemistry that operate on a different range of days as well as aging procedures for NMC,[16] LCO,[17] and LFP,[5,20] respectively.

(D–E) Correlation and change in and for all cells for NMC, LCO, and LFP, respectively. A linear relationship can generally be observed for all three groups of cells, with a high correlation coefficient for all three datasets, irrespective of the testing operating conditions. For NMC, the linear fit is almost perfect, which suggests that SOH_C and SOH_E are interchangeable for this chemistry. For LCO, correlation is high, albeit lower than the other two chemistries. For (F) it should be noted that the capacity-voltage feature used to predict capacity degradation was a measure of SOH_E, which can be observed to decrease faster compared to SOH_C.[5] However, SOH_E was not observed to deteriorate faster than SOH_C in the other two datasets.

regions of the battery operation where deterioration is highly perceptible, and therefore extracted to generate features. To that end, for the CC and CV processes, because of the diverse electrode chemistry and aging conditions of the datasets, observations from the certain pre-selected regions of the voltage and current data utilized. For the LFP cells, 30 s worth of lower portion of voltage data during CC and upper portion of current data during CV are considered.; Data spanning 3.65V–4.2V during CC for 25 min or less (as low as 70 s depending on the SOH of the cell) and the whole current profile worth two to 3 h during the CV phase are considered for LCO cells. Finally, 1-h worth of voltage data during the CC phase and no CV phase are considered for NMC cells. Because the datasets in Table 1 are characterized by different charging rates, the measurements used for generating features during the CC-CV phase are sampled or selected accordingly to maintain consistency. The sampling rate for the features is an important design parameter as excessive data collection can result in a computationally prohibitive model. For higher C-rates, the battery reaches the maximum voltage sooner, so a limited number of measurements are available to sample. However, for lower C-rates, it is not necessary to sample measurements frequently since the

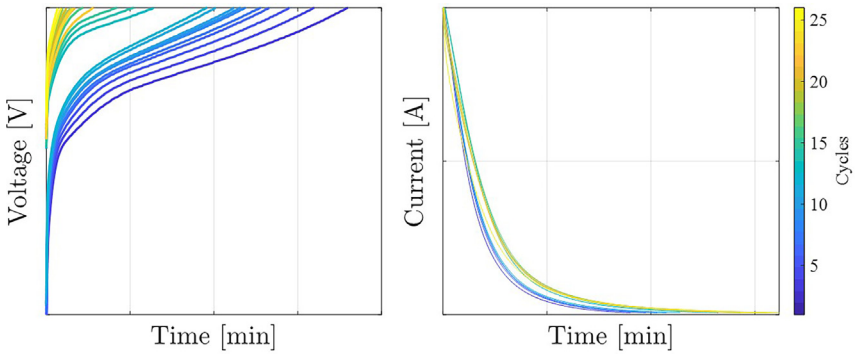


Figure 2. Current and voltage data for feature generation

Voltage curve during CC (left) and current curve during CV (right) for a sample cell from the LCO dataset[17] for feature generation

change in the voltage and current curves occur slowly, and high frequency sampling would lead to a large amount of data for the model to train on. The NMC and LCO datasets have >1 h charge durations with low current. For these datasets, we take either voltage or current measurements every 10 s to get sufficient data until end of CC or CV. As for the LFP dataset, where the C-rates are much higher, we take measurements every 2 s in a 30 s window for CC and 60 s for CV to capture. This also allows us to demonstrate the capability of this approach in both slow and fast charge regimes, as well as a limited window of measurements. Figure 2 illustrates the voltage vs. time and current vs. time curves over many CC-CV cycles of a representative cell in the LFP dataset.[5,20] During the CV portion of the charge profiles at the known voltage, current signals are also analyzed for feature extraction. Note that for the NMC dataset,[16] the cells did not undergo CV, hence CV is not considered in this case.

Feature generation and selection

For each of the CC and CV pair of datasets, 7 simple statistical measures are used to characterize the data, which are summarized in Figure 3A. The statistics are applied to voltage during CC and current during CV, respectively. The main benefit of using said statistics is their simplicity in calculation during operation. In addition, they do not rely on derivatives, as instead seen in incremental capacity and differential voltage analysis methods, which can introduce noise without using a slow characterization procedure.[21] Lastly, the features are shifted by subtracting the same metric calculated at the start of life.

Once the features are generated, a feature selection technique is used to eliminate redundant and noisy features. This reduces the dimensionality of the model and allows for efficient and accurate estimation of SOH. There are a variety of selection techniques, such as filtering, wrapping, and fusion, with each varying in purpose and complexity.[22] Here, we use the fastest filter method via correlation coefficient, such as Pearson or Spearman, to quickly determine which features correlate best with the predicted variable.

The Pearson correlation coefficient assumes a Gaussian distribution for the vectors and a linear relationship between two variables. This assumption is sometimes not true because of the nonlinear nature of capacity degradation. Hence, the Spearman coefficient is found to be more suitable and therefore used. The equation is as follows:

$$\rho = 1 - \frac{6\sum_{i=1} d_i^2}{N(N^2 - 1)} \qquad \text{(Equation 3)}$$

where ρ is the Spearman's rank correlation coefficient, i is the i-th value in the corresponding feature vector, d is the difference between the two sorted ranks of features and SOH, and N is the number of data points in the dataset. This expression looks for a monotonic relationship between features and responses, meaning the magnitude of ρ increases as the response either increases or decreases as the feature increases. Once the magnitude of ρ was found for all features, we select 10 features with highest value, while also looking for dissimilar features ($|\rho|<0.8$ between features), as done in.[15] An illustration is shown in Figure 3B for the LFP

A

Statistics	Meaning
Mean	Average over recorded values
Median	Central value of distribution
Sum	Adding all recorded values
Standard Deviation	Measure of spread with respect to mean
Variance	Measure of spread
Kurtosis	Measure of deviation from normal distribution at extreme values
Interquartile range	Measure of spread between 25% and 75% percentile

B

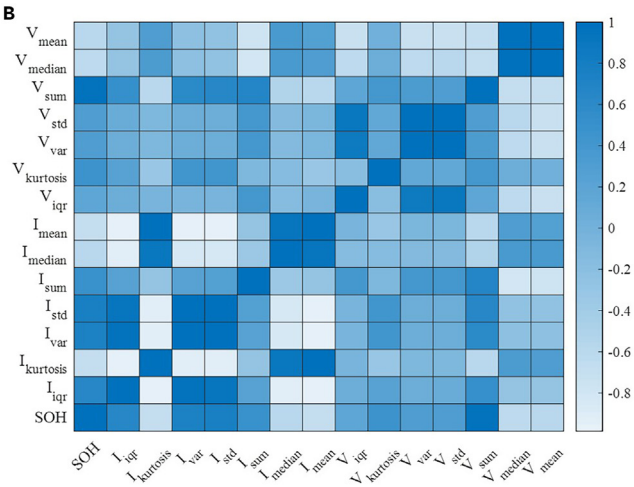


Figure 3. Feature generation and selection

(A) List of statistics applied on the voltage and current measurements. The selected statistics are chosen for their ease of calculation, and their definitions are briefly explained.

(B) Heatmap displaying the relationship between the 14 features and SOH for a sample subset of 120 cells from the LCO dataset. A value for Spearman coefficient $\boldsymbol{\rho}$ close to 1 represents a stronger correlation, or similarity, between the two metrics. Of these candidates, the best combination of features is selected based on their uniqueness and strong predictive capability.

dataset. An example of the relationship between features and SOH for representative cells is shown in Figures S1–S3 for NMC, LCO, and LFP cells, respectively, in the Supplemental Information section.

Scalable Gaussian process regression

GPR is a nonparametric probabilistic algorithm used to make predictions based on Gaussian distributions of features and response.[23] Specifically

$$GPR\ (\mu(x), k(x, x')), \qquad \text{(Equation 4)}$$

is a Gaussian process (GP) with mean function $\mu(x)$ and covariance function $k(x,x')$, where x and x' are training inputs. If each set of the inputs has a joint Gaussian distribution, then the whole set of inputs forms a joint Gaussian distribution. This means that the predicted response has a distribution, with M and k defining the response and its uncertainty. Further details are provided in the Experimental Procedures section.

Although GPR models are effective, they suffer from lack of scalability, since for model training there are $O(n^3)$ and $O(n^2)$ computation and storage requirements (in big-O notation), whereas the cost of a single prediction is $O(n^3)$, where n is the number of datapoints being trained on. This is a significant problem for second-life battery health assessment and other practical applications where large numbers of batteries need to be processed[18] in a short period of time. To mitigate this, one approach is to use bagging (bootstrap aggregating), an ensemble learning technique used mainly for random forest algorithms but transferable to other machine learning algorithms, including GPR.[24] From a large dataset m, different bags of size n are created by randomly sampling from the dataset with replacement, meaning some examples can be selected more than once. Once the m sampled datasets are created, a GPR model is trained on each dataset to create m models. For subsequent predictions then, these models are combined using some aggregating technique, such as a simple average or weighted average of the predicted output. It should be noted that this is different from Bayesian committee machines, which instead uses scalable GPs, where all of the training data is split into subsets instead of random sampling with replacement.[25]

Bagging is attractive for its reduction of variance because it creates multiple models instead of relying on a single model, which means if one of the models is overfitting the training data then the other models can work together to improve the prediction accuracy of the test dataset. It also allows for parallel computing because each of the bagged models undergoes training and prediction separately. Each of the GPR models can make their own predictions, and when merged, their net combined performance improves noticeably.[24] We use a weighted average to account for the predictive capability of each of the bags (see Experimental Procedures section).

The computational time needed to train the bagged model can be reduced significantly by using bags smaller than the whole dataset, as well as using parallel computing features. In general, the bagged approach reduces the amount of training time because there is no need to perform optimization on a large feature set. With bagging, computational burden becomes $O(mn^3)$, meaning that with smaller selection of m and n GPRs can be simplified exponentially while still achieving strong accuracy.[26] Specifically, the data used by the models can be compared using the following factor reduction of data (FRD) metric:

$$\mathrm{FRD} = \frac{N}{m \cdot n} \qquad \text{(Equation 5)}$$

where N is the number of datapoints in the original dataset. For constant N, the FRD can be increased by changing m and n. A larger FRD leads to a significant increase in model speed for training and prediction when combined with parallel computing.

Model performance

To measure model performance, an approximate 70-30 training-test split is randomly created based on individual cells. In the context of battery aging with data collected over a long period of time, the model is trained on the entire aging trajectory of 70% of cells and tested on the remaining 30% of cells. In addition, a GPR model without bagging, referred to as the baseline model, is developed to highlight the performance improvement of the proposed GPR with bagging approach.

Feature selection is conducted on the training set and features with low correlation are eliminated, as observed in the heatmap in Figure 3B. Then, the bagged models are created and trained. Finally, the models perform capacity assessments on all available test data, and the results are aggregated using an averaging rule. An example of the assessment process is shown in Figure 4. This process is repeated several times for smaller datasets to simulate a greater number of tested cells. To evaluate model performance, we use Root-Mean-Square Percent Error (RMSPE) and Mean Percent Error (MPE). This combination allows us to report model performance based on the presence of outliers and overall model performance. Formal definitions of RMSPE and MPE are described in the Experimental Procedures section.

SOH estimation results

For all datasets, m was initially selected to be half of the number of training cells, with sample size n being close to the average number of characterization cycles for each cell, then modifications were made based on performance. The selected values of m and n for each chemistry is shown in Table 2.

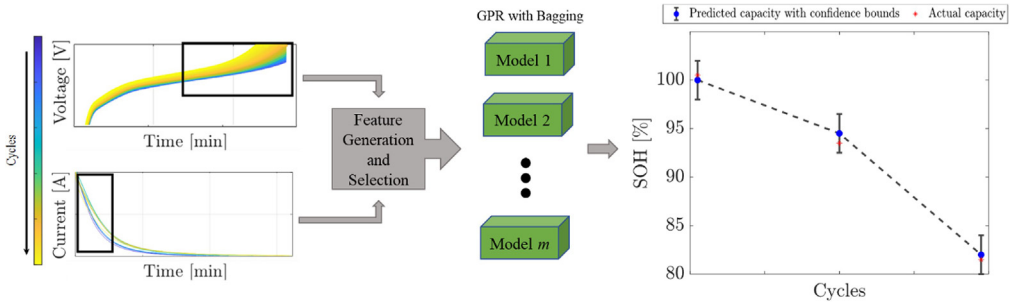


Figure 4. Schematic of ensemble-based SOH estimation based on GPR with bagging

Current and voltage data collected over a single charging curve are extracted from a pre-specified window and processed to generate features. The features are passed through the bagged models to generate a SOH estimation with uncertainty bounds, which over time generates a capacity fade envelope.

The distribution of SOH error for all individual training and test datapoints are shown in Figures 5A–5E compiles the MPE and RMSPE of individual cells into boxplots, Table 3 lists the model performance metrics in terms of median and mean of RMSPE and MPE Boxplots in Figure 5 are used instead of conventional histograms to demonstrate performance on individual cells in addition to compare the different chemistries. Error mean and median are reported to describe the distribution, in comparison to the baseline in Table 3. Average training and prediction time for the 70-30 splits are measured using an i7-9750H CPU @ 2.60 GHz and the parallel computing feature in MATLAB with 4 workers.

LCO dataset

14 cells are selected for training and 6 were used for testing in each iteration of assessment. Assessments are made to the end of available data (during CC phase: 3.65V to end of CC, and for CV phase: full curve). We select 7 bags with sample size 30 to reduce the computational time to train the algorithm. 300 assessments are made based on 50 iterations. The median RMSPE and MPE are below 1.5% as shown in Figures 5A and 5B. Training time reduction of 2 orders was also found for this dataset. In this dataset the sum of the voltage values was an important feature as shown in Figure S1 in the Supplemental Information section. This may be because of the long charge characterization, in that the CC regime was significantly longer and hence the feature generation captured more voltage points to add up in early cycles of the battery. This dataset shows the capability of SOH estimation far past the standard end of life of 80%, as data is available until 50% SOH.

NMC dataset

This dataset has the least cells; however, the 70-30 split still turns out to generate valid results with 6 cells used for training and 2 cells for testing in each iteration. 3 bags were created, each with 20 sample points, which reduced the training data to approximately $^1/_8$ of its original size for each iteration, resulting in an exponentially faster solution. Because there are not many cells the process is repeated 150 times to get a total of 300 estimates. The results in Table 3 indicate that the bagged GPR estimation is more accurate than the baseline. Table 3 also reports that there was a training time reduction of 2 orders of magnitude compared to the baseline model. Some of the higher RMSPE errors are likely because of the sudden drop in SOH past 50% in one of the cells; MPE is significantly lower than the other datasets because the effect of this single anomaly is small. The low error can also be attributed to the high correlation coefficient (>0.95) between the CC features and SOH for cells, which is shown in Figure S2 in the Supplemental

Table 2. Values of *m* and *n* used in model implementation for each cathode chemistry

Chemistry	*m*(number of models)	*n*(sample size)	FRD (factor reduction)
NMC	3	20	8
LCO	7	30	2
LFP	20	200	23

FRD in training data is found by dividing the typical amount of training data by the number of examples used in the bagged approach. This leads to an exponential decrease in training time.

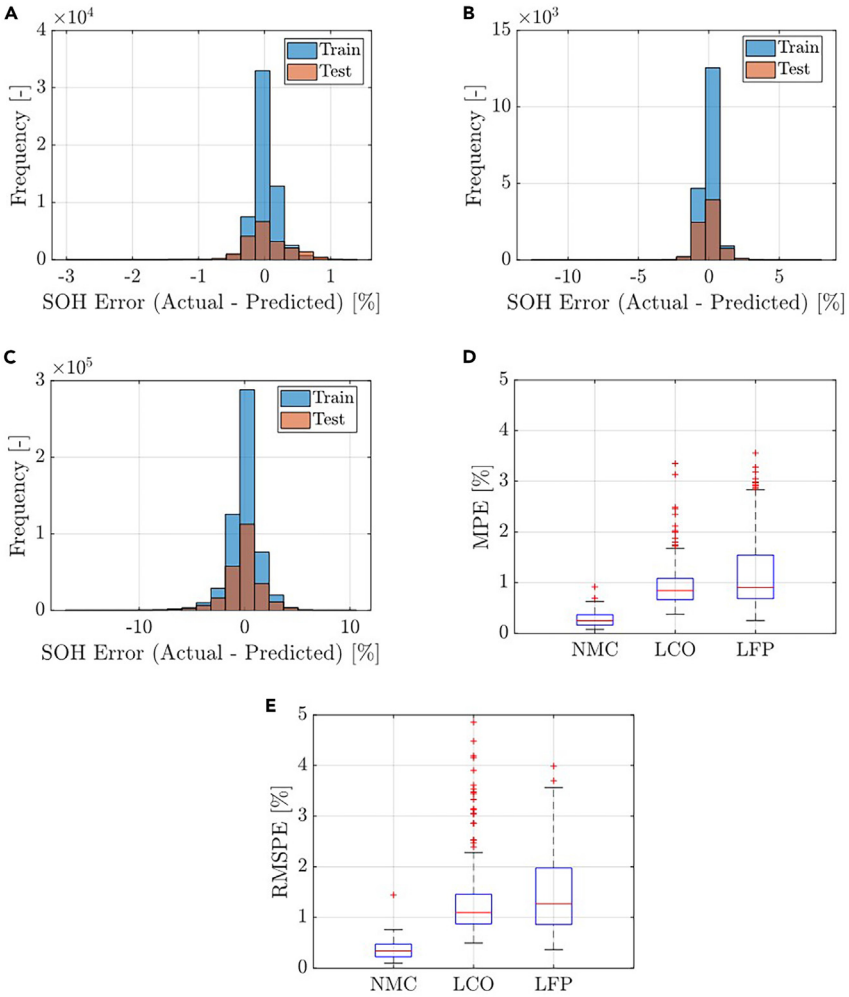


Figure 5. SOH estimation error distribution and statistics

(A–C) Distribution of SOH error with 20 bins for all individual training and test datapoints for NMC, LCO, and LFP cells, respectively.

(D and E) Boxplots show the median, 25th, and 75th percentiles, and the outliers (points past the interquartile range) of 300 iterations of test cells for MPE and RMSPE, respectively.

Information section. It should be reiterated that these cells do not undergo a CV regime during the charging process, meaning that while we look for the 10 best features there are only 7 features used.

LFP dataset

Because this dataset has the most data, the benefits of implementing an ensemble-based SOH estimation framework based on GPR with bagging are observed from reducing the computational time without sacrificing significant accuracy. The baseline model was impossible to store in memory with the original training data, so 2,000 datapoints were instead randomly sampled out of about 90,000 datapoints for each training iteration. Four cells were also anomalous in measurements, so they were excluded from analysis, leaving 165 cells. A training test split uses 115 cells for training and 50 cells for testing. Training data for the bagged models was reduced to less than 1/20 of a typical training set size. There is a training time reduction of 2 orders of magnitude compared to the baseline model, as shown in Table 3. For each test set there are 50 cells, so with 6 iterations there are 300 assessments made, whose error is shown in Figures 5D and 5E. The median MPE was under 1%, which means that for all the three chemistries the ML model was able to achieve under 1% MPE based on the median. The error is slightly higher compared to the other

Table 3. Comparison of the median and mean of RMSPE and MPE, and training/prediction times for all three datasets between baseline and bagged

	NMC		LCO		LFP	
Metric	Bagged	Baseline	Bagged	Baseline	Bagged	Baseline
$RMSPE\text{–}_{median\ (\%)}$	**0.3384**	0.35132	**1.099**	1.109	**1.266**	1.2922
$RMSPE_{mean\ (\%)}$	**0.2464**	0.39763	**1.277**	1.5409	1.475	**1.4624**
$MPE_{median\ (\%)}$	0.286	**0.26088**	**0.839**	0.88566	**0.907**	0.93055
$MPE\text{-}_{mean\ (\%)}$	**0.28486**	0.29896	**0.925**	1.194	1.286	**1.0958**
Training Time (s)	**0.024061**	2.0465	**0.1613**	3.1871	**4.1123**	123.8961
Prediction Time (s)	**0.021394**	0.067256	0.031035	**0.028035**	**4.2562**	4.2789

All datasets are tested to have a total of 300 test cells based on randomized bags. Training time and prediction time are reported as averages of multiple 70-30 splits of the data.

two datasets, which is to be expected given the wide variety of fast charging protocols and large size of the dataset.

Effect of *m* and *n*

In bagging techniques, there is a convention to choose *n* to be the same size as the training data whereas *m* is arbitrarily selected.[27] However, for GPs it is impractical to use large datasets because of the increasing training time and storage space, as well as the diminishing improvements in accuracy.[26] Hence, we analyze the effect of changing *m* and *n* in a set domain to see how much improvement in speed is achieved. First, the sample size *n* is selected, then *M* bags are created, where *M* is the greatest number of bags used. Then, we train all *M* models and use small batches of size *m* out of the *M* models for assessment. This methodology allows for comparisons between performance of bagged models with the same *n* but different *m*, as the same subsamples are used and performance is measured largely based on *m*, not random sampling. The process is repeated 10 times for each dataset.

Because the datasets are of different sizes, the number of bags and sample size are different for each chemistry, which is outlined in Table 4. Beyond the values selected, we notice diminishing changes in performance. The results for MPE and RMSPE for different combinations of *m* and *n* are observed in Figure 6.

In general, it can be observed in Figure 6 that increasing *m* and *n* decreases the average MPE and RMSPE. Of interest, the NMC dataset, which only contains 8 cells, had worse estimation with larger bags, albeit the increase in error is not significant. This may be because of the high predictive capability of the CC features and thus, larger number of bags results in making estimations more overconfidently. Increasing the number of bags past 10 did not have significant improvements, meaning that it is possible to make good SOH assessments with as little as 10 sample points each in 10 bags, which reduces the amount of training data by a factor of 4 and exponentially decreases computational time.

With the LCO dataset, the average RMSPE and MPE error quickly decrease as the number of bags and sample size increase. The data needed to make better estimation was relatively more than the NMC dataset, which is to be expected because of the variable aging procedures. The average MPE is below 0.8% and RMSPE was close to 1% even with a relatively small *m* and *n* of 10 and 80, for example, with less memory usage compared to using all data (0.802 megabytes vs. 1.13 megabytes based on data from 14 cells). For LFP, using $n = 850$ and $m=20$ results in error less than 1.5% in average $RMSPE_{mean}$ and 1.18% in average MPE_{mean}.

DISCUSSION

Battery data is becoming more important and abundant than ever because of the rapid growth of the lithium-ion battery industry. This calls for an immediate need of efficient approaches to diagnose the SOH of batteries, especially for repurposing purposes. The proposed novel method consists of (1) using simple statistics to describe a small window of voltage and current data, (2) fast feature selection, and (3) bagging multiple GPR models.

Table 4. Range of *m* and *n* used for the different datasets

Dataset	*m*	*n*
NMC	2 to 100	10 to 150
LCO	2 to 50	10 to 150
LFP	2 to 80	50 to 1000

Capacity estimation is made based on pure current and voltage charging data from datasets of diverse cathode chemistry undergoing different aging mechanisms. This work highlights the opportunity to develop data-driven based algorithms for quick and accurate estimation of capacity, which can find use in sorting retired batteries for second life applications despite their cycling history. The algorithm is an ensemble method containing GPR with bagging that outperforms conventional GPR based models both in terms of accuracy and computational time. It should be noted that the regions of voltage and current data used by the model to estimate the health can be adjusted. Furthermore, the feature generation and ranking or selection process should work as presented, provided that within the selected region, changes in voltage and current are visible over time. Aging assessment of LCO and NMC dataset showed to be more accurate likely because of the relatively similar degradation over time between cells.

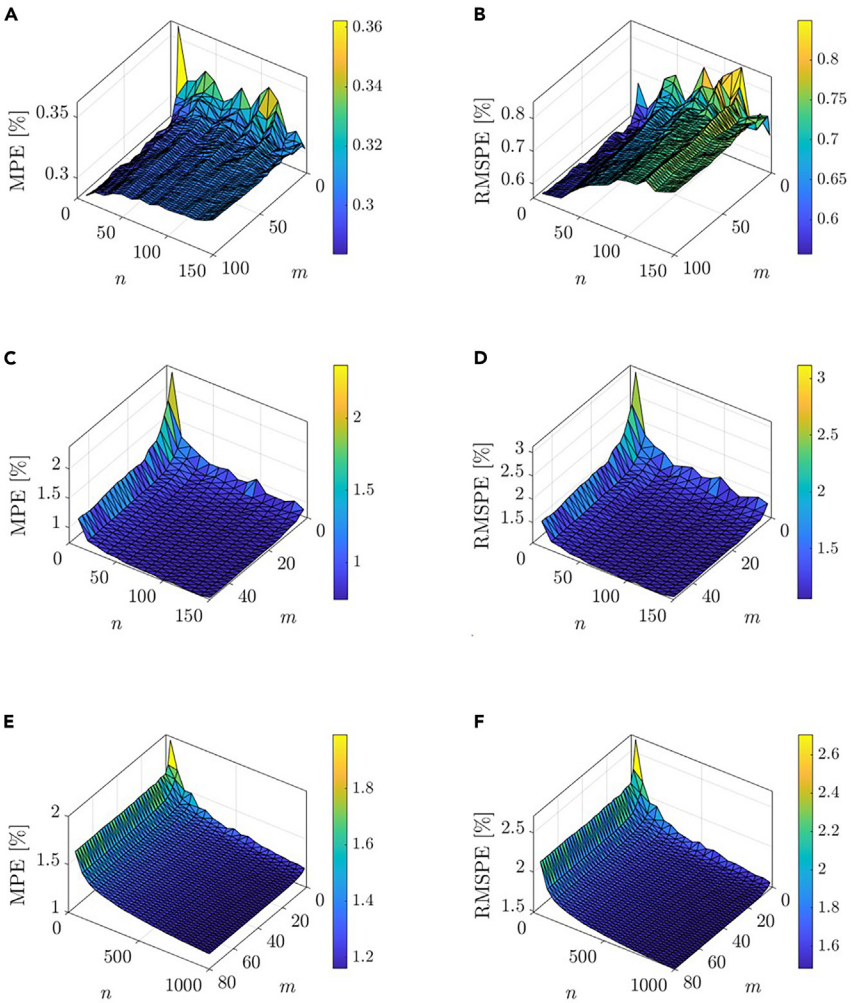


Figure 6. SOH estimation errors as a function of bag sizes and sample size

MPE_{mean} and $RMSPE_{mean}$ as a function of different bag sizes and sample size for (A) (B) NMC, (C) (D) LCO, and (E) (F) LFP, respectively. The color bar corresponds to the error value. The plotted error is an average of 10 iterations.

CellPress
OPEN ACCESS

Based on multiple training-test partitions, average and median RMSPE and MPE performance errors are found to be less than 1.48% and 1.29%, respectively. A consideration to be made also is the trade-off with uncertainty. Using similar data can cause overconfidence in models, while on the other hand, using not enough data will lead to loss in accuracy. The search space for *m* and *n*, parameters of the bagged method, provides guidance on this trade-off because it is possible to easily observe the change in performance metrics.

Ultimately, the bagged approach can save on memory and processing power even with a rich amount of data. This can be expanded to practical applications because of the way the data is collected on the charging curve and serves as an effective diagnostic technique to evaluate feasibility of second-life applications. We acknowledge that there are other ways to scale Gaussian processes in the literature, such as the Bayesian committee machine or approximating the kernel function, and these introduce interesting areas that can be explored in the context of battery degradation.[18,25,26]

LIMITATIONS OF THE STUDY

For future work, the scalability of data-driven models, in addition to accurate and safe health assessment is a key design parameter for transitioning models to industrial applications and achieving sustainability goals. Moreover, a chemistry-agnostic GPR with bagging algorithm can be designed as an extension of this work using richer datasets representative of real-world EV usage condition.[28]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Datasets
  - Nickel manganese cobalt oxide (NMC)
  - Lithium cobalt oxide (LCO)
  - Datapreprocessing
  - Gaussian process regression details
  - Predictions with boostrap aggregating
  - Model performance metrics

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106547.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightful comments.

AUTHOR CONTRIBUTIONS

A.T. trained, validated and implemented the bagged GPR model. A.T., A.A., and S.O. conceived the idea and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 17, 2022
Revised: January 9, 2023
Accepted: March 28, 2023
Published: March 31, 2023

REFERENCES

1. Vetter, J., Novák, P., Wagner, M.R., Veit, C., Möller, K.C., Besenhard, J.O., Winter, M., Wohlfahrt-Mehrens, M., Vogler, C., and Hammouche, A. (2005). Ageing mechanisms in lithium-ion batteries. J. Power Sources *147*, 269–281.

2. Moy, K., Lee, S.B., Harris, S., and Onori, S. (2021). Design and validation of synthetic duty cycles for grid energy storage dispatch using lithium-ion batteries. Advances in Applied Energy *4*, 100065.

3. Engel, H., Hertzke, P., and Siccardo, G. (2019). Second-life EV Batteries: The Newest Value Pool in Energy Storage (McKinsey & Company).

4. Hu, C., Jain, G., Tamirisa, P.A., and Gorka, T. (2014). Method for estimating capacity and predicting remaining useful life of lithium-ion battery. In 2014 International Conference on Prognostics and Health Management, pp. 1–8.

5. Severson, K.A., Attia, P.M., Jin, N., Perkins, N., Jiang, B., Yang, Z., Chen, M.H., Aykol, M., Herring, P.K., Fraggedakis, D., et al. (2019). Data-driven prediction of battery cycle life before capacity degradation. Nat. Energy *4*, 383–391.

6. Xing, Y., Ma, E.W., Tsui, K., and Pecht, M.G. (2013). An ensemble model for predicting the remaining useful performance of lithium-ion batteries. Microelectron. Reliab. *53*, 811–820.

7. Yang, D., Wang, Y., Pan, R., Chen, R., and Chen, Z. (2017). A neural network based state-of-health estimation of lithium-ion battery in electric vehicles☆. Energy Proc. *105*, 2059–2064.

8. Roman, D., Saxena, S., Robu, V., Pecht, M.G., and Flynn, D. (2021). Machine learning pipeline for battery state of health estimation. Nat. Mach. Intell. *3*, 447–456.

9. Mansouri, S.S., Karvelis, P.S., Georgoulas, G., and Nikolakopoulos, G. (2017). Remaining useful battery life prediction for UAVs based on machine learning. IFAC-PapersOnLine *50*, 4727–4732.

10. Nuhic, A., Terzimehic, T., Soczka-Guth, T., Buchholz, M., and Dietmayer, K. (2013). Health diagnosis and remaining useful life prognostics of lithium-ion batteries using data-driven methods. J. Power Sources *239*, 680–688.

11. Liu, J., and Chen, Z. (2019). Remaining useful life prediction of lithium-ion batteries based on health indicator and Gaussian process regression model. IEEE Access *7*, 39474–39484. https://doi.org/10.1109/ACCESS.2019.2905740.

12. Richardson, R.R., Osborne, M.A., and Howey, D.A. (2017). Gaussian process regression for forecasting battery state of health. J. Power Sources *357*, 209–219.

13. Yang, D., Zhang, X., Pan, R., Wang, Y., and Chen, Z. (2018). A novel Gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve. J. Power Sources *384*, 387–395.

14. Yu, J. (2018). State of health prediction of lithium-ion batteries: Multiscale logic regression and Gaussian process regression ensemble. Reliab. Eng. Syst. Saf. *174*, 82–95.

15. Greenbank, S., and Howey, D. (2022). Automated feature extraction and selection for data-driven models of rapid battery capacity fade and end of life. IEEE Trans. Ind. Inf. *18*, 2965–2973.

16. Birkl, C. (2017). Oxford Battery Degradation Dataset 1 (University of Oxford).

17. Bole, B., Kulkarni, C.S., and Daigle, M. (2014). Adaptation of an Electrochemistry-Based Li-Ion Battery Model to Account for Deterioration Observed under Randomized Use Annual Conference of the PHM Society (Vol. 6, No. 1).

18. Aitio, A., and Howey, D.A. (2021). Predicting battery end of life from solar off-grid system field data using machine learning. Joule *5*, 3204–3220.

19. Schweiger, H., Obeidi, O., Komesker, O., Raschke, A., Schiemann, M., Zehner, C., Gehnen, M., Keller, M., and Birke, P. (2010). Comparison of several methods for determining the internal resistance of lithium ion cells. Sensors *10*, 5604–5625.

20. Attia, P.M., Grover, A., Jin, N., Severson, K.A., Markov, T., Liao, Y., Chen, M.H., Cheong, B., Perkins, N., Yang, Z., et al. (2020). Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature *578*, 397–402.

21. Bloom, I.D., Jansen, A.N., Abraham, D.P., Knuth, J., Jones, S.A., Battaglia, V.S., and Henriksen, G.L. (2005). Differential voltage analyses of high-power, lithium-ion cells: 1. Technique and application. J. Power Sources *139*, 295–303.

22. Hu, X., Che, Y., Lin, X., and Onori, S. (2021). Battery health prediction using fusion-based feature selection and machine learning. IEEE Transactions on Transportation Electrification *7*, 382–398.

23. Rasmussen, C.E., and Williams, C.K. (2009). Gaussian processes for machine learning. In Adaptive computation and machine learning.

24. Chen, T., and Ren, J. (2009). Bagging for Gaussian process regression. Neurocomputing *72*, 1605–1610.

25. Tresp, V. (2000). A bayesian committee machine. Neural Comput. *12*, 2719–2741.

26. Liu, H., Ong, Y., Shen, X., and Cai, J. (2020). When Gaussian process meets big data: a review of scalable GPs. IEEE Transact. Neural Networks Learn. Syst. *31*, 4405–4423.

27. Martínez-Muñoz, G., and Suárez, A. (2010). Out-of-bag estimation of the optimal sample size in bagging. Pattern Recogn. *43*, 143–152.

28. Pozzato, G., Allam, A., and Onori, S. (2022). Lithium-ion battery aging dataset based on electric vehicle real-driving profiles. Data Brief *41*, 107995.

29. André, M. (2004). The ARTEMIS European driving cycles for measuring car pollutant emissions. Sci. Total Environ. *334*, 73–84.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
LCO dataset	Bole et al.[17]	https://www.nasa.gov/content/prognostics-center-of-excellence-data-set-repository
NMC dataset	Birkl.[16]	https://ora.ox.ac.uk/objects/uuid:03ba4b01-cfed-46d3-9b1a-7d4a7bdf6fac
LFP datasets	Severson et al.[5] Attia et al.[20]	https://data.matr.io/1/
Software and algorithms		
Matlab	MathWorks	R2021a
Machine learning model for state of health estimation	This paper	https://doi.org/10.5281/zenodo.7651573

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to the corresponding author, Simona Onori (sonori@stanford.edu).

Materials availability

This study did not generate new materials.

Data and code availability

This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table. All original code has been deposited to Zenodo and is publicly available as of the date of the publication. DOIs are listed in the key resources table. Any additional information required to reanalyze the data reported in this paper is available from the lead contactupon request.

METHOD DETAILS

Datasets

Datasets from online repositories were used in this work. Each of the datasets used unique aging experiments and different battery chemistries – namely, LFP, NMC, and LCO. Aging data used in this paper for all cells is shown in Figures 1A–1C.

Nickel manganese cobalt oxide (NMC)

Th NMC dataset[16] used 8 Kokam SLPB533459H4 NMC cells, which aged via the ARTEMIS driving schedule.[29] Characteristic data is taken for every 100 cycles with a 1C full charge-discharge cycle and a constant-current OCV test. The discharge capacity data is available hence it is used for SOH calculations.

Lithium cobalt oxide (LCO)

The LCO dataset uses statistically random discharge for battery deterioration.[17] The cells used were LG Chem. 18650 LCO cells. The cells were divided into multiple groups of 4, each undergoing a unique, randomized charging and discharging procedure at room temperature. A characteristic charge-discharge cycle at 2A took place periodically, allowing for a comparison between aging procedures. Capacity data is unavailable, so the discharge curve is integrated to obtain SOH measures.

Lithium iron phosphate (LFP)

In this group of cells, 124 commercial high-power LFP A123 APR18650M1A cells were aged via a full two-step fast charging cycle and 4C discharge.[5] Another batch of 45 cells from the same manufacturer underwent a 10 minute 4-step fast charge cycle and 4C discharge.[20] Once cells reached 80% nominal capacity a

1C charging regime followed by constant voltage (CV) charge was used to fully charge the cells in both datasets. There are three different batches of cells and each batch differing by the amount of rest taken between charging and discharging phases.[5] Discharge capacity data is available, so we use these measurements as the expected response.

Datapreprocessing

For data processing on MATLAB, erroneous measurements, such as battery voltage exceeding cutoff values, were first removed. An interpolation scheme is then used on the voltage and current charging curves to obtain regular measurements. The frequency of measurements used (measurements are needed every 2–10 seconds) for the ML model is slower than the frequencies used for lab data, which is often measured less than every second. Thus, a simple linear interpolation was enough for collecting accurate measurements of battery voltage and current. Measurements are then collected at the desired frequency based on the starting voltage or current value.

Gaussian process regression details

The Matern 5/2 covariance function with automatic relevance determination is used for all three datasets since it is able to adapt to different smoothness throughout the regression problem.[12,23] The function is as follows:

$$k_f(x, x') = \sigma_f^2 \frac{2^{1-v}}{\Gamma(v)} \left(\sqrt{2v} \ \frac{x - x'}{\xi} \right)^v K_v \left(\sqrt{2v} \ \frac{x - x'}{\xi} \right), \qquad \text{(Equation 6)}$$

where x and x' are any two points of the same variable, σ_f^2 is the signal variance of the input variables, Γ is the gamma function, K_v is the modified Bessel function, ξ is the length scale, and $v = 5/2$ which simplifies the expression considerably. The hyperparameters that need to be optimized are the length scale, a measure that outlines how far extrapolations can be made, and the signal variance, which determines the spread of the joint distribution. Essentially, optimizing the set of hyperparameters allows the GPR model to fit a set of input variables X to model a Gaussian with respect to the response y. The hyperparameters are optimized by maximizing the log marginal likelihood function, which allows for automatic tradeoff between bias and variance. The equation is as follows:

$$L = -\frac{1}{2}\log\left(\det\left(k_f + \sigma_n^2 I\right)\right) - \frac{1}{2}(y - H\beta)^T \left[K_f + \sigma_n^2 I\right]^{-1} (y - H\beta) - \frac{N}{2}\log(2\pi), \qquad \text{(Equation 7)}$$

where k_f is the selected covariance function, σ_n^2 is the noise variance of the prediction, I is the identity matrix, H is the basis function (assumed to be 1), and β is the coefficient for the basis function. For predictions of variables far from the training set, the state of health prediction will revert to $H\beta$. MATLAB's fitrgp function is used for optimizing the hyperparameters.

Predictions with boostrap aggregating

For bagging, the predictions of each model must be combined to generate a prediction. While weightless aggregation is possible, it has been shown that a weighted prediction is often better.[24] For this work, predicted responses are weighted based on the standard deviation. The weight function is as follows:

$$w_a = \frac{1}{\sigma_a}, \qquad \text{(Equation 8)}$$

where a is the a-th GPR model in the m model set for a particular prediction, σ is the associated error standard deviation. In general, this means that a more unconfident prediction is punished more, and it has been shown that a weighted average performs better.[27] The standard deviation is used instead of the variance since the variance more heavily favors or punishes individual models, which leads to significantly overconfident predictions with large m and n.[26] This expression is multiplied with the corresponding prediction to establish a weighted average, which also has a weighted standard deviation as follows:

$$y_{pred} = \frac{\sum_{a=1}^{m} w_a y_a}{\sum_{a=1}^{m} w_a} \qquad \text{(Equation 9)}$$

$$\sigma_{pred} = \sqrt{\frac{Z\sum_{a=1}^{m} w_a \left(y_a - y_{pred}\right)^2}{(Z-1)\sum_{a=1}^{m} w_a}}, \quad \text{(Equation 10)}$$

where y_{pred} is the aggregated SOH prediction, y_a is the prediction made by the a-th model out of the m-model set, σ_{pred} is the predicted standard deviation and Z is the number of nonzero weights (which is almost always the same as m in our implementation, but is included for greater generalizability).

Model performance metrics

Formally, RMSPE and MPE are defined as follows:

$$\text{RMSPE}\ (\%) = \sqrt{\frac{1}{c} * \sum_{j=1}^{c} \left(\frac{y_{pred,j}}{y_{exp,j}} - 1\right)^2} * 100\% \quad \text{(Equation 11)}$$

$$\text{MPE}\ (\%) = \sum_{j=1}^{c} \left|\frac{y_{pred,j}}{y_{exp,j}} - 1\right| * \frac{100\%}{c}, \quad \text{(Equation 12)}$$

where j is the j-th datapoint in the c characteristic cycles of a single cell, y_{exp} is the expected SOH from the discharge data and y_{pred} is the weighted average of the predicted SOH.