A Data-Driven Snapshot Method for State of Health Modelling and Diagnostics of Maritime Battery Systems

Erik Vanem[1, 2], Maximilian Bruch [3], Qin Liang [1], David Vicente Reyes Gonzalez [3], Øystein Åsheim Alnes [4]

[1] *DNV Group Research & Development, Høvik, Norway*
Erik.Vanem@dnv.com
Qin.Liang@dnv.com

[2] *Department of Mathematics, University of Oslo, Oslo, Norway*
erikvan@math.uio.no

[3] *Fraunhofer ISE, Freiburg, Germany*
maximilian.bruch@ise.fraunhofer.de
david.reyes.gonzalez@ise.fraunhofer.de

[4] *DNV Maritime, Høvik, Norway*
oystein.alnes@dnv.com

ABSTRACT

Battery systems are increasingly being used for powering ocean going ships, and the number of fully electric or hybrid ships relying on battery power for propulsion and manoeuvring is growing. In order to ensure the safety of such electric ships, it is of paramount importance to monitor the available energy that can be stored in the batteries, and classification societies typically require that the state of health of the batteries can be verified by independent tests - annual capacity tests. However, this paper discusses data-driven diagnostics for state of health modelling for maritime battery systems based on operational sensor data collected from the batteries as an alternative approach. There are different strategies for such data-driven diagnostics. Some approaches, referred to as cumulative damage models, require full operational history of the batteries in order to predict state of health, and this may be impractical due to several reason. Thus, snapshot methods that are able to give reliable estimation of state of health based on only snapshots of the data streams are attractive candidates for data-driven diagnostics of battery systems on board ships. In this paper, data-driven snapshot methods are explored and applied to a novel set of degradation data from battery cells cycled in laboratory tests. The paper presents the laboratory tests, the resulting battery data, shows how relevant features can be extracted from snapshots of the data and presents data-driven models for state of health prediction. It is discussed how such methods could be utilized in a data-driven classification regime for maritime battery systems. Results are encouraging, and yields reasonable degradation estimates for 40% of the tested cells. This is greatly improved if data from the actual cell is included in the training data, and indicates that better results can be achieved if more representative training data is available. Nevertheless, improved accuracy is required for such snapshot methods to be recommended for ships in actual operation.

1. INTRODUCTION

Electric or hybrid ships using batteries have been an increasingly attractive alternative for many shipping segments, most notably ferries and offshore supply vessels, with significant environmental benefits and large potential for fuel, cost and emission savings. Moreover, electrification of the ship fleet is completely aligned with societal and regulatory ambitions for emission reduction and a change to more environmentally friendly technologies for maritime transport.

The maritime industry has always been concerned with safety at sea, and the safety of battery-powered ships is no exception. The risk of fire and explosion are obvious for battery powered ships, and these are controlled by risk mitigation measures and safety regulations. Another central aspect of the safety of electric ships is to ensure that the available energy stored in the batteries is sufficient to cover the demand for safely operating the vessel. Loss of propulsion power in a critical situation can lead to collision or grounding accidents with potentially severe consequences. Therefore, a reliable estimation and prediction of the actual available energy of a maritime battery system is crucial.

Battery systems are ageing, meaning that the energy storage capacity degrades by calendar time and by charge/discharge cycles. This degradation affects both the amount of charge that can be stored in the battery as well as the power that can be delivered. Battery degradation also affects fire safety and thermal runaway properties (Geisbauer, Wöhrl, Mittmann, & Schweiger, 2020; Ren et al., 2019), but this paper focuses on data-driven diagnostics for state of health (SOH) estimation. Thus, the monitoring of the degradation of capacity for maritime battery systems based on sensor data will be addressed in this paper.

Classification societies typically require annual capacity testing for ships utilizing batteries for propulsion or manoeuvring in order to ensure that the estimated State of Health estimated by the battery management system (BMS) is accurate and reliable (DNV, 2021a, 2021b; DNV GL, 2016). There are some challenges with this approach, however, and data-driven methods to predict SOH are believed to be an attractive alternative if they can be demonstrated to work satisfactorily. From a practical point of view, the annual capacity test is time consuming and typically requires that the ship is taken out of normal operation for the duration of the test. Moreover, the accuracy of the test is questionable due to several factors influencing the results, such as variability in loads, temperatures and Depth of Discharge (DOD). Maritime battery systems are typically designed for a 10-year lifetime while ships are normally designed for 25-30 years. Hence, the ship will typically outlive the onboard battery system, which may need to be replaced. When battery systems are approaching their end of useful life (EOL) reliable estimation of SOH will become increasingly important, both from a safety point of view, but also from pure economical considerations.

A recent literature survey on data-driven models for SOH estimation presented an overview of various approaches and grouped them into a few generic categories (Vanem, Bertinelli Salucci, Bakdi, & Alnes, 2021; Vanem, Alnes, & Lam, 2021). One important distinction that was made is between cumulative methods and snapshot methods.

Cumulative methods refer to methods that rely on the full loading history of the batteries in order to predict current SOH. Such methods can relate information such as number of equivalent full cycles (EFC) or the total energy throughput the battery has experienced, combined with other stress factors such as temperature, C-rate and variations in state of charge (SOC) to maximum available capacity. In essence, such methods can model the accumulated degradation by establishing a relationship between the individual cycles and the *change* in SOH, i.e., ΔSOH. The actual SOH after n cycles can then be estimated as the cumulative sum of such differences, i.e., $SOH_n = SOH_0 + \sum_{i=1}^{n} \Delta SOH_i$, where SOH_0 is a known initial capacity; typically 100%. Although this is an attractive approach, with potentially accurate and reliable results, it has some challenges. For example, the full operating history of the batteries are needed, and it will be challenging to handle large data gaps. For a maritime battery system onboard ocean going ships, it may be difficult to guarantee uninterrupted data streams throughout the lifetime of the battery system. Moreover, for very large battery systems, the amount of data can easily be enormous, putting strict requirements on ship to shore connectivity, storage and computational capacity for the data-driven models. One example of a cumulative data-driven method for SOH prediction is the battery.ai tool (Xue, Zhou, Luo, & Lam, 2022); for other examples see e.g. (Nuhic, Terzmehic, Soczka-Guth, Buchholz, & Dietmayer, 2013; You, Park, & Oh, 2016; Nuhic, Bergdolt, Spier, Buchholz, & Dietmayer, 2018; Xu, Oudalov, Ulbig, Andersson, & Kirschen, 2018; S. Li, He, Su, & Zhao, 2020).

Snapshot methods, on the other hand, refer to methods that can make SOH predictions from just brief snapshots of the data without requiring the full cycling history. Such methods can use regression models where for example features extracted from partial charging or discharging curves or incremental capacity curves are used as covariates. Some examples of methods that exploit such features can be found in e.g. (Weng, Cui, Sun, & Peng, 2013; Feng et al., 2013; Zheng, Zhu, Lu, Wang, & He, 2018; Jiang, Dai, & Wei, 2020). It is noted that there are several challenges with this approach as well, and it may not be straightforward to account for the effects of varying conditions on the charge/discharge curves, e.g., varying temperatures and current rates. Notwithstanding, with such models it would be sufficient to receive batches of the data at regular intervals, which would be much more practical from a third-party verification point of view. Hence, if such models can be established that perform well enough, they may be the preferred approach for independent verification and validation of onboard SOH prediction routines. Such models typically include various regression-type models that would need to learn the relationship between the extracted features and SOH from a training dataset. The simple method proposed in (Plett, 2011), which is a linear regression model based on Coulomb counting and accounting for measurement uncertainties, do not need training data, however, but it is heavily dependent on accurate SOC estimation. In this paper, snapshot methods based on the raw data of measured currents, voltages and temperatures rather than derived quantities such as SOC will be explored and applied to a novel dataset from laboratory battery cycling tests.

Other approaches to capacity monitoring include direct measurement techniques and state-space models with observers (see e.g. (Niu, Wang, Liu, & Zhang, 2022) for a recent example).

2. LABORATORY BATTERY CYCLING TESTS

Two different types of battery cells have been subject to cycling tests at Fraunhofer's laboratory in order to generate degradation data. Two types of cylindrical 18650 cells, i.e. energy cells (henceforth denoted DDE) and power cells (henceforth denoted DDP), have been cycled according to a specified test matrix. According to the test matrix, individual cells have been cycled within specified lower and upper voltage limits, with specified charge and discharge current rates, and at specified controlled temperatures. Varying these parameters for different cells yields different degradation rates.

In the experiments, the battery cells are cycled continuously according to these specifications, interrupted at regular intervals to perform check-ups and capacity measurements. These check-ups include pulse tests and charge and discharge capacity measurements by way of Coulomb counting over a deep cycle at low current rates. Hence, there will be observations of capacities at certain points in time for all cells. This is illustrated in Figure 1, which shows the measured capacities from these test procedures as a function of the number of equivalent full cycles (EFC) for the two types of cells. As can easily be observed, the degradation of the cells varies considerably according to how they have been cycled. It is noted that the cells in this experiments have been charged and discharged according to a constant-current-constant voltage (CCCV) scheme: the cells are charged/discharged with constant current until the cut-off voltage, where the cells continue to charge/discharge at constant voltage with a current that gradually decreases towards zero.

3. DATA DESCRIPTION

Values of current, voltage and temperature are sampled continuously, resulting in high-resolution time-series of these variables throughout the experiment. From these raw measurements, different derived variables can be calculated as well, such as cumulative throughputs, cycle counts and equivalent full cycles. An example of time series of the raw measurements of current, voltage and temperature is shown in Figure 2 for an arbitrary cell. According to the test-matrix, this particular cell should be cycled between SOC = 50% to 10%, corresponding to voltage limits approximately 3.70 and 3.23 V, respectively, with discharge and charge C-rates of 0.75 and 0.2, respectively and at a temperature of 22 $^{\circ}C$. The regular cycling and the check-ups are easily discerned in the figure. It can also be observed that the cycling has been interrupted at certain times during the experiment.

Similar measurements are collected from a total of 65 individual cells; 35 DDE cells and 30 DDP cells. This constitutes the datasets used in this study for establishing data-driven models for state-of-health and capacity estimation. It is noted that these data will be made available for others for research purposes upon requests from the authors[1].

4. FEATURES EXTRACTION FROM SNAPSHOTS OF DATA

4.1. Extracting Charge and Discharge Curves

In order to establish data-driven snapshot methods for prediction of state of health, the data need to be pre-processed, so that selected features can be extracted from the raw time series. First, different filters are applied in order to extract the charge and discharge curves from the regular cycling part of the data. This is done by first identifying where in the time-series regular cycling starts and ends, and then filtering out individual charge and discharge half-cycles by examining the currents and when the current changes sign. Additional checks and filters are applied in order to remove spikes and correct for erroneously categorized data. The individual charge and discharge cycles within each regular cycling period are then numbered, starting at 1 for the first cycles following a capacity measurement or check-up. In this study, features have then been extracted from the second charge and discharge curves after such a check-up. This choice is made based on two considerations: one wants features that are as close to known capacities as possible (i.e., close in time and EFC to the capacity measurements) and features far enough away from the check-ups so that the effect of the pulse-testing on the battery cells have diminished. Considering the second set of cycles after a test is a trade-off between these two considerations. Examples of extracted charge and discharge curves for an arbitrary cell are shown in Figure 3. The different colours correspond to different regular cycling periods. The measured capacity preceding each period of regular cycling is also indicated in the figures.

Some interesting observations can be made from these plots. First, it is clearly seen that the charge and discharge curves change as the battery degrades. From the figure to the left, where only the second cycle after each test is shown, it can be seen that the curves change notably. From the rightmost figure, it is observed that also within a regular cycling interval, the charge and discharge curves change gradually. It is also observed that some of the charging curves behave slightly differently and do not start at the same voltage level. This happens to be from the first charge cycles, immediately following a check-up, and this verifies the choice of using the second cycles. The difference between the second and third sets of cycles, on the other hand, is much smaller.

When the individual charge and discharge half-cycles have been extracted from the time series, they need to be matched with results from the corresponding capacity tests. Results of this matching is illustrated in Figure 4, where the matching is illustrated in terms of both equivalent full cycles (EFC) and

[1] Contact the second author to get access to these data

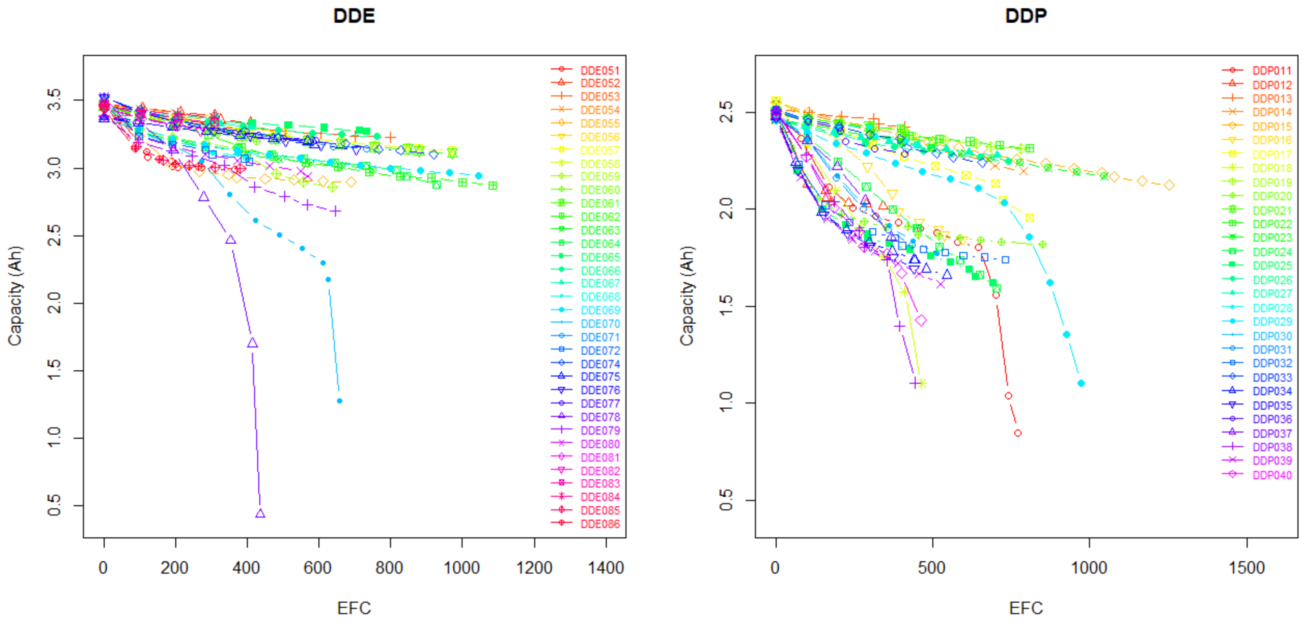

Figure 1. Measured capacity as a function of equivalent full cycles; DDE and DDP cells

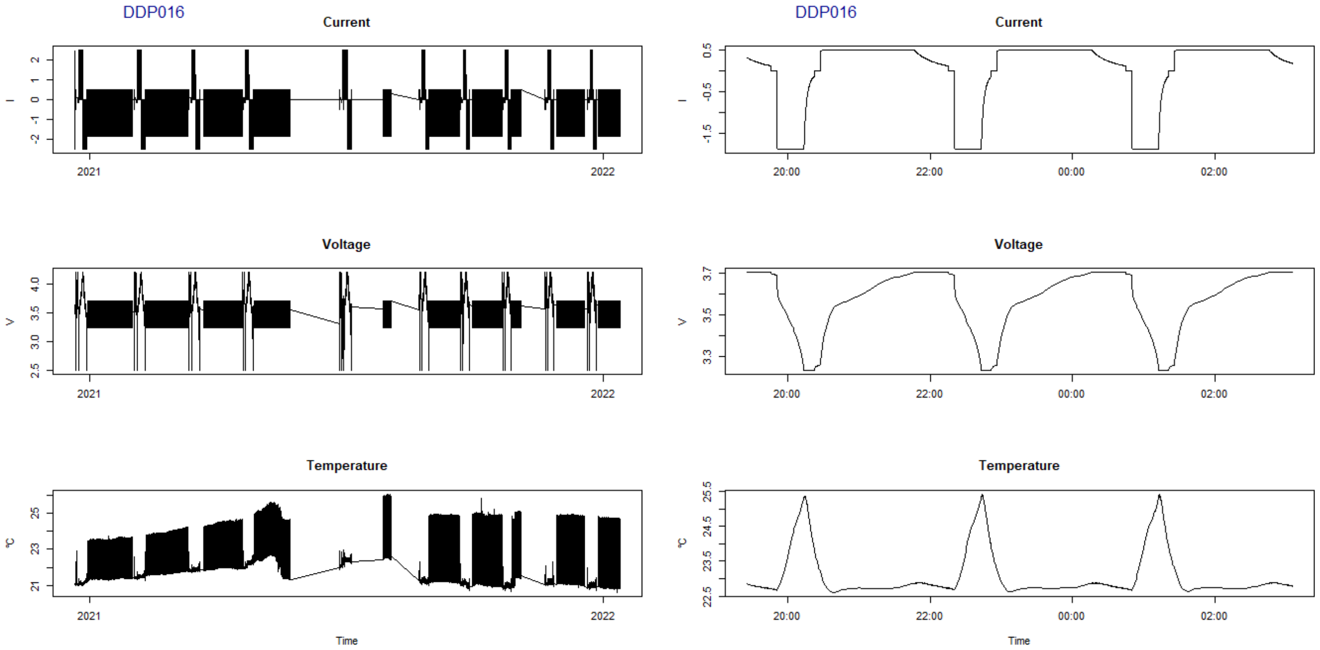

Figure 2. Example of measured time series from the cell cycling; currents (top), voltages (middle) and temperature (bottom); full time series (left) and zooming in on one of the periods with regular cycling (right)

time for two arbitrarily selected cells. Vertical red lines indicate the EFC/time of the capacity measurement and blue vertical lines represents the EFC/time of the second charge cycle. It can be observed that the cycles and the capacity measurements are close in both EFC and time in most cases. In some cases, where the cycling has been interrupted for some reason, there might be some time between the test and the cycle, but they will still be close in EFC. At any rate, for the purpose of this study, it will be assumed that the capacity from the preceding capacity test is approximately the same as the actual capacity during the second charge/discharge cycle. Note also that the same capacity will be assumed for the charge and dis-

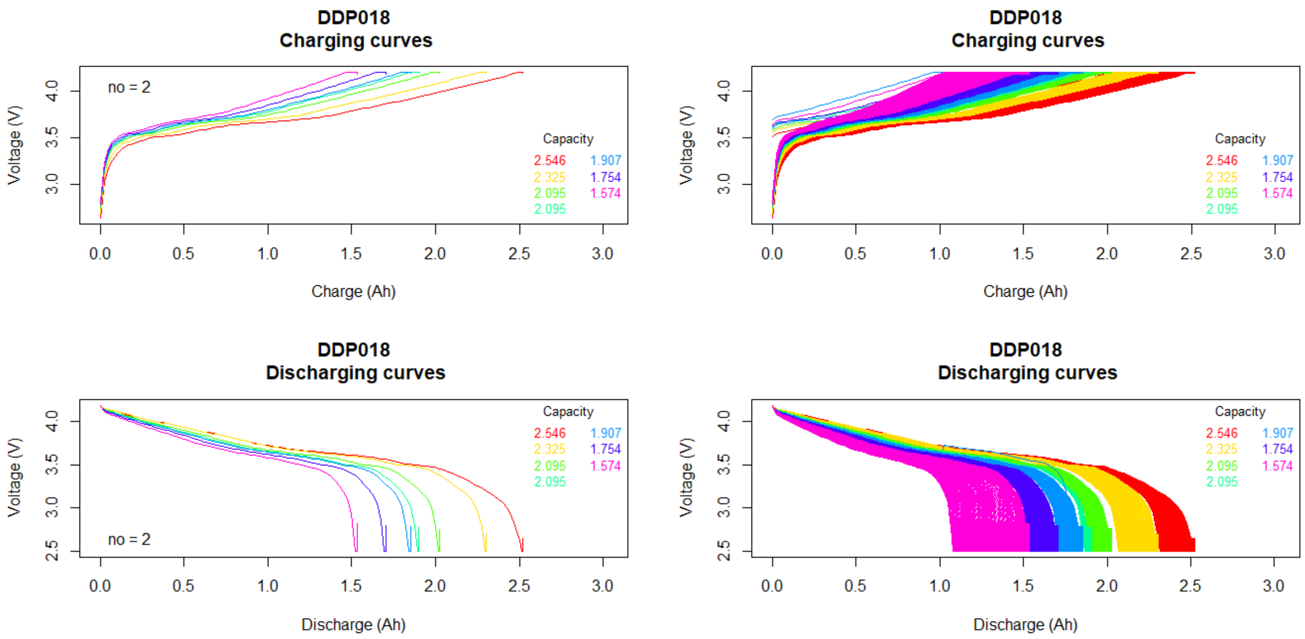

Figure 3. Extracted charge and discharge curves from the raw time series for an arbitrary cell. Only the second curves after each test (left) and all curves (right)

charge cycle, and the effect of the degradation over a single charging half-cycle is assumed negligible.

4.2. Extracting Features from the Charge and Discharge Curves

The next step is to extract particular features from these curves. Several alternative features can be used, and in this paper, features related to the current rate, temperature and energy throughput between selected voltage ranges will be utilized. That is, for each selected cycle, the mean, minimum and maximum temperature as well as the mean current are used as covariates. Moreover, the total energy throughputs between voltage ranges in steps of 0.1 V, as illustrated for an arbitrary cell in Figure 5, are used as additional explanatory variables (a similar idea was explored in (Z. Deng et al., 2022)). Simple linear interpolation have been used to estimate the cumulative throughput at the voltage limits (in terms of Ampere-hours (Ah)). It is noted that even though the constant-voltage phase of the charging/discharging cycles might contain information that can be related to the degradation state of the cells, features have only been extracted from the constant-current phase in this study. One reason for this is that these are the features deemed most likely to be found in data from battery systems in actual operation onboard ships. This is illustrated in Figure 5, where the black line represent the complete charge/discharge and the red points correspond to measurements during the constant current part, from which the voltage-based features are calculated. Other features suggested in the literature, include features from derivative curves and from probability density functions from time spent in different voltage ranges, see e.g. (Weng et al., 2013; Feng et al., 2013; Zheng et al., 2018; Jiang et al., 2020), and feature extraction is also discussed in (Y. Deng et al., 2019; Guo, Cheng, & Yang, 2019; Shu et al., 2020).

In this way, a set of snapshot-based features are extracted from the time-series data and simple prediction models will be trained on these to estimate the capacity and subsequently the State of Health of the battery cells. In total up to 44 features are collected and the overall dataset of extracted features contains 281 samples for the DDE cells and 269 samples for the DDP cells. However, it should be noted that not all cells have information for all covariates. The various cells have been cycled between different voltage limits, and therefore have different subsets of the voltage-based features. Hence, the feature matrix is sparse, and this represent an additional challenge. It means that the effective number of samples available for training is reduced, and at different degrees for the different cells. For most of the models presented in the following, only complete cases are used for training. This means that values are needed in the training data for all covariates that are relevant for the cell to be predicted. This will be further elaborated in the analysis and results section of this paper.

5. DATA DRIVEN MODELS

A number of rather simple statistical models are employed in this study to predict the capacity of the battery cells based

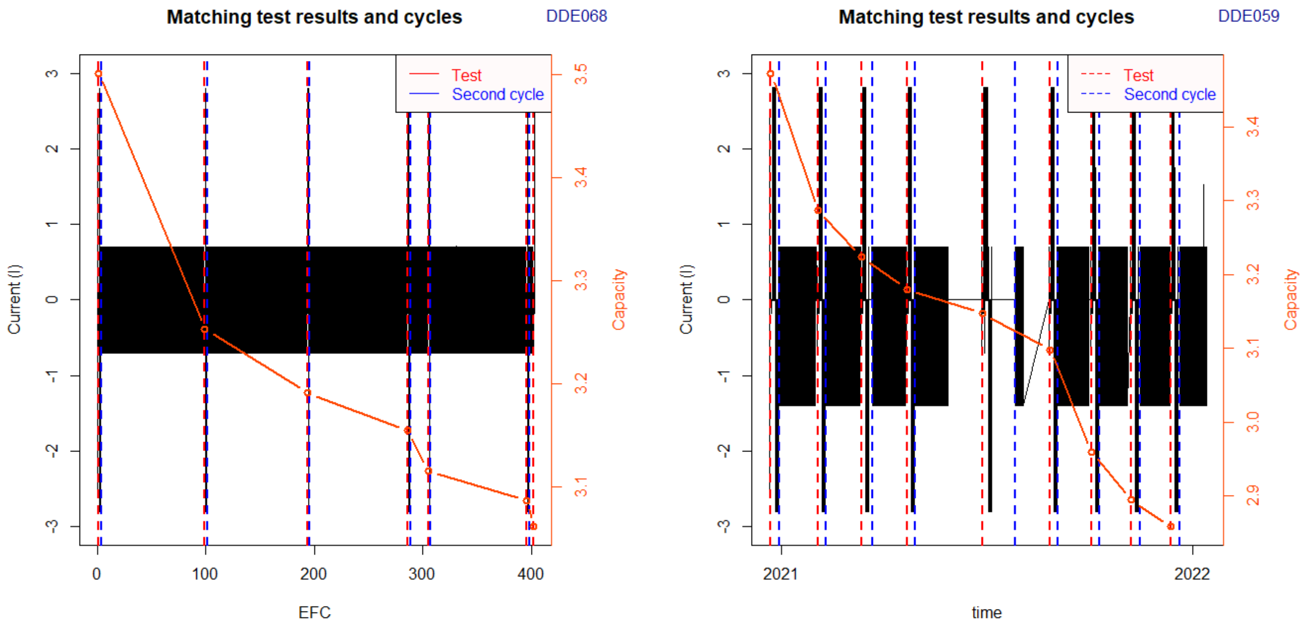

Figure 4. The extracted charge and discharge cycles in the raw time series have been matched with results from capacity measurements. In terms of EFC (left) and in terms of time (right)

on snapshot features. The amount of training data is not sufficient to train more complicated machine learning models such as neural networks and deep learning. Moreover, simple prediction models have the advantage that they are more easily interpretable, and that they are less prone to overfitting. At any rate, the following data-driven models are explored in this study:

- Linear regression (Linear)
- Linear regression with missing covariates (Miss)
- Regression tree (RT)
- Random forest (RF)
- Generalized additive models (GAM)
- Ridge regression (Ridge)
- Least absolute shrinkage and selection operator regression (Lasso)
- Multivariable fractional polynomial regression (MFP)
- Support vector regression (SVM)

Reference is made to standard textbooks for full mathematical descriptions of the various models, and in the following, a crude qualitative description will be given.

Linear regression models assume a linear relationship between the response variable y (capacity in this case) and the individual covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$, i.e.

$$y = f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon, \qquad (1)$$

where the β's are regression coefficients estimated from data and ε is an error term, typically assumed zero-mean Gaussian. Standard linear regression requires data for all relevant covariates. However, the linear regression with missing covariates model tries to account for missing covariate values by imputation: values of missing covariates can be predicted based on the available data and these predictions can be used in the regression model.

Regression trees represent a different way of obtaining a prediction rule for the response variable. The input covariate space is subdivided into several nodes by making splits for selected covariates. Simple prediction models are then applied within each terminal node, typically the nodal mean. Such models are very flexible, but may be prone to overfitting. A random forest is a model that combines several regression trees, trained on different subsets of the data, and combines them into a random forest that makes predictions based on the ensemble of trees. This is normally found to reduce the risk of overfitting from individual trees.

Generalized additive models are predictive models based on fitting nonlinear functions to each individual covariate, typically using splines. Hence, more flexible relationships between the response variable and the covariates can be found. That is, the linear terms $\beta_i x_i$ in eq. (1) is replaced by terms of the form $f_i(x_i)$, where f_i can be nonlinear functions.

Ridge regression and lasso are linear regression models, where additional regularization constrains are applied in order to shrink the regression coefficients and avoid overfitting.

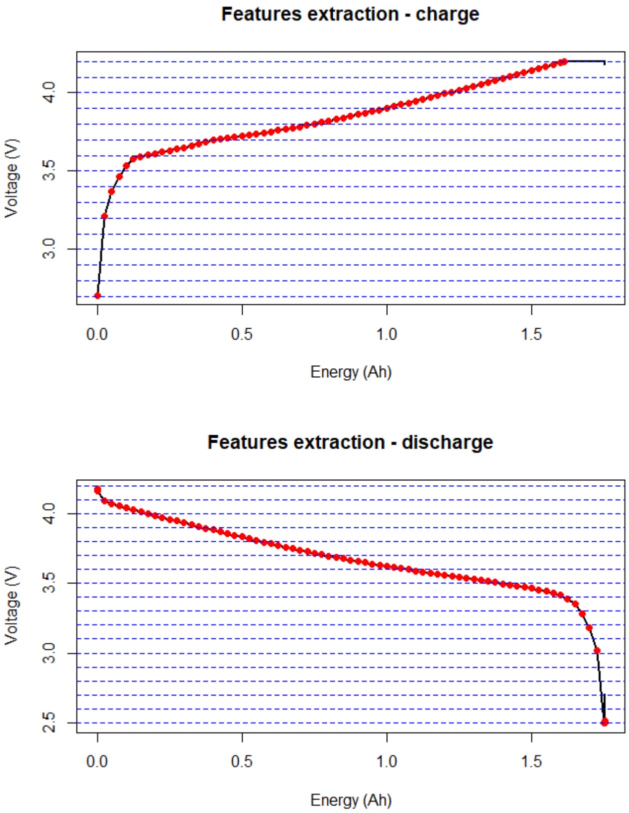

Figure 5. Energy throughput (Ah) accross selected voltage ranges are calculated and used as explanatory variables

The difference between ridge and lasso is how the penalty term is defined: in ridge regression the penalty term is proportional to the square of the magnitude of the coefficients whereas for lasso the penalty is proportional to the absolute sum of the coefficients. Multivariable fractional polynomial regression is again similar to linear regression, but searches for an optimal polynomial transformation of the individual covariates. Such methods have previously been applied to battery data in (Bertinelli Salucci, Bakdi, Glad, Vanem, & De Bin, 2022). Finally, support vector regression is based on finding a hyperplane in a higher dimensional space that can be used to make predictions. So-called support vectors are used to find this and are the data points closest to the hyperplane. For further details on these regression models, reference is made to e.g. (Hastie, Tibshirani, & Friedman, 2009).

In addition to these regression models, two types of ensemble predictions will be calculated. The first is simply the ensemble mean, i.e. the mean prediction from all the individual models, and the second is a weighted ensemble prediction, where the weighted average of the individual predictions are calculated and the weights are calculated based on the relative root mean square error (RMSE) of the individual predictors. That is, the models that obtain a lower RMSE on the test data will obtain a higher weight.

It is also noted that for some of the cells, the amount of training data was insufficient to train some of the more complicated models such as the GAM and MFP models. In these cases, the model that cannot not be estimated is replaced by the linear model (which is a special case of both the GAM and the MFP model).

6. ANALYSIS AND RESULTS

The various regression models described above, as well as the ensemble models, are applied to predict the actual capacity of all the 65 DDE and DDP cells for which cycling data are available. For each cell, two sets of predictions are made. First, all the data are used as training data to fit the models, and these are then used to predict the capacity for the cells. Note that even though data from all cells are used for training, the training data will not be identical for each cell, since they have different sets of explanatory variables. Only the voltage ranges that are relevant for the cell in question are included in the training data. In the second set of predictions, data from the cell that is to be predicted is removed from the training data. This gives out-of-sample predictions that are more comparable to predictions on real operational data.

When applying the various regression models on data from all the cells it turns out that results are quite good for most of the cells. However, when data from the cell itself is removed from the training data, results are more varying. Predictions might be reasonably good for some cells but not so for many others. For some cells, some of the models perform well, whereas others might give poor predictions for the same cells. This is the case for both DDE and DDP cells, and there are no clear trend as to what models perform best (this will be further elaborated below). Some examples of predictions that were good for most of the models are shown in Figure 6. For each of the cells, predictions are compared with observed capacities for both the case where data from the cell itself were part of the training data and the case where the models were trained on data from other cells only. In both cases, error metrics in terms of root mean square error (RMSE) and mean error (ME) are included in the plots. As can be observed, whether data from the actual cell are part of the training data or not has a significant impact on the prediction accuracy. Figure 7 shows some examples where some models yield reasonable results but where other models give poor predictions, and Figure 8 shows some examples where predictions were generally poor across models.

Similar results are obtained using the same predictive models but with only a subset of all the features described above. For example, similar analyses have been carried out on only the charging features, disregarding the features extracted from the discharge curves. Moreover, analyses have been made

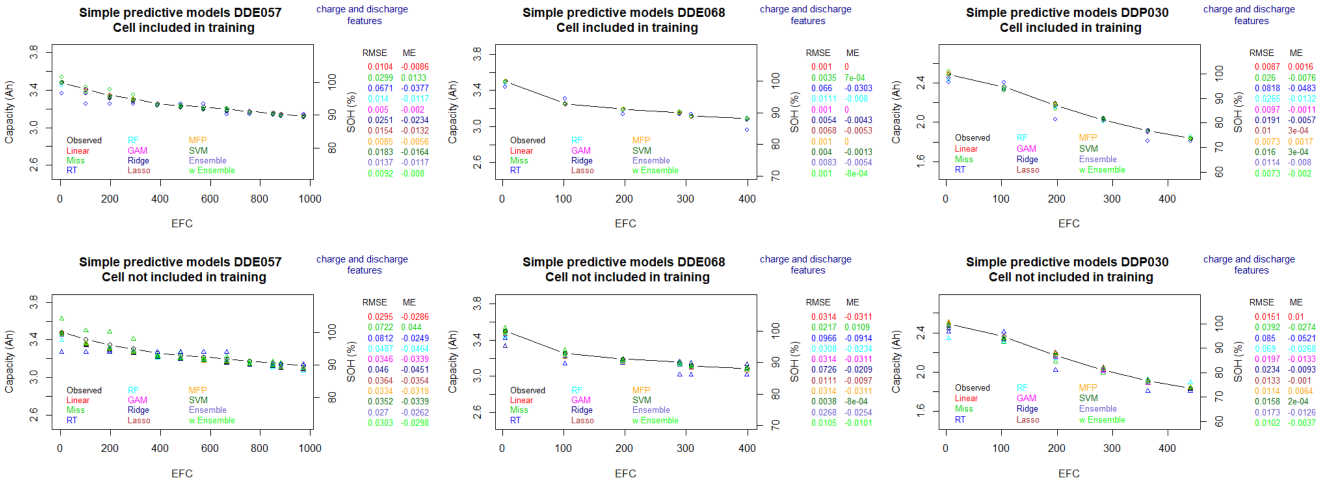

Figure 6. Data-driven predictions based on snapshot features with reasonable accuracy for most models

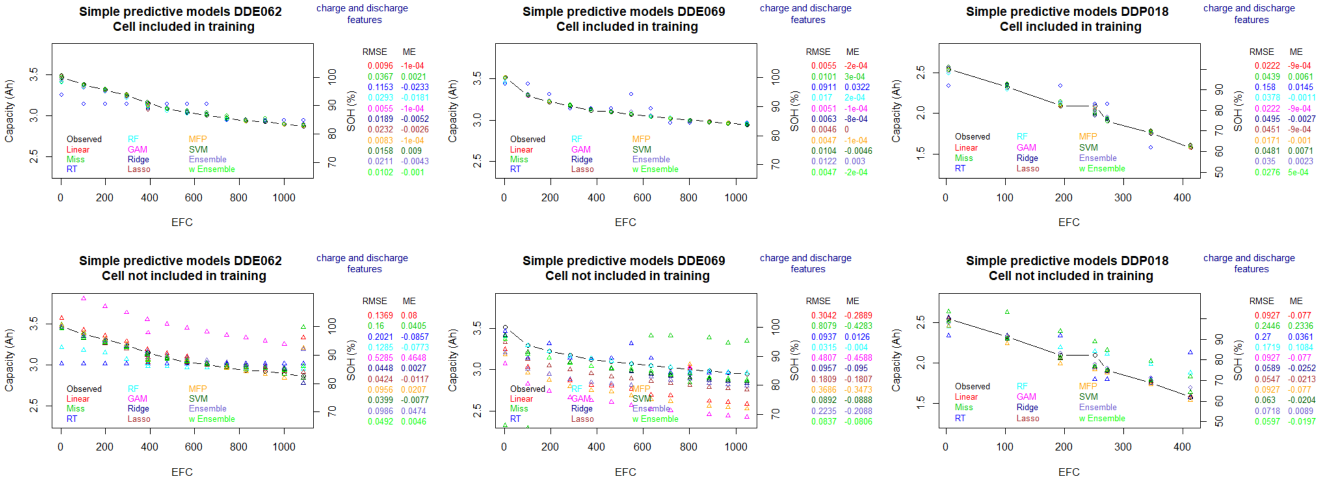

Figure 7. Data-driven predictions based on snapshot features with reasonable accuracy for some models and poor for others

where only the mean temperature is included, disregarding the temperature variation described by the minimum and maximum temperatures experienced during the charging and discharging. Another reduction in features that has been explored is to remove some voltage ranges, e.g. the first and last voltage ranges in the charge and discharge curves. Although the numerical predictions vary from case to case, the overall observations remain the same, i.e., that the simple snapshot models are able to predict well the degradation for some cells, but not for all. For example, including discharge features in addition to the charging features seems to slightly improve predictions for the DDP cell, but to slightly impair predictions for the DDE cells.

6.1. Performance Evaluation

The results above illustrate that the simple predictive models based on snapshot features may and may not perform satisfactorily for an arbitrary cell. In order to evaluate the model performance, there is a need to decide what level of accuracy is acceptable from a practical point of view. In current classification rules for electric ships (DNV GL, 2020), it is stated that the annual tests should be within ± 5% of the value presented by the battery management system (BMS). Hence, it may be assumed that an error around 5% from data-driven models would be acceptable. Thus, for an energy cell (DDE) with nominal capacity around 3.5 Ah, an RMSE below 0.175 Ah could be regarded as acceptable. Similarly, for a power cell (DDP) with a nominal capacity around 2.5 Ah, an RMSE value below 0.125 Ah could be deemed acceptable.

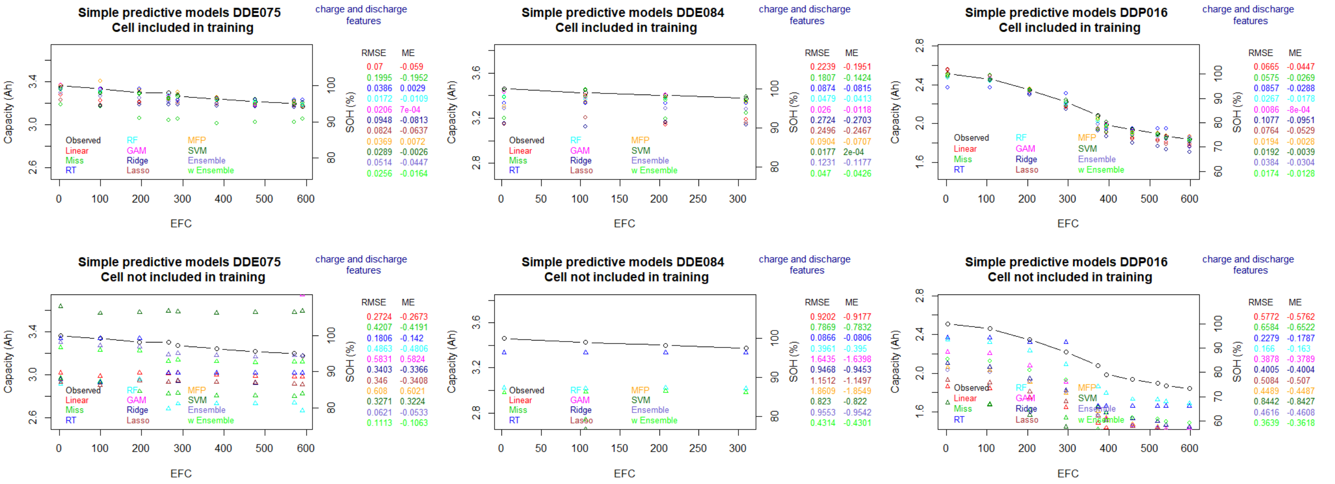

Figure 8. Data-driven predictions based on snapshot features with poor accuracy for most models

6.1.1. Performance Across Cells

One may want to evaluate the performance of the data-driven snapshot methods across the different cells, for example by calculating the average RMSE from the different predictive models. Then one could compare the average performance for each cell with the test parameters – voltage ranges, C-rates and temperatures – of the cells to see which influence the results most. This is illustrated in Figure 9, for both the DDE and the DDP cells. The orange markers present the average RMSE for each cell when the data from that cell is included in the training data and the red markers present the results when data from that cell have been excluded from the training. The vertical bars correspond to the voltage range the cells have been cycled between (right axis), and the two colours of each vertical bar correspond to the C-rate (leftmost colour) and temperature (rightmost color) of the different cells, respectively, as indicated by the colour legends in the plots. These results pertain to analyses when only the charging features are included, but the outcome is very similar for the other analyses that have been carried out.

Results from the models presented in this study, when only the features from the voltage curves are included, indicate that 16 of the 35 DDE cells (46%) have RMSE < 0.175 and that 10 of the 30 DDP cells (33%) have RMSE < 0.125. However, if the cell to be predicted has been included in the training data, then these ratios increase to 34/35 (97%) for the DDE cells and 28/30 (93%) for the DDP cells. Although it is obviously not good enough that predictions are reasonably accurate only for some of the cells, it is encouraging to observe that the rather simple predictive models based on a few snapshot features perform acceptable on nearly half of the cells. Moreover, if the training data include data from the actual cells to be tested, the simple snapshot methods perform acceptable for more than 95% of the cells. This indicates that the simple models are indeed able to model the dependencies between these covariates and the capacity of the cells, provided sufficient representative training data are available.

6.1.2. Performance Across Models

One may also try to evaluate the performance across predictive models by comparing the average RMSE for all cells, as well as counting the number of times a particular model performs best and when it performs worst. This is summarized in Table 1, for a set of analyses with only the charge-based features. It is observed that there is no clear winner. Looking at the mean RMSE for the results obtained when the predicted cell is included in the training data, it is observed that most models perform quite well. However, when these data are excluded from the training, all models do considerably worse, and the best candidate models (ignoring the ensemble predictions) would be regression tree, random forest, ridge regression and lasso. Considering results not trained on the predicted cell only, it appears that the regression tree is the model which most often performs best. However, this also turns out to be one of the models that most often performs worst. On the other hand, ridge regression and lasso never perform worst, but also rather seldom perform best. Overall, it is difficult to select one model that performs best overall, and results vary considerably between cells. Possibly, an ensemble method could be regarded as the best approach, but this is also very sensitive to very wrong predictions from individual models. Indeed, the very high average RMSE for the MFP model is from a very few predictions that are completely wrong, probably as a result of extrapolating a polynomial fit.

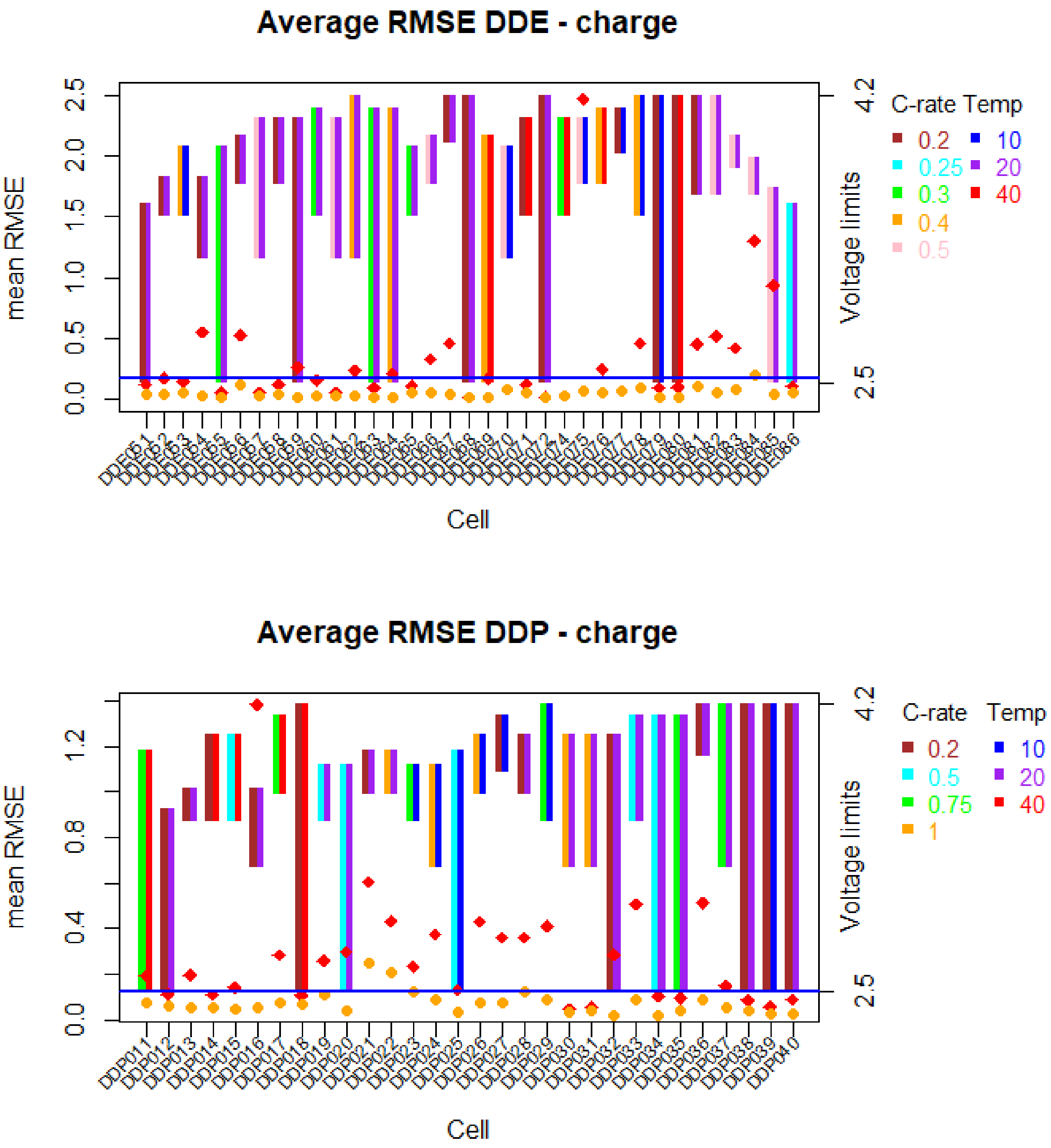

Figure 9. Average RMSE and test parameters for each cell

7. DISCUSSION

7.1. Importance of Representative Training Data

One explanation for the varying results from the snapshot methods applied in this study could be the limited training data that have been available. Since the various cells have been cycled differently, the data for the different cells do not contain the same covariates, and only the ones relevant for the cell to be predicted could be included in the models. This leads to varying amount of training data for the different cells,

Table 1. Performance across models; average RMSE for all cells; number of times a model performs best and worst

	Linear	Missing X	RT	RF	GAM	Ridge	Lasso	MFP	SVM	Ensemble	w Ens
average RMSE for all cells											
With cell	0.0699	0.120	0.126	0.0409	0.0188	0.0933	0.0747	0.0344	0.0233	0.0515	0.0236
Without cell	0.308	0.373	0.240	0.243	0.414	0.245	0.246	152	0.354	17.0	0.138
Number of times individual models perform best											
With cell	2	1	0	3	41	0	1	2	11	0	4
Without cell	6	5	10	5	7	4	7	3	4	2	12
Number of times individual models perform worst											
With cell	0	22	34	0	0	7	2	0	0	0	0
Without cell	4	17	15	1	10	0	0	10	8	0	0

and may in particular lead to lack of representative training data for some cells. For example, some cells have been cycled between 2.5 and 4.2 V and contain all voltage ranges, corresponding to 17 charge-based features. For these cells, only data from other cells that include all voltage ranges can be included in the training set, and this significantly reduces the amount of training data. In one example, this means a reduction in training samples from 261 to 44 complete cases. For cells cycled over a narrower voltage range, the number of features is reduced, and the amount of training data is effectively higher. For example, one cell cycled between 3.85 and 4.12 V contains 6 covariates with an effective training set of 96 complete cases; another cell cycled between 3.22 - 3.72 V contains 8 covariates with a training dataset of 86 samples. In all cases, the limited amount of representative training data is a possible explanation for varying results, and one possible remedy could be to enlarge the dataset by doing more laboratory tests.

This sparsity of covariates in the training data is the main motivation to try out the missing covariate model. With such a model, rather than disregarding all samples without the full set of covariates the idea is to also use this additional information by imputing values for the missing covariates based on the existing ones. However, as it turns out, this is not successful for all cells, and this model often turns out to perform worst. This is presumably due to the number of missing covariates in many cases. If only one or a few covariates are missing in a sample, it may be possible to accurately impute the missing value. However, in many cases in this dataset, a large portion of the covariates are missing in many samples, and it is unrealistic to believe that one should be able to impute them all accurately. In some examples, the number of missing covariates is similar to the number of existing ones.

The importance of representative training data is also clearly seen by looking at the few duplicate cells. In the experiment, a few pairs of cells have been cycled according to the same test parameters. Hence, even though the data from the cell that are to be predicted are removed from the training set, data from another cell that have experienced a similar load pattern are available. In this study, duplicate cells are DDE057/DDE061, DDE068/DDE072, DDP030/DDP031 and DDP038/040, and for all of these cells all the individual models perform very well even when the cell itself is not included in the training. This can also be seen by studying Figure 9. Again, this highlights the importance of representative training data to obtain good predictions.

7.2. Possibilities for Improvements

Although it is overall encouraging that the data-driven snapshot methods perform well for many of the battery cells, it is obviously not satisfactory that they fail to predict accurately for all cells. In the following, some ideas of how to improve the results are discussed.

7.2.1. Extending the Training Data

The obvious solution for improving the data-driven methods is to extend the training data. As presented above, if a cell that has experienced similar loading history is included in the training data, as is the case with the duplicate tests, the models perform very well. However, extending the training data means increasing the number of expensive and time-consuming laboratory test, which is not always an option. One alternative could be to supplement the training data with data from other experiments of similar cells, if available. However, care should be taken to ensure that the cells are sufficiently similar, as differences in chemistry and form factors may influence the degradation patterns.

Another approach could be to, rather than expanding the cycling tests, to focus testing more on the data needed for training snapshot methods and to design the experiments somewhat differently. Typically, tests are performed according to a static test matrix, where one wants to investigate how different stress factors such as varying temperature, C-rate and Depth of Discharge influence the degradation. This will typically be features in cumulative damage models and such tests are useful for training such models. However, for snapshot methods, the features will typically not be these stress factors but will be extracted from the charge and discharge curves. In this case it may not be necessary that the cells are cycled according to static test parameters, and it will be more important to ensure that all cells have a full set of covariates, for example that they are experiencing the same voltage ranges. One could assess which voltage ranges the battery system is expected to experience most frequently during normal operation, and focus on these voltage range, for varying temperatures and current rates, in the experiments. If all charge and discharge cycles contain these selected voltage intervals, the training data will be richer and could possibly lead to better predictions from snapshot methods even without extending the number of tests. One could also apply more dynamical loading in order to better explore the variation in the other covariates with a limited number of tests. Looking at the voltage ranges for the various cells indicated in Figure 9, it is obvious that it is not possible to find a subset of the voltage ranges that is included in all the cells for the available dataset. Thus, a lot of data is disregarded for each cell, and this way of testing is wasteful if the data are to be used for snapshot methods.

7.2.2. Additional features

The models established in this study are based on a small set of covariates extracted from the measured time series of current, temperature and voltage. In addition to mean current and mean, minimum and maximum temperatures, covariates related to the energy throughput at selected voltage intervals are the main explanatory variables. These have been extracted directly from the charge and discharge curves. It is possible that improved predictions can be achieved by using different or additional features. One example could be to try to extract features from derivative curves, such as dQdV or dVdQ curves. Such derivative curves are known to exhibit peaks at certain voltage levels, that may change in both location and amplitude as the battery cell degrades, and could be possible features in data-driven models. An example of such derivative curves is shown in Figure 10, and it is clearly observed that the curves change character as the cell degrades. However, there are a lot of uncertainty associated with obtaining such curves, e.g. related to the differentiation of noisy, discrete measurements as discussed in (Feng et al., 2020), and even though there are several approaches to estimate such curves (Han et al., 2019; Lin, Cabrera, Yu, Yang, & Tsui, 2020; X. Li, Wang, Zhang, Zou, & Dorell, 2019; He, Wei, Bian, & Yan, 2020), it is far from straightforward to obtain smooth curves with well-defined properties. Moreover, the information extracted from the charge and discharge curves would undeniably contain the same information as the derivative cures, and it is not obvious that predictions would be improved by including additional features from the derivative curves.

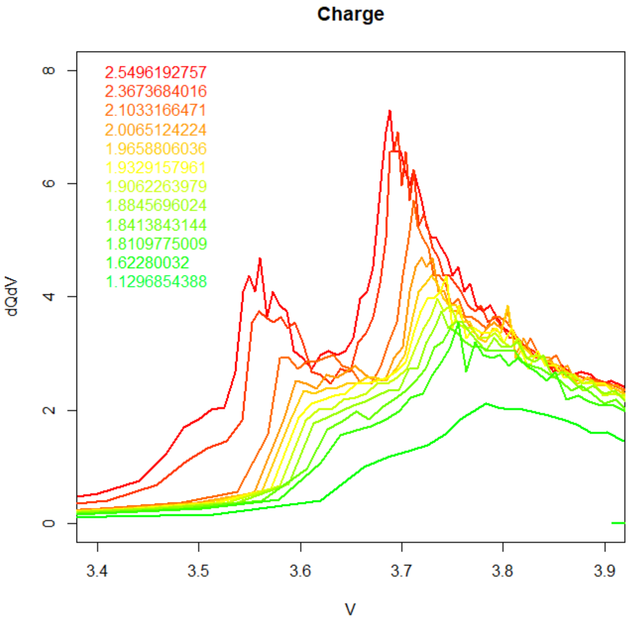

Figure 10. Example of derivative curves, dQdV

Another way of extracting features from the charge and discharge curves is to fit parametric functions to the curves, and use the parameters as features. Two examples are polynomial or piecewise linear curves. However, it is not straightforward to obtain good fit, and for example with polynomial functions, one would typically require rather high order polynomials to fit the curves well.

7.2.3. Additional Data Pre-processing, Fine-Tuning and Different Predictive Models

Careful pre-processing of the data has been performed in order to extract the charge and discharge curves from the raw time-series data and subsequently to extract the features from these curves. A number of filters have been applied to remove spikes in the data and to obtain error-free training data. However, more careful pre-processing could possibly improve the data quality and remove residual errors in the data. Examples of such errors are that only parts of the charge/discharge cycles are included in the data or that charge/discharge curves could be associated with wrong capacity measurement. It is always possible that such noise can be removed by additional pre-processing and yield improved results.

It is also possible to more carefully fine-tune the predictive models. Some of the models include several hyperparameters that needs to be specified. For many of the methods applied in this paper, hyperparameters are selected by cross-validation and grid-searches, but even more optimal values can possibly be found. Yet another option to explore is to use different predictive models. In this study, several simple data-driven models are employed, and more complicated machine learning models such as neural networks and deep learning models have not been tried out. However, it is believed that with the limited amount of training data available, it will be difficult to fit more complicated models, and even the more complicated models used in this study, most notably the GAM and the MFP models, are found to sometimes have difficulties.

Hence, notwithstanding several possibilities for improving the predictions by adding more features, more preprocessing, fine-tuning and exploring different models, it is believed that the most important potential for improvement in this study is related to the available training data. Hence, further efforts will be directed towards integrating results from other tests in order to extend the training data.

7.3. Snapshot Methods for Data-Driven Diagnostics of Ships in Operation

The objective of this study is to explore and establish data-driven models for diagnostics and state of health prediction of maritime battery systems. If such methods can be demonstrated to yield reliable results, they may be used to replace annual capacity tests as a class requirement for electric and hybrid ships. In this context, snapshot methods are believed to be a much more attractive solution compared to cumulative damage models, which would require data from the whole operational history of the ship. This is difficult to guarantee for ships at sea, with possible intermittent and limited ship-to-shore connectivity and data storage capabilities. Snapshot methods, on the other hand, would only require mere snapshots of the data streams at regular intervals.

In addition to requirements on prediction accuracy and reliability, for snapshot methods to be accepted by class as an alternative to annual capacity testing, a natural class requirement would be that sufficient and relevant training data are available. Hence, laboratory testing might be required prior to installation for the particular cell type. This is associated with a notable cost, but it is believed that test results could be reused for different installations of the same or similar battery systems. Moreover, as elaborated above, it is believed that if experiments are carefully designed with snapshot methods in mind, it should be possible to obtain richer training data without prohibitively extensive and expensive laboratory testing. As has been demonstrated by this study, the quality and representativeness of the training data is crucial, and further research is needed in order to specify such data requirements within a data-driven classification regime.

8. Summary and Conclusion

This paper has presented results from a study on snapshot methods for data-driven state of health modelling and monitoring of battery systems onboard ships in operation. If such models can be demonstrated to work well, they will represent a huge benefit to the maritime industry compared to annual capacity testing or data-driven modelling using cumulative damage methods. The paper has illustrated how charge and discharge curves extracted from raw measurements of currents and voltages are influenced by battery degradation and how relevant features can be identified from these. Moreover, a set of simple statistical models are explored for predicting the actual capacity of the batteries.

The overall results are encouraging, and demonstrates that simple statistical models based on a limited set of features easily obtained from sensor data are able to predict the degradation in battery capacity. With representative training data, illustrated in this study by including data from the cell to be predicted in the training data, the models yield reasonable results in more than 95% of the cases. If data from the cell are excluded from the training set, results are still reasonable for 40% of the cells. This is encouraging, but needs to be improved before such methods can be recommended as an alternative to current class requirements of annual capacity tests.

The results clearly illustrate the sensitivity of data-driven models in general, and snapshot methods in particular, to available training data. It is believed that lack of sufficient representative training data is the main explanation for the varying results, and further efforts will be directed towards extending the data set and more carefully designing experiments with snapshot methods in mind.

Acknowledgment

This work has partly been carried out within the DDD BATMAN project, supported by MarTERA and the Research Council of Norway.

References

Bertinelli Salucci, C., Bakdi, A., Glad, I. K., Vanem, E., & De Bin, R. (2022). Multivariable fractional polynomials for lithium-ion batteries degradation models under dynamic conditions. *Journal of Energy Storage*, *52*, 104903.

Deng, Y., Ying, H., E, J., Zhu, H., Wei, K., Chen, J., … Liao, G. (2019). Feature parameter extraction and intelligent estimation of the state-of-health of lithium-ion batteries. *Energy*, *176*, 91-102.

Deng, Z., Hu, X., Xie, Y., Xu, L., Li, P., Lin, X., & Bian, X. (2022). Battery health evaluation using a short random segment of constant current charging. *iScience*, *25*(5), 104260.

DNV. (2021a). *Rules for classification: Ships.* (DNVGL-RU-SHIP)

DNV. (2021b). *Rules for classification: Ships. part 6 additional class notations. chapter 2 propulsion, power generation and auxiliary systems.* (DNV-RU-SHIP Pt.6

Ch.2)

DNV GL. (2016). *DNV GL handbook for maritime and offshore battery systems* (Tech. Rep. No. 2016-1056). Author.

DNV GL. (2020). *Rules for classification: Ships. part 6 additional class notations. chapter 2 propulsion, power generation and auxiliary systems.* (DNVGL-RU-SHIP Pt.6 Ch.2)

Feng, X., Li, J., Ouyang, M., Lu, L., Li, J., & He, X. (2013). Using probability density function to evaluate the state of health of lithium-ion batteries. *Journal of Power Sources*, *232*, 209-218.

Feng, X., Merla, Y., Weng, C., Ouyang, M., He, X., Liaw, B. Y., … Offer, G. (2020). A reliable approach of differentiating discrete sampled-data for battery diagnosis. *eTransportation*, *3*, 100051.

Geisbauer, C., Wöhrl, K., Mittmann, C., & Schweiger, H.-G. (2020). Review of safety aspects of calendar aged lithium ion batteries. *Journal of the Electrochemical Society*, *167*, 090523.

Guo, P., Cheng, Z., & Yang, L. (2019). A data-driven remaining capacity estimation approach for lithium-ion batteries based on charging health feature extraction. *Journal of Power Sources*, *412*, 442-452.

Han, X., Feng, X., Ouyang, M., Lu, L., Li, J., Zheng, Y., & Li, Z. (2019). A comparative study of charging voltage curve analysis and state of health estimation of lithium-ion batteries in electric vehicle. *Automotive Innovation*, *2*, 263-275.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

He, J., Wei, Z., Bian, X., & Yan, F. (2020). State-of-health estimation of lithium-ion batteries using incremental capacity analysis based on voltage-capacity model. *IEEE Transactions on Transportation Electrification*, *6*(2), 417-426.

Jiang, B., Dai, H., & Wei, X. (2020). Incremental capacity analysis based adaptive capacity estimation for lithium-ion battery considering charging condition. *Applied Energy*, *269*, 115074:1-12.

Li, S., He, H., Su, C., & Zhao, P. (2020). Data driven battery modeling and management method with aging phenomenon considered. *Applied Energy*, *275*, 115340.

Li, X., Wang, Z., Zhang, L., Zou, C., & Dorell, D. D. (2019). State-of-health estimation of li-ion batteries by combining the incremental capacity analysis method with grey relational analysis. *Journal of Power Sources*, *410-411*, 106-114.

Lin, C., Cabrera, J., Yu, Y., Denis, Yang, F., & Tsui, K. (2020). SOH estimation and SOC recalibration of lithium-ion battery with incremental capacity analysis & cubic smoothing spline. *Journal of The Electrochemical Society*, *167*, 090537.

Niu, G., Wang, X., Liu, E., & Zhang, B. (2022). Lebesgue sampling based deep belief network for lithium-ion battery diagnosis and prognosis. *IEEE Transactions on Industrial Electronics*, *69*(8), 8481-8490.

Nuhic, A., Bergdolt, J., Spier, B., Buchholz, M., & Dietmayer, K. (2018). Battery heath monitoring and degradation prognosis in fleet management systems. *World Electric Vehicle Journal*, *9*(3), 39:1-20.

Nuhic, A., Terzmehic, T., Soczka-Guth, T., Buchholz, M., & Dietmayer, K. (2013). Health diagnosis and remaining useful life prognostics of lithium-ion batteries using data-driving methods. *Journal of Power Sources*, *239*, 680-688.

Plett, G. L. (2011). Recursive approximate weighted total least squares estimation of battery cell total capacity. *Journal of Power Sources*, *196*, 2319-2331.

Ren, D., Hsu, H., Li, R., Feng, X., Guo, D., Han, X., … Ouyang, M. (2019). A comparative investigation of aging effects on thermal runaway behavior of lithium-ion batteries. *eTransportation*, *2*, 100034.

Shu, X., Li, G., Shen, J., Lei, Z., Chen, Z., & Liu, Y. (2020). A uniform estimation framework for the state of health of lithium-ion batteries considering feature extraction and parameters optimization. *Energy*, *204*, 117957.

Vanem, E., Alnes, Ø. Å., & Lam, J. (2021, November-December). Data-driven diagnostics and prognostics for modelling the state of health of maritime battery systems – a review. In *Proc. annual conference of the prognostics and health management society 2021 (phm 2021).*

Vanem, E., Bertinelli Salucci, C., Bakdi, A., & Alnes, Ø. Å. (2021). Data-driven state of health modelling – a review of state of the art and reflections on applications for maritime battery systems. *Journal of Energy Storage*, *43*, 103158.

Weng, C., Cui, Y., Sun, J., & Peng, H. (2013). On-board state of health monitoring of lithium-ion batteries using incremental capacity analysis with support vector regression. *Journal of Power Sources*, 36-44.

Xu, B., Oudalov, A., Ulbig, A., Andersson, G., & Kirschen, D. S. (2018). Modeling of lithium-ion battery degradation for cell life assessment. *IEEE Transactions on Smart Grid*, *9*(2), 1131-1140.

Xue, Y., Zhou, H., Luo, Y., & Lam, J. (2022, March). Battery degradation modelling and prediction with combination of machine learning and semi-empirical methods. In *Proc. 12th International conference on Power, Energy and Electrical Engineering (CPEEE 2022).*

You, G.-w., Park, S., & Oh, D. (2016). Real-time state-of-health estimation for electric vehicle batteries: A data-driven approach. *Applied Energy*, *176*, 92-103.

Zheng, L., Zhu, J., Lu, D. D.-C., Wang, G., & He, T. (2018). Incremental capacity analysis and differential voltage analysis based state of charge and capacity estimation for lithium-ion batteries. *Energy*, *150*, 759-769.