# Contents

# 1 ABSTRACT

Lithium-ion battery safety is the single most crucial issue in the reliability of electric vehicles. The early state of health estimation and life prediction ensure the safe operation and timely detection of battery faults. This study aims to use voltage, capacity degradation, and EIS technique for predicting the cell's End-of-Life. Firstly, our method can accurately predict a cell's remaining useful life at any point in its lifetime, starting from the 10th cycle for short life cells and the 40th cycle for long life cells. Secondly, we have made significant advancements in predicting remaining useful life through EIS data by generating new features and studying the degradation pattern with aging. Our results improve upon the existing literature in both the techniques


# 2 Introduction

With increasing innovation and environmental awareness, there has been a revolution from traditional gasoline-powered vehicles to electric vehicles (EVs). This revolution is also leading to innovation in the field of Li-ion batteries. Li-ion batteries with their higher energy and power density are the first choice for the EVs. The battery on an electric car is a proven technology that will last for many years. The question is, how many years? To answer this question and to inform a customer in well advance about its battery end of life (EOL), the battery remaining useful life (RUL) and state of health (SOH)is estimated. When the battery capacity reaches 80% of its initial capacity, battery reaches its EOL. Once an EV battery loses its capacity to power a vehicle, it can power a home or building by contributing to a battery storage system. The performance of Li-ion batteries deteriorates with time and use due to the degradation of their electrochemical constituents, resulting in capacity and power fade [1]. The Li-ion battery ages mainly due to three degradation mechanism: 1) Loss of Lithium Inventory (LLI), 2) Loss of Active Material (LAM), 3) Increase of cell impedance. LLI is the consumption of lithium-ion in the parasitic reaction, such as the formation of solid electrolyte inter-phase and lithium plating. LAM is the loss of electrode due to mechanical stress leads to the reduction of lithium storage sites. Cell impedance increases due to the growth SEI layer [2]

Nowadays, there is a boom in the use of Machine Learning (ML) in every field. ML has also gained significant importance in the field of Li-ion batteries with its straightforward approach as compared to physics, electrical- thermal models. There is no need to solve any tedious mathematical differential equations. With the ML, one doesn't need to see inside the battery for degradation, which is also dangerous. ML method requires data collected at some defined experimental conditions, which is used to predict life.

Cell aging is correlated with the capacity fade and power fade. Capacity fade is the deterioration of cell capacity with the cycles. Power fade is due to an increase in the impedance of the cell. To check the correctness of the above degradation method with the cell aging, two different datasets are selected 1) Severson dataset [3] 2) Cambridge dataset [4], one for each degradation approach. The capacity fade approach is performed on the Severson datasets while the power fade approach on the Cambridge datasets.

Despite the opportunities seen in the ML technique and the advances made in RUL prediction research, there have been only a few applications that have done well in the market. The technology has not matured enough for broad usage by the public in delicate applications. Some limitations should be overcome before its applications can be brought outside laboratories. In the current work, some of these problems have been resolved in both the degradation approach.

# 3 Severson Dataset

## 3.1 Discussion

The current method is built to generate the result after using the first 100 cycles with an error of around 9.5% [3]. The results are improved by halving the use of initial cycles to 50 to predict EOL by using knee-point and knee-onset in capacity degradation curves of lithium-ion cells [4]. There are two drawbacks in the models mentioned above: 1) These are using a large number of initial cycles for their predictions. 2)These can make predictions only once, i.e., after 100 cycles in the case of [3] and after 50 for [4]. We need something more reliable to make predictions with a very less number of initial cycles and any number of times. The problem in one-time prediction is that if the cell is used at different discharge and charge rates, its EOL will differ every time.

To build a reliable model, we first need to analyze the Severson dataset. After looking at the Severson dataset, we would notice that most of the cells have their EOL between 500-1000 in the training data. Now looking at the figure of the "full model of Observed and predicted cycle lives of Severson paper," we can notice that the Severson model is predicting EOL accurately for cells with their EOL between 500-1000. For cells with their EOL above 1500, it's predicting its EOL around 1200. The Severson model performs well for the only cells that lie in EOL of 500-900 because most of the training cells lie in this range. If we choose our training data correctly, we can even get good results with just an initial 10 cycles for cells with EOL less than 1000. ML model result mainly depends on how we choose our training and testing dataset. We spend most of our time finding new features, methods, and algorithms to improve outcomes, all of which go in vain without choosing a correct dataset. To verify this, we use training and testing sets of the Severson dataset in distinct ways.

We classify training set (41 cells) of Severson dataset into 2 category 1) Short life cell (EOL $<$ 1000) 2) Long life cell (EOL $>$ 1000). Similarly, we classify the secondary set (40 cells) into the 2 categories. We used 31 short life cells from the training set and 22 short life cells from the testing set for training of short life cells. Similarly, for long life cell, 10 cells from the training set and 18 cells from the secondary set. Now we will show the importance of dividing the dataset into these two categories. We have opted for two different approaches; each one has its own advantage.

**Approach 1**

In this approach we used Random Forest Regression with mean, variance, minimum of change in discharge voltage curves between the $100^{th}$ and $1^{st}$ cycle: $\Delta\mathrm{Q}_{100-1}(V)$ = $\mathrm{Q}_{100}(V)$ - $\mathrm{Q}_1(V)$ as features. On short life cells, we get an MAE of 6.5% and 8.5% on long-life cells, making an aggregate of (22*6.5%+18*8.5%)/40=7.4%. When we use all training set without categorizing into short and long life cells, we get an error of 13.5% on secondary set. This shows that choosing a correct dataset for a model is an essential part. Comparing our results with Severson datasets, they have got an MAE of 9.5%, so there is a decrease of 2.1% error. While looking at individual categories, we are getting an MAE of just 6.5% for short life cells. Moreover our model can make continuous prediction till its EOL, when we use mean, variance, minimum of change in discharge voltage curves between the $\mathrm{i}^{th}$ and $1^{st}$ cycle: $\Delta\mathrm{Q}_{i-1}(V)$ = $\mathrm{Q}_i(V)$ - $\mathrm{Q}_1(V)$, for i>=100 as feature. MAE decreases as the i in $\Delta\mathrm{Q}_{i-1}(V)$ increases. When we use a different ML algorithm (Lasso) for short life cells training and testing with the same features, the error decreases to 5%. This further decreases the aggregate MAE to (22*5.2%+18*8.5%)/40= 6.7%

**Approach 2**

In this approach, we have lowered i in $\Delta\mathrm{Q}_{i-1}(V)$ from 100 to 10 for short life cells and 40 for long life cells with the trade of MAE. When we plot variation in capacity with the cell cycles, there are 2 main observations: 1) The capacity is increasing for initial cycles till $10^{th}$ cycle and at $10^{th}$ cycle its maximum as can be seen in table 1 figures. 2) The capacity starts decreasing from $10^{th}$ cycle, and then it falls till $40^{th}$ cycle, there is local minima at cycle number 40 in the table 2 figures. These two observations give us the incentive to use them as a feature in our model.

We use mean, variance, and minimum of $\Delta\mathrm{Q}_{10-1}(V)$ as features for short-life cells. This model uses RFR and gives an MAE of 11.5%, as shown in figure 1a. While for Long life cells, we use mean, variance, and minimum of $\Delta\mathrm{Q}_{40-10}(V)$ as features, which gives MAE of 13.85% and RMSE of 211 cycles. Looking at the RMSE of Long life cells, its nearly the same as in Severson's paper, and the cells with EOL above 1500 are doing better than the Severson model (figure 1b).

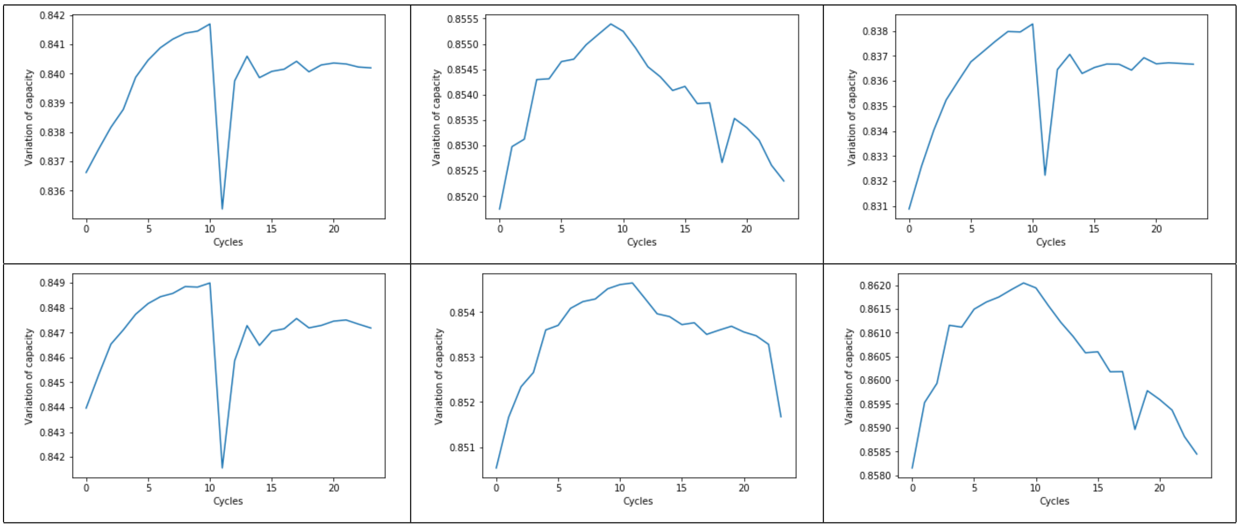



Table 1: Variation of capacity with cell cycles in short life cell

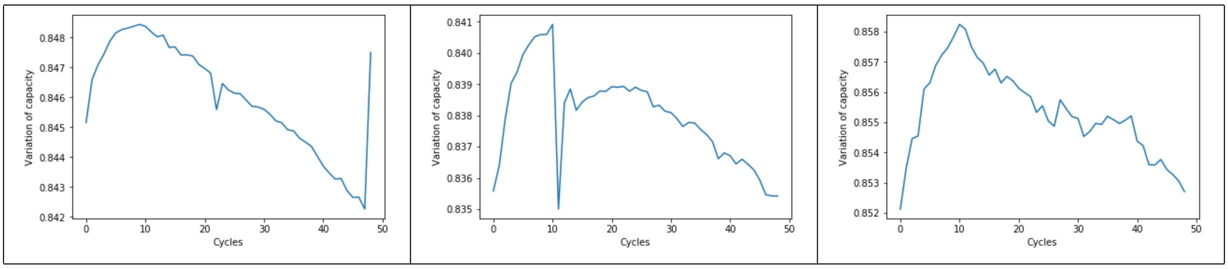
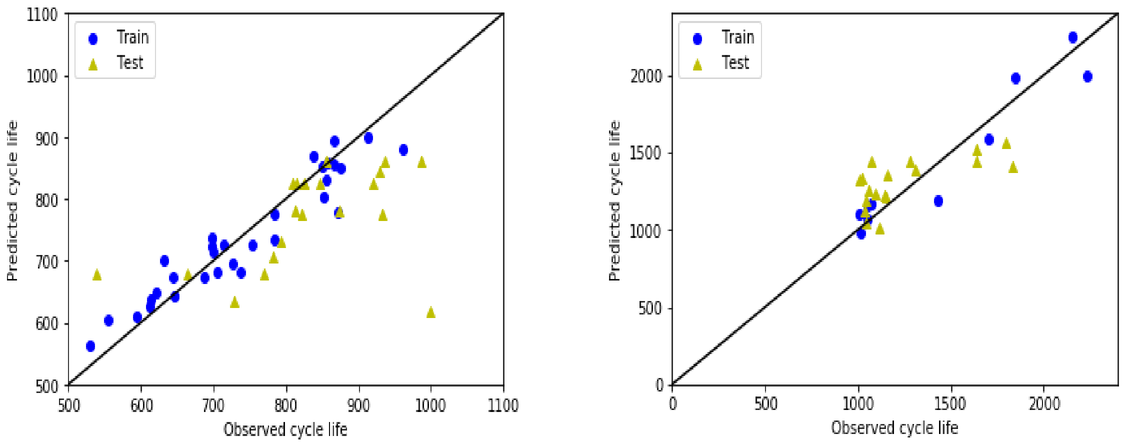



Table 2: Variation of capacity with cell cycles in long life cell




(a) Short life cells

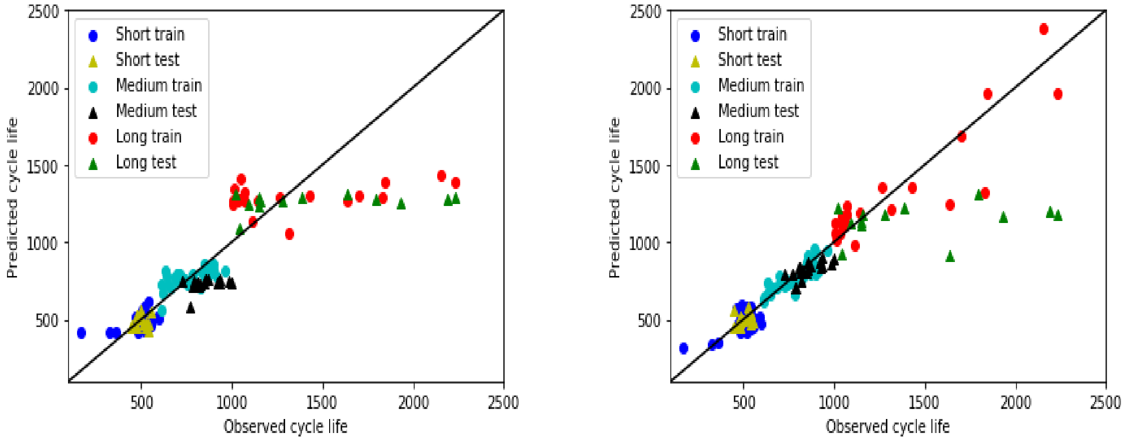
(b) Long life cells

Figure 1: Observed and predicted cycle lives

## 3.2 Early Work

Before reaching the above results, a lot of techniques and ML algorithms were tried to improve the results. Linear regression, Decision Tree, Random Forest, Lasso, Neural Network, Bayesian approach, Support Vector Machine, Relevance Vector Machine were tested to refine the outcome, but there was no significant drop in error. When the ML algorithm didn't work, new features were generated. The mean and variance of change in discharge voltage curves between the i$^{th}$ and j$^{st}$ cycle ($\Delta Q_{i-j}(V)$) for i,j belongs to 100<i<j<10, were generated to see their correlation with the RUL of a cell. In one model, all these 8000 features were concatenated and reduced to 5 features through principal component analysis (PCA) to generate new features that can have a high correlation with the RUL of a cell.

The dataset was also divided into 3 categories 1) Short life cell (EOL $<$ 600) 2) Medium life cell (600 $<$ EOL $<$ 1000) 3) Long life cell (EOL $>$ 1000). Datasets can be categorized by the first 3 cycles of cells. First, we classify a cell into one of these 3 categories and then make predictions with the model in which the cell lies. This model's beauty was that it could accurately predict at any point in life time starting from the 10$^{th}$ cycle for the short life cells, as can be seen in plots of table 3. It makes correct prediction for medium life cells till its EOL after 75$^{th}$ cycle. To make prediction, variance and mean of $\Delta Q_{i-1}(V)$ work great as features in RVM. The plots in figure 2 will give us a better idea of ML predictions. Figure 2a is when mean, variance of $\Delta Q_{10-1}(V)$ is used a feature, we can see that short life cells are fitting well while medium and long life cells are giving the same RUL for every cell. Now looking at figure 2b with mean, variance of $\Delta Q_{100-1}(V)$ as features, medium life cells result have improved drastically, and the long-life cells are performing the same as Severson paper results. In the case of Severson model cell with EOL below 1200 perform better which compensate for the bad performance of cell with EOL greater than 1200, overall resulting in MAE of 9.5%




(a) $\Delta Q_{10-1}(V)$ (b) $\Delta Q_{100-1}(V)$

Figure 2: Variation of data with i=100 and i=10 in $\Delta Q_{i-1}(V)$

By dividing datasets into 3 groups, we have decreased the range of EOL for training and testing datasets. We can observe in table 3 plots that using this approach, short life cells can predict very accurately right after the 10$^{th}$ cycle to EOL. The result can be improved for medium and long life cells if we use a data diverse in term on their EOL and homogeneous in term of the number of cells between a particular range of EOL (e.g., 10 cells between 700-800, 10 between 800-900).

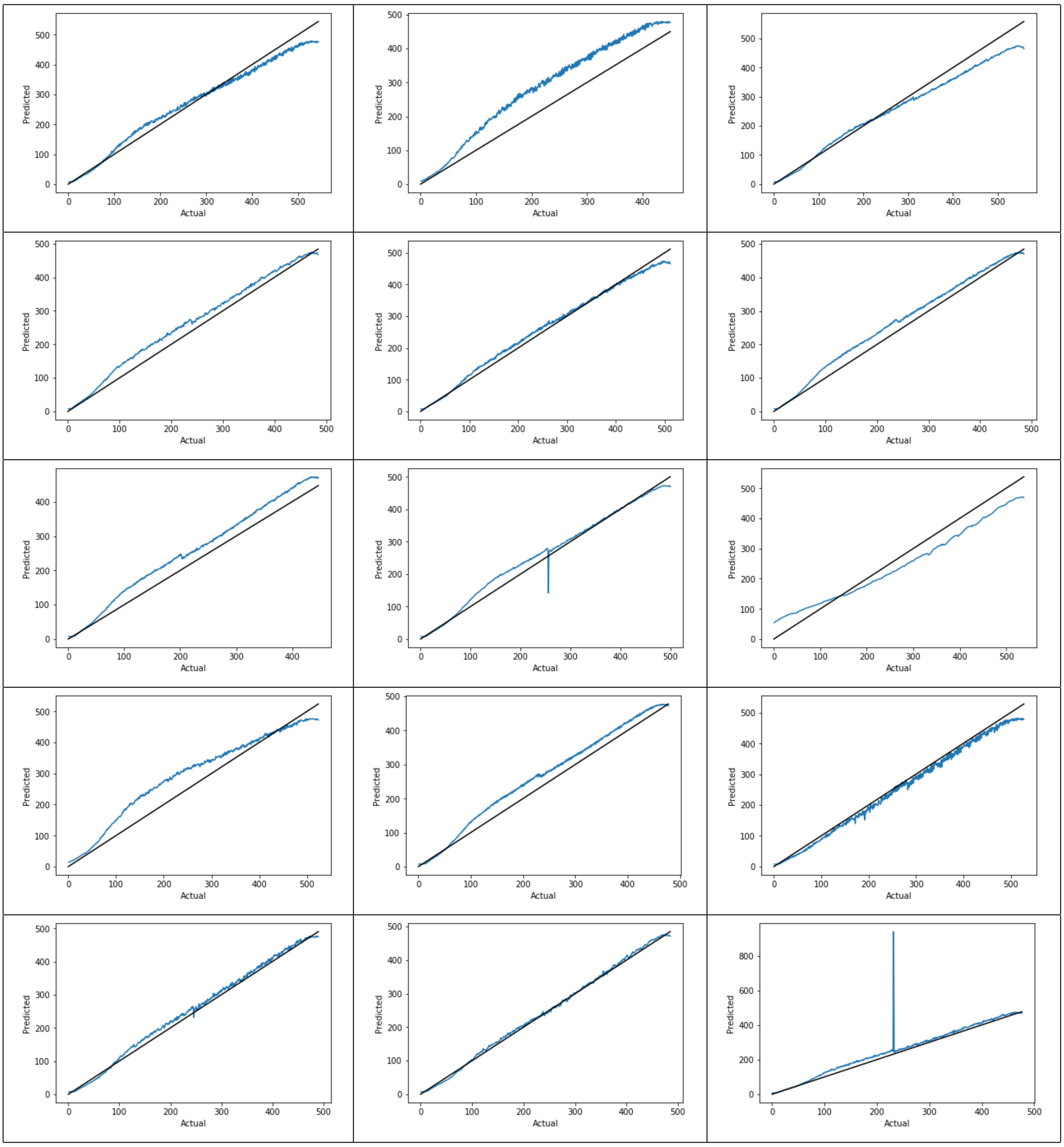


Table 3: Testing set of short life cell

| Cells - EOL | RMSE (cycles) by our model | RMSE (cycles) in Zhang Paper |
|---|---|---|
| 45C02 - 396 | 28.6 | 32 |
| 35C02 - 252 | 28.8 | 36.5 |
| 25C05 - 150 | 10.5 | 15.5 |
| 25C06 - 120 | 13.7 | 18 |
| 25C08 - 38 | 9.7 | 5 |

Table 4: Comparison of our results with Cambridge paper's result

# 4 Cambridge Dataset

## 4.1 Discussion

Electrochemical Impedance Spectroscopy (EIS) is an electrochemical analysis technique that finds much application in studying batteries, fuel cells, and many other systems. EIS involves looking at the impedance characteristics of an electrochemical system over a range of frequencies. Power fade is caused by the increase in the impedance of a cell with the cell aging. EIS is used to measure the impedance at a spectrum of frequencies (range of 0.02 Hz–20 kHz), which is used to know the health of a battery [5]. In [5], they have accurately predicted the RUL of batteries at three different constant temperatures, at any point in its life, from a single impedance measurement.

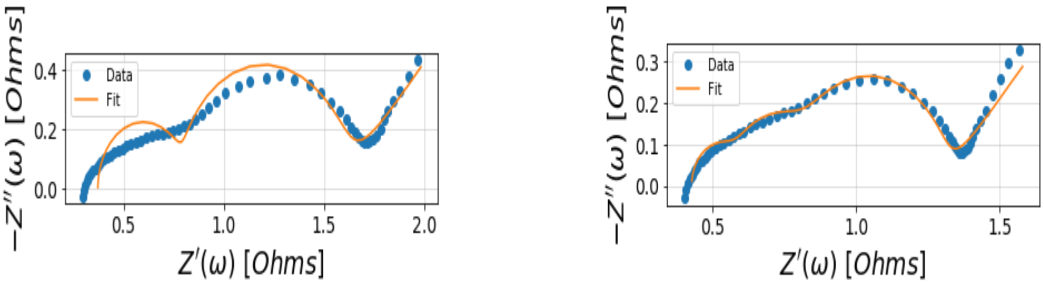
The results in [5] paper is quite fascinating. We will show how a new feature and different ML algorithms have improved the results. The Cambridge dataset contains 12 (45 mAh Eunicell LR2032) LiCoO 2 /graphite cells cycles at three different temperature 25 °C (25C01–25C08), 35 °C (35C01 and 35C02) and 45 °C (45C01 and 45C02) but the first 30 cycles of all cells underwent at 25 °C before different temperatures were set. The cells underwent 2C CC (constant current) discharge up to 3V and 1C CC-CV (constant current–constant voltage) charge up to 4.2 V with EIS measurement at every even-numbered cycle. The model is trained with a mixed set of temperature cells, four 25 °C (25c01 - 25c04), one 35 °C (35C01), and one 45 °C (45C01) cells, and all others as a testing set. Relevance Vector Machine (RVM) algorithm is used to predict the RUL of cells with new features( [$Z_{re}(2)$ - $Z_{re}(1)$, $Z_{re}(3)$ - $Z_{re}(2)$, ... $Z_{re}(60)$ - $Z_{re}(59)$, ... $Z_{im}(2)$ - $Z_{im}(1)$, ... $Z_{im}(60)$-$Z_{im}(59)$]). The new features and old features are used to give improved results, as summarized in the table 4. Looking at the results, the model performs well with the new features, but as the EOL decreasing it has started affecting the performance of the model. The algorithms used in the ML model require sufficient and heterogeneous training data to achieve sound statistical learning. When the training sample is not large enough, statistical learning is not possible, constrained by the curse of dimensionality, or over-fitting will be inevitable. As the training data contains just one cell with EOL less than 50, that's why the model has started performing badly with low EOL life cells. With a different ML algorithm, the model achieves an RMSE (cycles) of 16 cycles for 35C01, but the same algorithm performed badly on other cells.

## 4.2 Early Work

It's challenging to perform an EIS experiment at different frequencies for calculating impedance. The plan was to use Impedance.py (a Python package for analyzing electrochemical impedance spectroscopy (EIS) data) [7] to predict impedance at different frequencies by training the model on a given data set particular spectrum of frequencies. This impedance was further used to improve the prediction of the RUL of a cell.

### 4.2.1 Impedance Model

Impedance.py allows us to build an equivalent circuit model (ECM) of a cell. The parameters (R, C) of ECM are predicted after fitting the equivalent circuit to a spectrum of frequencies. These R, C values give the total impedance of the ECM. Many ECM models (Randles, 2RC, 3RC) were fitted on the data to find out the best ECM, which can emulate the behavior of a cell. It can be seen in figure 3 that 3 RC-Warburg ECM fits better than 2 RC-Warburg ECM. The increase of impedance with the cycles was observed by plotting the Nyquist plot, as shown in figure 4. The predicted 3RC-Warburg model started fitting poorly as the cell ages, three semi-circles in figure 3b, become two with aging. One set of parameters(R, C) starts dominating over the others, which changes three semi-circles to two semis.




(a) 2 RC-Warburg impedance

(b) 3 RC-Warburg impedance

Figure 3: 3 RC-Warburg model fit better than 2 RC-Warburg model

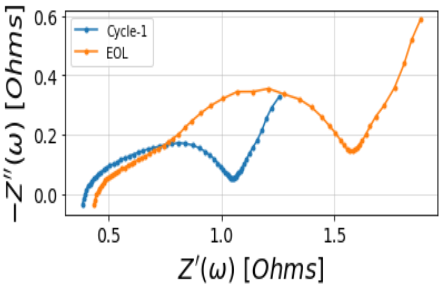
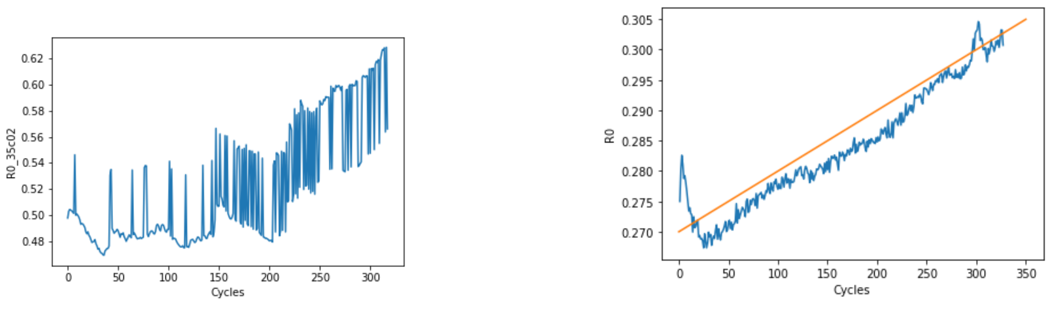



Figure 4: Change in parameters as the cell ages




(a) R0 calculated by Impedance.py

(b) R0 calculated manually

Figure 5: R0 calculated manually through code is much smoother than calculated from Impedance.py

3RC-Warburg model of impedance.py has nine parameters; these can be used as features for predicting RUL. The idea of using predicted ECM parameters as a feature for RUL prediction was dropped due to the generation of highly fluctuating parameters. We verify this by choosing four parameters in the Nyquist plot, which are easy to calculate manually 1) R0 2) R0+R1 3) Local maxima of semi-circles 4) Slope of Warburg impedance line. The plots of these four parameters are very smooth compared to parameters calculated by Impedance.py. The comparison in the plot of R0 of cell 35C01 can be seen in fig 5. Earlier, we thought there is noise in the Cambridge dataset, but we reached this conclusion after verifying manually. We use only these four parameters as a feature, which gives nearly the same result as Cambridge Paper. This shows that if we can correctly predict these nine ECM parameters, these could be used as features in the model. If the problem of highly fluctuating parameters will be fixed in the near future, then Impedance.py can be used to study the effect of degradation (CL, LLI, LAM) with cell aging of cells [6]. Impedance.py gives great insight into the EIS.

# 5 Conclusion

Machine Learning has given us incentives for its future use to research in batteries with its reliable algorithms. We use two different datasets in the two various fields (capacity degradation and EIS) to build two different models to predict the RUL at any point in its lifetime. In the Severson dataset, we divided the dataset into two categories based on cell life, which give our model high power to predict RUL at an early stage of life with accuracy better than existing literature. Using some new EIS features and a different ML algorithm has improved the results of Cambridge datasets. We fit the EIS data using impedance.py to generate high fidelity ECM and study patterns in a cell aging. There is a future scope to accurately make EIS data at any frequency by training with a limited spectrum of frequencies. The parameters of ECM have also shown the potential to be used as a feature in the ML algorithms.

# 6 References


[1] Mikolajczak Celina. Lithium-ion batteries hazard and use assessment. US: Springer; 2011.

[2] Li, Y., Liu, K., Foley, A. M., Zülke, A., Berecibar, M., Nanini-Maury, E., . . . Hoster, H. E. (2019). Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review. Renewable and Sustainable Energy Reviews, 113, 109254. doi:10.1016/j.rser.2019.109254.

[3] Severson K, Attia P, Jin N, Perkins N, Jiang B, Yang Z, et al. Data-driven prediction of battery cycle life before capacity degradation. Nat Energy 2019;4:1–9.doi: 10.1038/s41560-019-0356-8.

[4] Fermín-Cueto, P., Mcturk, E., Allerhand, M., Medina-Lopez, E., Anjos, M. F., Sylvester, J., amp; Reis, G. D. (2020). Identification and machine learning prediction of knee-point and knee-onset in capacity degradation curves of lithium-ion cells. Energy and AI, 1, 100006. doi:10.1016/j.egyai.2020.100006.

[5] Zhang, Y., Tang, Q., Zhang, Y., Wang, J., Stimming, U., amp; Lee, A. A. (2020). Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. Nature Communications, 11(1). doi:10.1038/s41467-020-15235-7.

[6] Pastor-Fernández, C., Uddin, K., Chouchelamane, G. H., Widanage, W. D., amp; Marco, J. (2017). A Comparison between Electrochemical Impedance Spectroscopy and Incremental Capacity-Differential Voltage as Li-ion Diagnostic Techniques to Identify and Quantify the Effects of Degradation Modes within Battery Management Systems. Journal of Power Sources, 360, 301-318. doi:10.1016/j.jpowsour.2017.03.042.

[7] https://impedancepy.readthedocs.io/en/latest/getting-started.html.