Report

Stratos Pateloudis

December 15, 2024

1 A pre-analysis from Severson et all paper

The goal of this project was to explore, and if available, establish relations between parameters of the batteries and their state of health. The first part concerned purely exploratory data analysis, modeling and establishing relations. The second part, involved machine learning and neural networks engineering and testing leading to accurate predictions of batteries characteristics, sometimes measurable and sometimes not. What is crucial to state here, is the fact that while the first part is based on real-world data, the second part is based on artificial data created with PyBAMM at Python.

1.1 Setup

I begin with a pre-analysis of the Severson et all paper, then I compare the model built from data analysis against machine learning (ML) techniques and I am able to construct a ML model that predicts the knee points and end cycle of a battery with high accuracy ($\geq 90\%$).

Let us start with setting up the playground. The data are available here. From a small reading on background information from the available literature, the double exponential fits these kind of data among the best possible ways. A similar path follows a double power function which I will not adapt.

A fit over the Discharge Capacity (QD) in the data using the double exponential function, with respect to the cycles is shown in Fig. 1. The fitting function is given by

$$f(t) = c_0 e^{-c_1 t} + c_2 e^{-c_3 t}, \tag{1}$$

where t represents the number of cycles.

Performing a correlation analysis over the whole data, I find a linear relation between "QD" and "QC" variables as it is shown in Fig. 2. However, here there is no useful information still. I will have to choose one variable to work with. At first site it seems that both "QD" and "QC" are equally good, but I will choose to work with "QD" [1]. The strategy I follow at this point contains the following steps:

- Perform a fit of the form (1) for each "healthy" battery and store the coefficients c_0, c_1, c_2, c_3 in a list. In the end, I have 118 batteries.

[1]Practically one variable might be preferable in a real-life scenario, but this remains to be discussed by an expert on batteries. For the moment it will not be important.

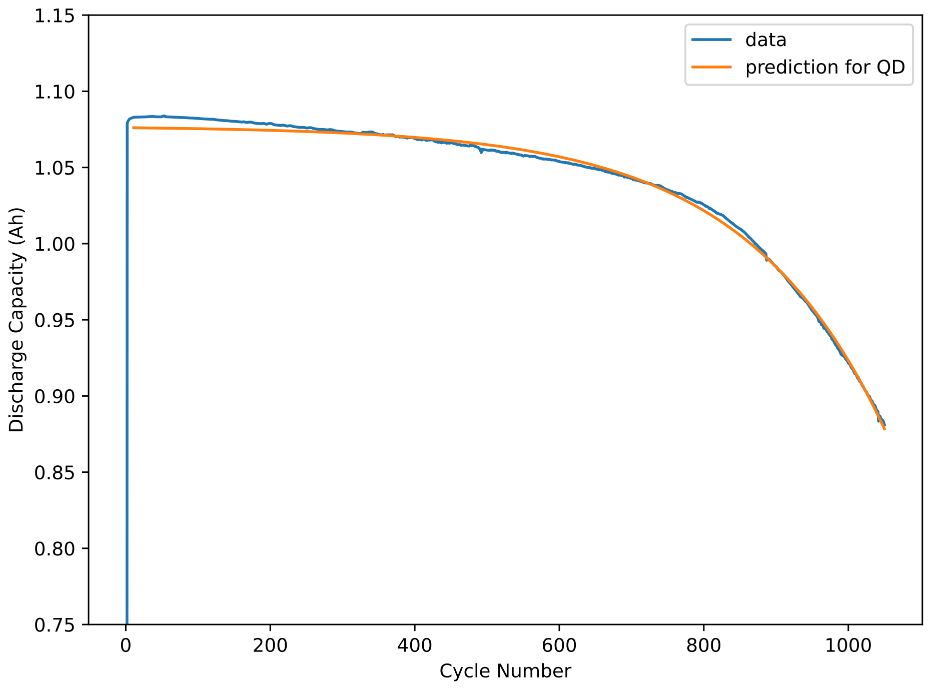

Figure 1: Data for one battery. Discharge Capacity (QD) with respect to the cycles and the fitting function. The same pattern holds for all batteries.

- Perform an analysis over the data to find/check correlations.
- It is important to find the knee-point, that is when the curve starts changing rapidly. So far there is no definition for this point but I am going to define it shortly following a proposal in the literature (see next section).
- Check whether there is a correlation between the knee-point and the predicted maximum cycle from the fitting model. Intuitively thinking one expects to find a correlation because when the curve changes, then one can do a linear fit over the rest of the curve and predict the maximum cycle.
- This will also give us a prediction for the maximum cycle (t_{maxQD}) until failure of the battery (that is $QD = 0$). This point-cycle (t_{maxQD}) corresponds to

$$t_{\mathrm{maxQD}} = \frac{\ln \frac{|c_2|}{c_0}}{c_3 - c_1}. \tag{2}$$

1.2 Detecting the knee-points

It is important to know the knee-points for each battery. One expects that this will help in return to predict the endpoint cycle of the battery. The caveat is that there is no agreement in the definition of the knee-point in the literature. Despite this, I will use the definition and the algorithm presented

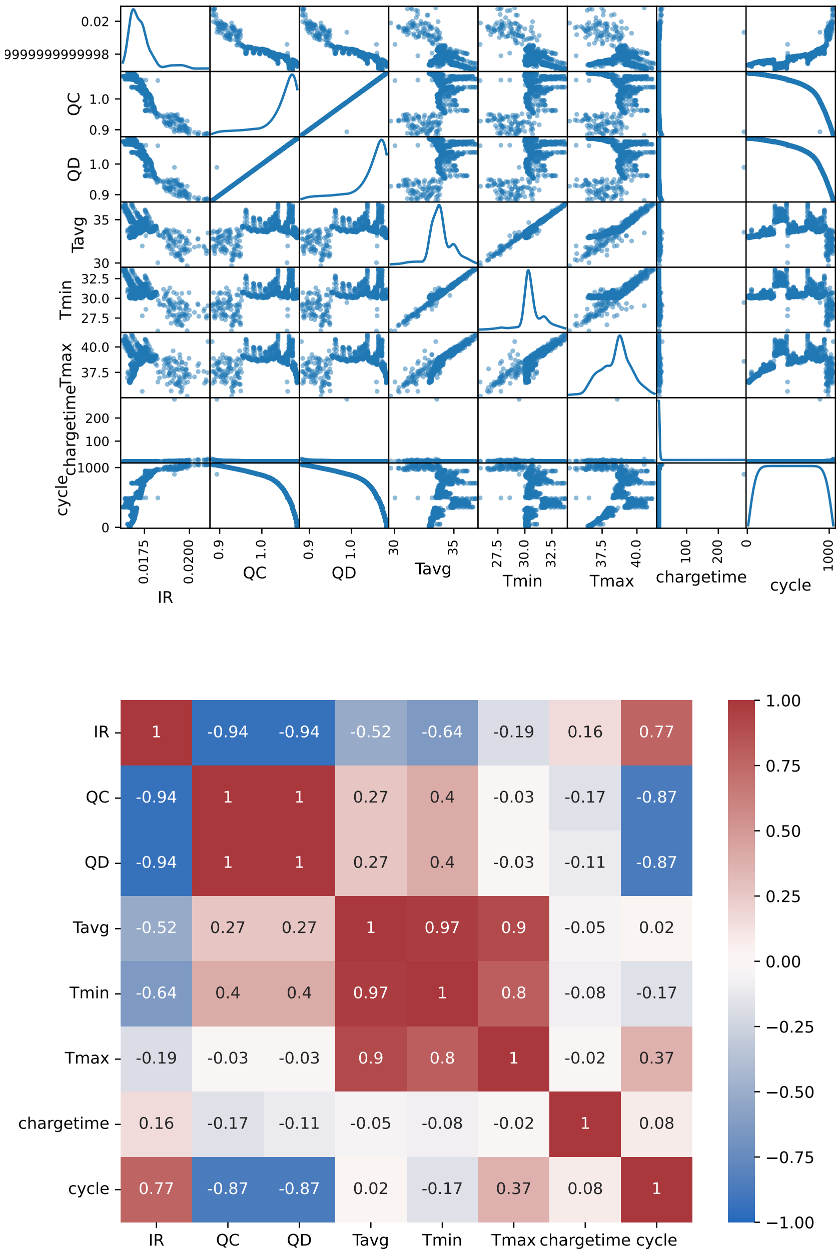

Figure 2: Correlation between data for a specific battery. The same pattern holds for all "healthy" batteries. [Up]: Possible correlations between the data. In particular there is a strong correlation between "QD" and "QC" shown as linear lines. [Down]: correlation parameters for each variable. Close to ± 1 means strong correlation, and close to 0 means no correlation.

here. As we will see, this method will suffice to capture the expected behaviour even when the data are noisy.

Doing so, I am able to detect the knee-points for all batteries. An example is shown in Fig. 3. I did this for all batteries and performed again a correlation analysis between the fitting data, t_{maxQD}

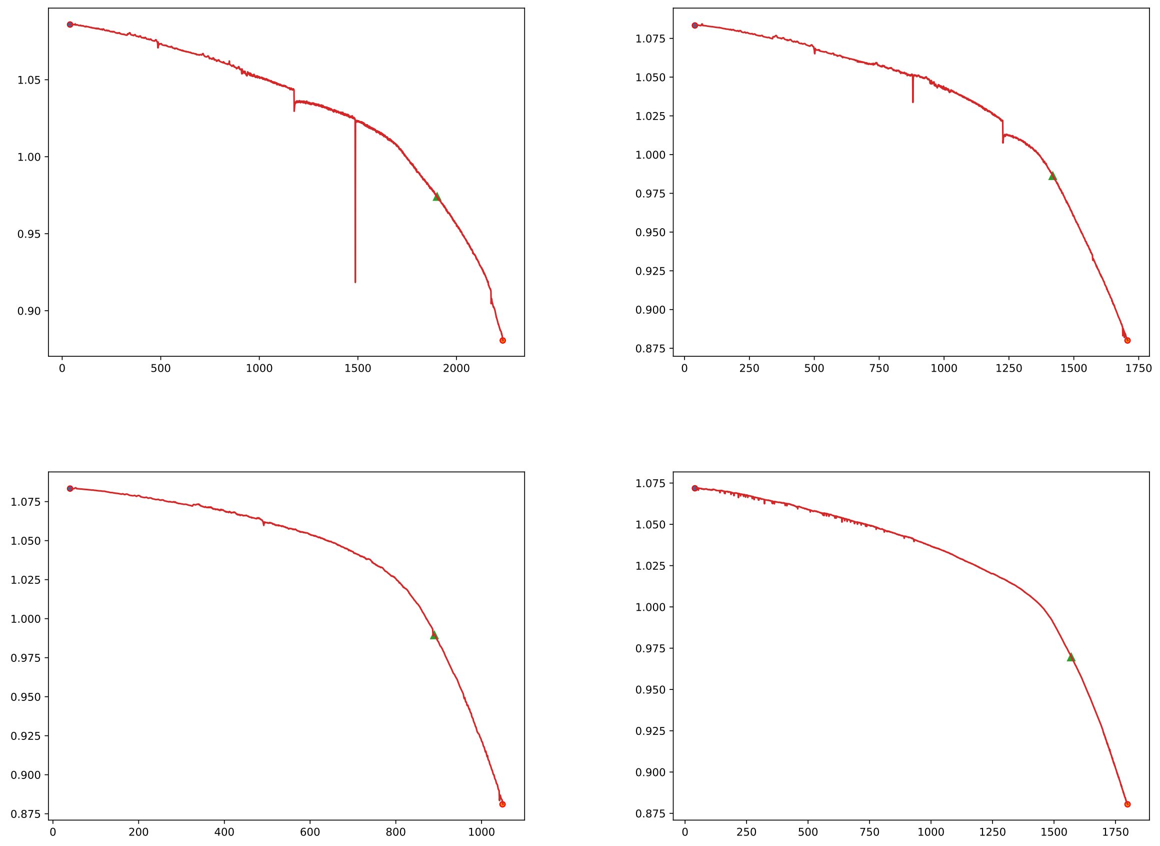

Figure 3: Knee points for batteries. Down-left is the battery presented in before in Fig. 1. It is important to note that this algorithm is valid even though the data are noisy like up-left. The knee-points are shown as green triangles emanating from the intersection of the tangent lines of the red (circle) points.

and t_{knee}. As expected a strong correlation between the knee-point and the end-cycle for each battery is observed. I show this in Fig. 4. A mild correlation between c_3 and t_{knee} or t_{maxQD} is also noted which intuitively makes sense because c_3 controls the fall of the function via the second exponential.

This suggests that a linear fit between the knee-point data and the predicted end-life cycles for all batteries. Before we do this, we note that the detection of knee points does not work correctly for batteries whose c_1 value is positive. These will result in outliers in data giving knee-points below 100 cycles which is aesthetically wrong. Therefore, we are going to remove these outliers and perform the fit. We get as a result a model of the form

$$t_{\text{maxQD}} = 1.43 \cdot t_{\text{knee}} + 37.5,$$

and the fit is shown in Fig. 5 with $R^2 = 0.94$. What this result suggests, is that as soon as we can detect the knee point we can predict the end life of the battery. Detecting the knee point currently is always done by measuring, and it is somewhat pathological because one needs to know the behaviour of the battery around EOL80. On the other hand it should be emphasized that the model is the best representative but not ideal as long as β in $t_{\text{maxQD}} = \alpha \cdot t_{\text{knee}} + \beta$ is not zero. On the other hand a mean error of around 37 cycles over several hundred (or even a thousand) cycles is not that bad. An

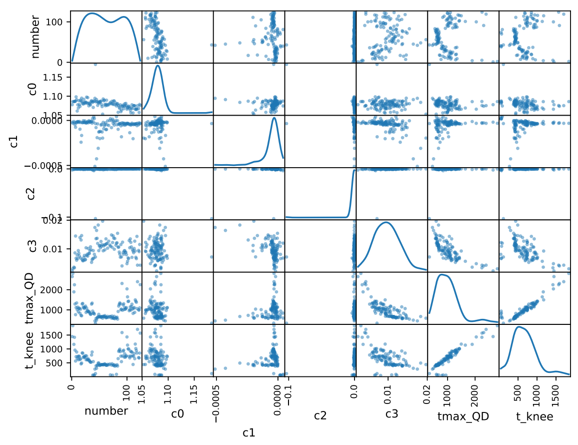
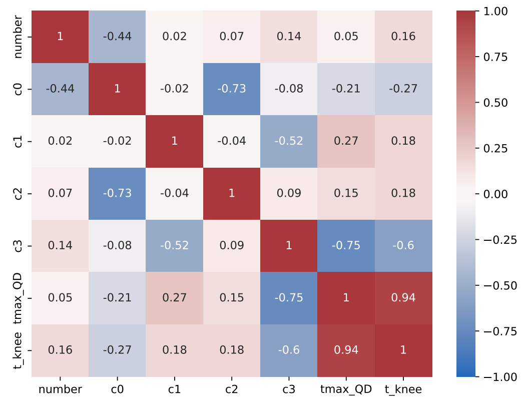
Figure 4: [Left]: Correlation between the characteristics of all batteries. [Right]: There is a strong correlation between the knee-point (t_{knee}) and the end cycle of each battery t_{maxQD}. Mild correlation between t_{maxQD}, t_{knee} and c_3 is noted.

ideal model would be of the form

$$t_{\text{maxQD}} = \alpha \cdot t_{\text{knee}}.$$

What these finding suggest is that as soon as we know, or can detect the knee point of a battery, then we know the end-cycle with high accuracy. Since this merely establishes an important result because

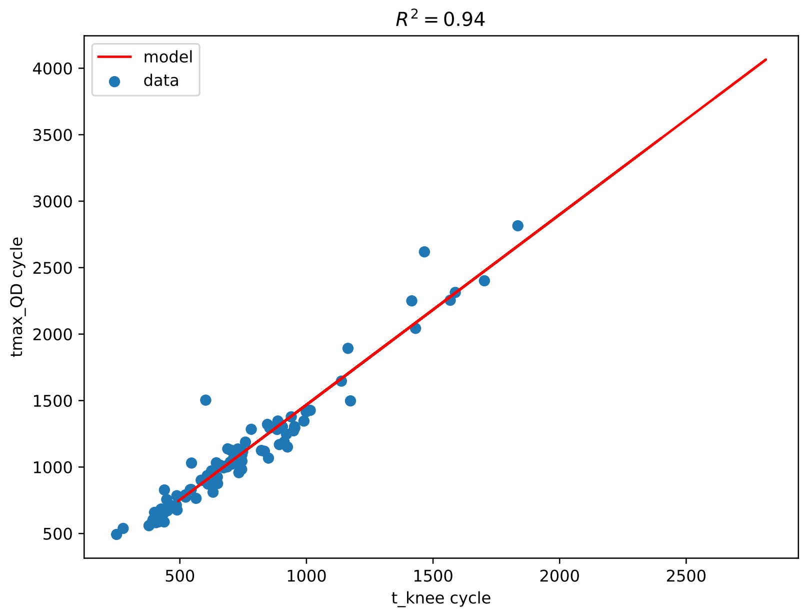

Figure 5: The fit between the predicted maximum cycle and the knee points for all batteries after removing the outliers.

it only provides an insight and is non-predictive we now turn to a more powerful and predictive technique.

1.3 Machine learning tools

An interesting insight is to try to predict the future of a battery, specifically the EOL80 points by using data from relatively few cycles in the beginning. To avoid an over-fitting analysis and given the lack of data the first 100 cycles appeared sufficient for this. By parametrising the measured data in a particular way one is able to train a model over these measurements and predict the future. This however, uses an evolution function, which describes the capacity (QD) at a given time and such a function is given by the double exponential

$$QD(t) = c_1 e^{c_2 \cdot t} + c_3 e^{c_4 \cdot t}. \tag{3}$$

The idea is to train a ML model to predict c_1, c_2, c_3, c_4 over the first 100 cycles and ask for which cycle (noted as t_{EOL80}) QD, hits EOL80. Such a model is succeeds with a relatively good accuracy. This is shown in Fig. 1.3.

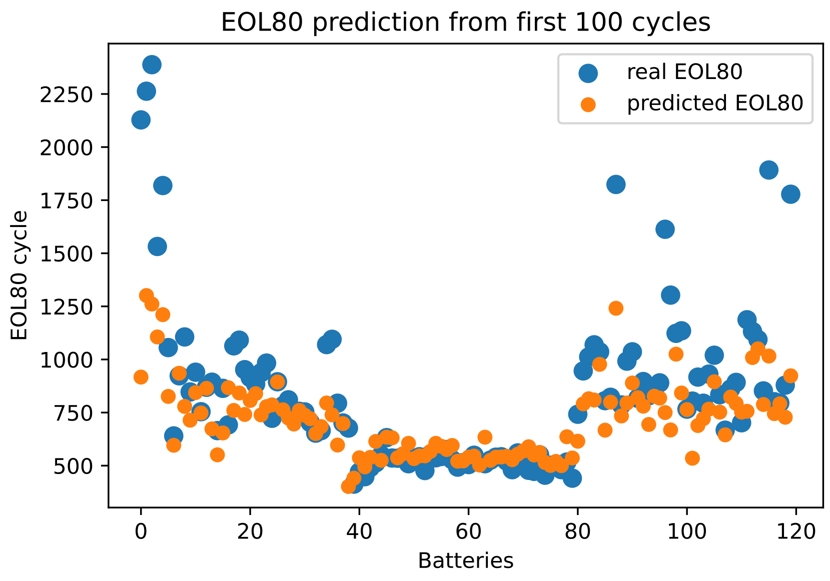

Figure 6: Machine learning prediction for each battery (x-axis) against the actual EOL80 cycle predicted by the fit of the double exponential model over QD.

The model cannot perform more accurate than that avoiding over-fitting given the lack of data available. In addition it worths stressing that one cannot entirely trust this particular model, than rather use it as evidence and a guide that, given a fair amount of data, such a task is possible.

This concludes the analysis and the willingness of people to contribute to this task, since there is no available resources. To this end, one would like to create data.

2 An analysis on artificial data

In this task we used artificial data produced by PyBAMM. The protocol of the battery (up the chemical components) was chosen to be quite close to the batteries used in Severson et all, however the lack of precise knowledge of the battery system and how to clone the chemistry in PyBAMM failed to exactly reproduce the system.

Nevertheless, I set-up a battery protocol with precise characteristics up to the chemistry which was used as a variable of the model.

For this task, a ML model was setted up to predict non-measurable, internal states of the battery such as Open Circuit Voltage (OCV) by using measurable variables, such as the Terminal Voltage (TV), the charging current, etc.

The key was to find a relationship between the OCV and TV which again was appearing as a specific function and the task again was to try predicting the fitting coefficients.

The artificial dataset was produced by PyBAMM, a ML model was trained over these data and tested against a *new*, unknown (to the model) dataset by predicting the fitting coefficients with accuracy 85%.

In this way, one is able to predict internal (unmeasurable) states of the battery system by knowing measurable variables such as TV. This would help, in return, to predict the state of health (SOH) of the battery, and therefore t_{EOL80}.

Given the goodness of fit of $f(t)$ from

$$OCV - TV = f(t), \tag{4}$$

the model not only can predict the current SOH, but also evolving with $f(t)$ one could now this in the feature by doing one measurement in the very beginning (see Fig. 2). This gives a powerful prediction mechanism, that given the TV of a battery at time t_0 one not only can predict the OCV at time t_0 accessing the SOH of the battery but it is also possible to predict the SOH at later times using the fitting function.

This reveals a physical prediction model for the SOH of a battery, but on the other hand it has limitations. First of all, the model is not perfect since not all the fits are perfect. In fact there are cases that the model predicts wrong OCV from the very beginning. This is perhaps a chemistry-related problem, as far as we understand, because not all parameters (and perhaps functions) are suitable across different chemistries.

Secondly, depending on the model, the predicting power varies. As it can be seen from Fig. 2, choosing the initial conditions $t_0 = 0$, nothing guarantees that OCV and TV evolve along the function $f(t)$. As this is the case for the bottom-right sub-figure, it is not for the bottom-left after 50 minutes, showing a clear chemistry-dependent pattern.

On the other hand, we can overcome this problem by doing specific measurements at time $t_1 > t_0$, setting a further predicting horizon and repeating the process. Currently t has units of seconds, but a new protocol can be setted-up for larger time units.

The question now, is whether or not we can lift the evolving function and replace it by an AI model. Even this might be possible to access OCV at the moment of measurement $t = t_0$, there is a strong doubt that it will have forecasting power, since this requires knowing a function along which the OCV evolves.

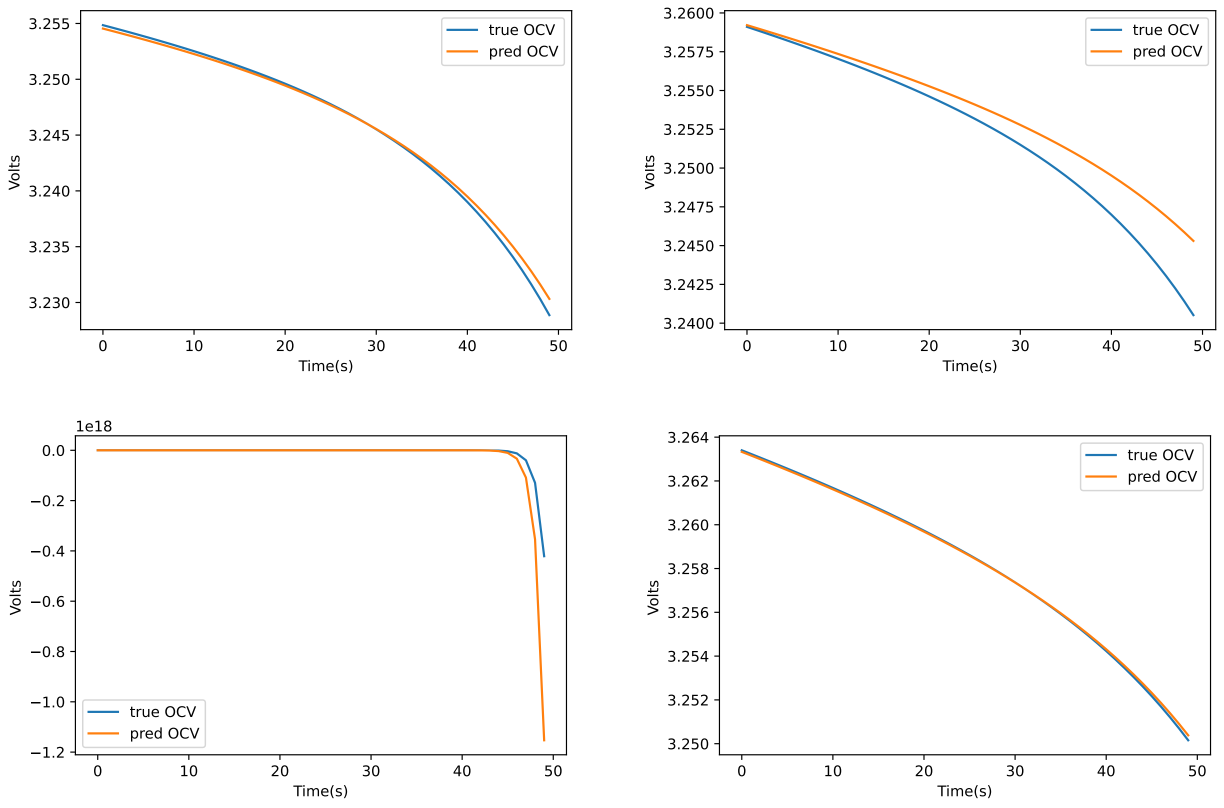

Figure 7: Machine learning prediction for the OCV of batteries against the real value, with different charging/discharging currents, chemistry, etc, which has $TV = 3.3V$ at $t_0 = 0$. Given this, one can predict how OCV will evolve. Bottom-left situation might seem as an outlier but in fact it is not. The chemistry of this battery, under specific charging/discharging currents, goes off the protocol concerning cut-off voltages well before the time interval shown here. This, in return, affects the fitting function which does not perform well in all time ranges.