Masterthesis

Machine Learning for a fully automated, self learning test factory for battery cells

Maschinelles Lernen für eine vollautomatisierte, selbstlernende Prüffabrik für Batteriezellen

Autor	●	**Valluru, Govilnag**
Datum	●	**03.08.2020**
Version	●	**1**

SYSTECS
Informationssysteme GmbH
Bahnhofstraße 11
70771 Leinfelden-Echterdingen
Deutschland

Telefon +49 711 722 312 10
www.systecs.com

Hinweise zum Datenschutz unter
www.systecs.com/datenschutz

Abstract

The Main Objective of this Master thesis is to develop a machine learning based battery model to predict the execution time for charging and discharging cycles of different types of Lithium-ion battery cells from different manufacturer's for different Test strategies for battery cell testing, in order to have an optimal planning of the test orders in a fully automated test bed for battery cells.

Lithium-ion-based battery systems are an efficient alternative energy storage system for electrically propelled vehicles. The requirements for lithium-ion based battery systems for use as a power source for the propulsion of electric road vehicles are significantly different from those batteries used for consumer electronics or stationary usage.

In the light of the rapid spread of hybrid electric vehicles and the emergence of battery and plug-in hybrid electric vehicles, a standard method for testing performance requirements of lithium-ion batteries is indispensable for securing a basic level of performance and obtaining essential data for the design of vehicle systems and battery back. For automobile application, it is important to note the usage specificity; i.e. the design diversity of automobile battery packs and systems, specific requirements for cells and batteries corresponding to each such designs. In order to accomplish this Test Factory for battery cell testing are built to research the behavior of battery cells like static and dynamic characterization, Parameterization of battery models etc.

Machine learning is one of the promising areas in order to develop surrogate models for nonlinear system modelling to predict or to classify the variables. Battery execution time has a nonlinear behavior with the given input parameter set. Battery modelling with the ML algorithms will be good solution to represent the digital twin of a Battery.

1 Contents

Acknowledgement

The Master Thesis opportunity I had with SYSTECS Informationssysteme GmbH was a great chance for me to learn and to develop professionally. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be part of it. I am also grateful for having a chance to meet wonderful people and professionals during my Master Thesis period.

I would like to express my gratitude to Olaf Meyer for his patience guidance, enthusiastic encouragement and belief in me during this Master thesis work. I express my deepest thanks to Dr. Thomas Zurawka for useful critiques and guidance.

Finally, I wish to thank my parents for their support and encouragement throughout my Master Thesis.

Table of Abbreviations

EV	Electric Vehicle
SOC	State of Charge
OCV	Open Circuit Voltage
Amps, A	Amperes
V	Volts
°C	Degree Celsius
IEC	International Electrotechnical Commission
BEV	Battery Electric vehicle
PHEV	Plug-In Hybrid Electric Vehicles
HEV	Hybrid Electric Vehicles
CV	Constant Voltage
C	Coulomb
mA	Milliamperes
U	Voltage
min	Minimum
ch	Charge
dch	Discharge
SCH	Standard Charge

SDCH	Standard Discharge
Ah	Ampere hour
CC	Constant Current
PDE	Partial Differential Equations
HPPC	Hybrid Pulse Power Characterization
ML	Machine Learning
AI	Artificial Intelligence
NASA	The National Aeronautics and Space Administration
SVM	Support Vector Machine
SVR	Support Vector Regression
SVC	Support Vector Classification
ID3	Iterative Dichotomiser 3
CART	Classification and Regression Tree
RSS	Residual Sum of Square
SSR	Sum of Square Residual
$\boldsymbol{R^2}$	Coefficient of Determination
$\boldsymbol{SS_{res}}$	Residual Sum of Square
$\boldsymbol{SS_{tot}}$	Total Sum of Square
RMSE	Root Mean Square Error

1 Introduction

The extension of battery application to automotive industries as a fuel for the Electric Vehicles (EV) has led to increase the research on the optimization of battery performance with a reduced battery cost. The Batteries to meet the market requirements for long-lasting high-energy performance for a dynamic charge and discharge processes and important characteristics for energy storage systems are the optimization of lifetime, safety, power, energy and costs. In order to evaluate the required conditions batteries must be tested with various testing strategies in a test bed with a controlled climate chamber and conduct the battery charging and discharging cycling procedures.

In a testbed for a battery cell there are various basic and lifetime tests (input/output test, Rated Capacity test, Cyclic Aging test etc.) are conducted. Some of the required specifications of the cells like nominal capacity, charging & discharging characteristics are mentioned in a datasheet provided by the corresponding battery cell manufacturer. In order to conduct different test strategies for different battery cells, an optimal plan to run the test orders is required. For this the execution time of battery cell charging and discharging times are to be determined and provided for each test bed in a Test Chamber of a Test Factory. In order to achieve this a battery model is to be designed to determine the execution time of battery cell charging & discharging from the history of measured data.

In this thesis concentrated mainly on the prediction of execution time of charging and discharging process for basic test methodology with general versatility, which serves a function in common primary testing of lithium-ion cells to be used in a variety of battery systems. If the execution time for different Lithium-ion battery cells is estimated, the test engineers can plan the implementation of tests in a test bed for each battery cell type.

In this chapter discussed about the aim of the thesis, software tools and data utilized to achieve the goal of the Master thesis. For battery model development utilized **MATLAB/Simulink** used and after finalizing the steps in MATLAB, implemented all the steps in Microsoft Azure Machine Learning Studio.

In the next chapter, **Battery**, discussed about the basics of battery cell, operation and their specification.

In the chapter, **Battery Test Factory**, discussed about the procedures of battery cell testing in a Battery test factory

In the chapter, **Battery Modelling**, discussed about various types of battery modelling types and finalizing a battery modelling to accomplish the aim of the thesis.

In the chapter, **Battery Model development**, discussed the steps & procedures to design a battery model.

In the chapter, **Results & Discussions,** presented the results of the designed model and their limitations are further discussed.

In the final chapter, **Conclusion**, concluded the aim of the thesis and further improvements.

2 Battery:

A battery [1] is a device that converts the chemical energy contained in its active materials directly into electric energy by means of Electro-chemical oxidation-reduction (redox) reaction. This type of reaction involves the transfer of electrons from one material to another through an electric circuit.

Figure 1: Lithium-Ion Cell

While the term Battery is often used, the basic electrochemical unit being referred to is the "cell" as shown in Figure 1. A battery consists of one or more of these cells, connected in series or parallel, or both depending on the desired output voltage and capacity.

The cell consists of three major components:

1. The anode or negative electrode—the reducing or fuel electrode—which gives up electrons to the external circuit and is oxidized during the electrochemical reaction.

2. The cathode or positive electrode—the oxidizing electrode—which accepts electrons from the external circuit and is reduced during the electrochemical reaction.

3. The electrolyte—the ionic conductor—which provides the medium for transfer of charge, as ions, inside the cell between the anode and cathode. The electrolyte is typically a liquid, such as water or other solvents, with dissolved salts, acids, or alkalis to impart ionic conductivity. Some batteries use solid electrolytes or gel-type polymer electrolytes, which are ionic conductors at the operating temperature of the cell.

The most advantageous combinations of anode and cathode materials are those that will be lightest and give a high cell voltage and capacity. Lithium, the lightest metal, with a high value of electrochemical equivalence, has become a very attractive anode as suitable and compatible electrolytes and cell designs have been developed to control its activity. The cathode must be an efficient oxidizing agent, be stable when in contact with the electrolyte, and have a useful working voltage. The electrolyte must have good ionic conductivity but not be electronically conductive, as this would cause internal short-circuiting. Other important characteristics are nonreactivity with the electrode materials, little change in properties with change in temperature, safety in handling, and low cost.

Electrochemical cells and batteries are identified as primary (non-rechargeable) or secondary (rechargeable), depending on their capability of being electrically recharged.

Primary Cells or Batteries: These batteries are not capable of being easily or effectively recharged electrically and, hence, are discharged once and discarded. The primary battery is a convenient, usually inexpensive, lightweight source of packaged energy for portable electronic and electric devices, lighting, digital cameras, toys, memory backup, Global positioning system devices, and a myriad of other applications.

Secondary or Rechargeable Cells or Batteries: These batteries can be recharged electrically, after discharge, to their original condition by passing current through them in the opposite direction to that of the discharge current. They are storage devices for electric energy and are known also as “storage batteries” or “accumulators.” Secondary battery is used or discharged essentially as a primary battery but recharged after use rather than being discarded. Secondary batteries are used in this manner as, for example, in portable consumer electronics, such as cell phones, laptop computers, power tools etc., for cost savings (as they can be recharged rather than replaced) and in applications requiring power drains beyond the capability of primary batteries. Electric vehicles (EVs) and plug-in hybrid PHEVs also falls into this category.

The chemicals in the battery will ultimately reach a state of equilibrium. In this state, the chemicals will no longer tend to react, and the battery will not generate any more electric current. At this point, the battery is considered ‘dead’.

2.1 Factors Effecting Battery performance:

Battery cell characteristics are varied dependent on the charge cycle, load cycle (Discharge cycle), over lifetime including many factors like internal chemistry, current drain, and temperature.

At low temperatures, a battery cannot deliver much power. In cold climates some car owners install battery warmers, representing a small electric heating pads that keep the car battery warm.

For a rechargeable battery lifetime means either the length of time a device can run on a fully charged battery or the number of charge/discharge cycles possible before the battery is dead.

Disposable batteries lose 8 to 20 percent of their original charge per year when stored in room temperature (20-30 °C) and this phenomenon is called as self-discharge. Self-discharge can happen due to non-current producing “side” chemical reactions that occur in cell when there is no load applied.

2.2 Operation of a cell:

Discharge: When the cell is connected to an external load, electrons flow from the anode, which is oxidized, through external load to the cathode, where the electrons are accepted, and the cathode material is reduced [2]. The electric circuit is completed in the electrolyte by the flow of anions (negative ions) and cations (positive ions) to the anode and cathode, respectively *Figure 2*.

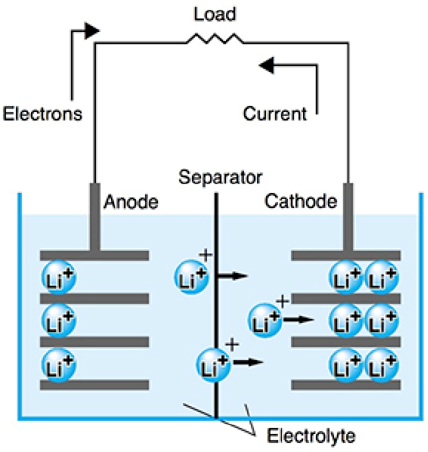


Figure 5: Discharge of a cell

<u>Charge:</u> During the Recharge/Charging of a rechargeable or storage cell, the current flow is reversed, and oxidation takes place at the positive electrode and reduction at the negative electrode, as shown in figure. As the anode is, by definition, the electrode at which oxidation occurs and the cathode the one where reduction takes place, the positive electrode is now the anode and the negative the cathode *Figure 3*.

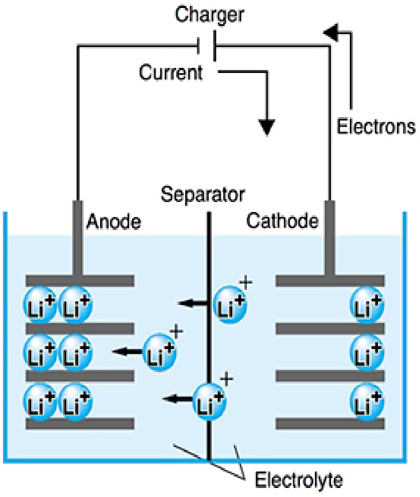


Figure 6: Charge of a cell

2.3 Battery Specifications:

Capacity:

Battery capacity is defined as total number of ampere-hours that can be withdrawn from a fully charged battery under specified condition.

$$C = I * T$$

C- Capacity of battery in Ah

I - Current in amperes

T- Discharge time in hours

State of charge:

State of charge is the current state of battery with respect to the available capacity during charging and discharging profiles. State of charge of a battery is defined as the ratio between the current capacity of the nominal capacity. It is expressed as the following equation.

$$SOC(t) = SOC_0 - \frac{1}{\eta * C_{nom}} \int_0^t I(t)\ dt$$

SOC -State of charge at time instant `t´

SOC_0 -Initial or full charged state

η -Coulomb efficiency of battery

C_{nom} -Nominal capacitance of the battery

$I(t)$ -Instantaneous current which is positive for discharging and negative for charging

Open Circuit Voltage:

Open circuit voltage (OCV) is the potential difference between the terminals of battery when no load is connected. For batteries, OCV is an important parameter to determine the State of charge of the cell, as it is in steady state. The accuracy depends on the characteristics between voltage over OCV curve.

Cut-off Voltage:

For a cell cut-off voltage is the voltage at which cell is considered fully discharged, beyond this discharge could harm the cell. The cut-off voltage is chosen so that maximum useful capacity of the battery is achieved. For testing the capacity of cell, a cut-off voltage of 1.0V is used. This cut-off voltage is defined by the manufacture and it is different for different types of battery cells.

Nominal Capacity:

Nominal capacity is the total Amp-hours available when the battery is discharged at a certain discharge current from 100% to the cut-off voltage. Capacity is calculated as product of discharge current (Amps) and time taken to discharge.

Nominal Voltage:

Nominal voltage is the average voltage a cell output when charged. The nominal voltage must be lower than the rated voltage the nominal voltage of a battery depends on the chemical reaction behind it. For Lithium polymer, Nominal voltage is 3.7V.

Rated Capacity:

Supplier's specification of the total number of ampere hours that can be withdrawn from a fully charged battery cell for a specified set of test conditions such as discharge rate, temperature, and discharge cut-off voltage.

Battery Pack:

Mechanical assembly comprising battery cells and retaining frames or trays, and possibly components for battery management.

State of Health:

State of health is the measure that gives the general condition of battery and its ability to deliver at specified performance in comparison to the fresh battery.

Rated Voltage:

Rated voltage is nothing but the voltage value that has given by the manufacturer representing the safest maximum voltage it can work without reducing cell life span.

3 Test Factory for Battery Cells:

In the next few years the automotive industries will bring variety of electric vehicles in to market. One of the essential parts in electric vehicles is Battery pack that consists of several battery cells. These battery cells before use must be extensively tested to be efficient for the operation of vehicles. For this battery testing factories are built in present and future to research the behavior of battery cells. As show in the *Figure 4* Test Factory mainly comprised of:

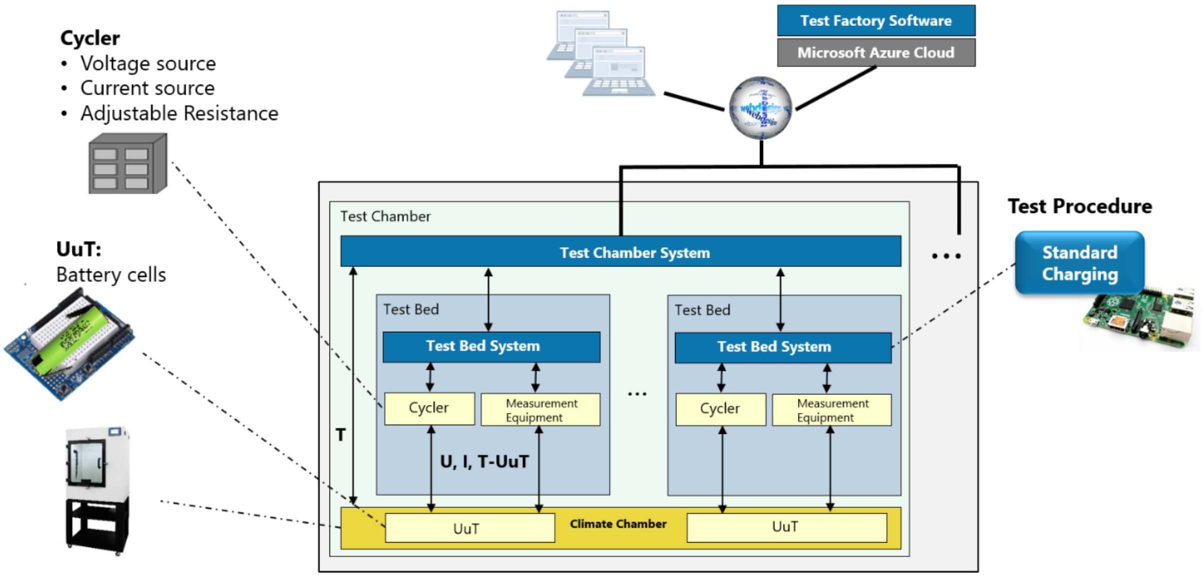


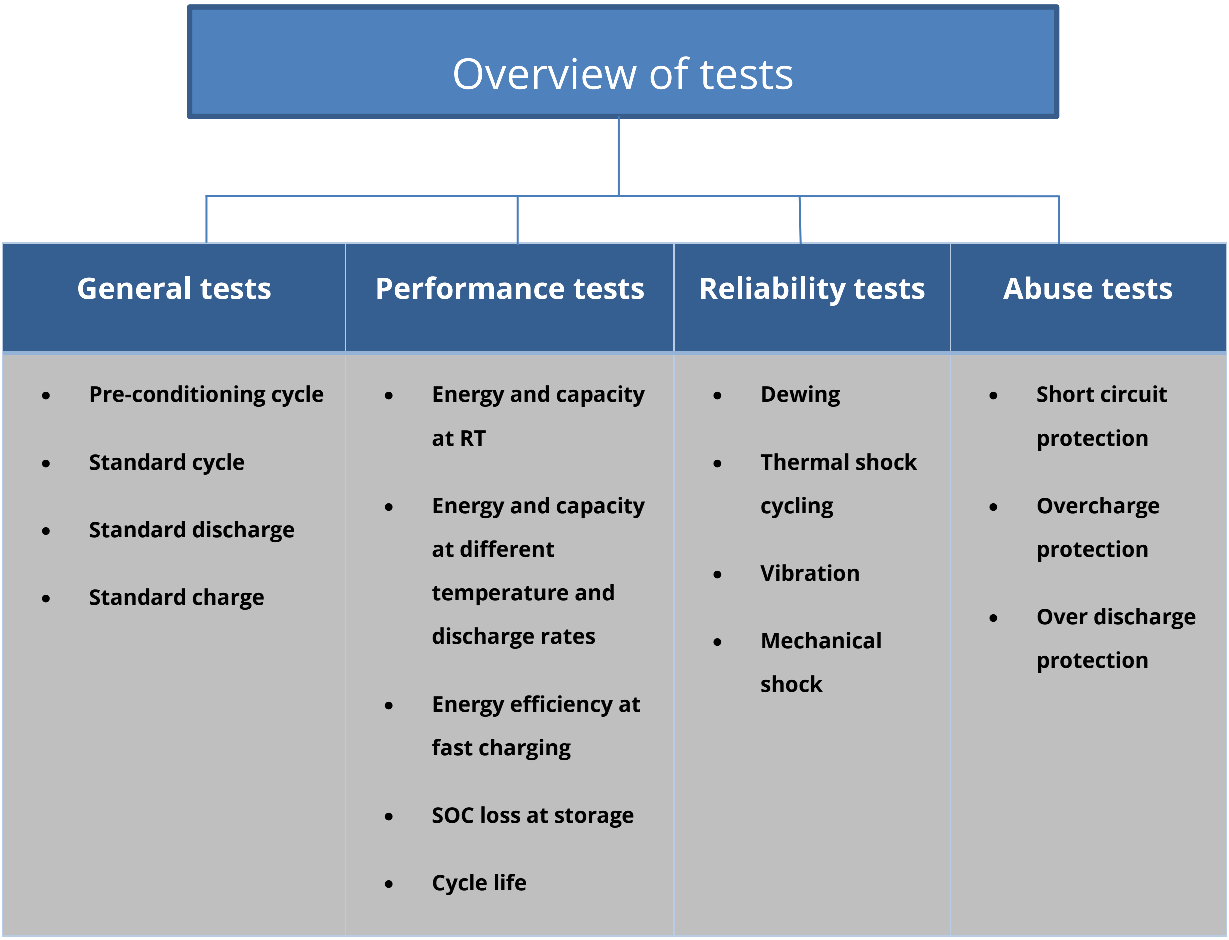
Figure 7 Schematic diagram of Test Factory set up

Test Chamber System: A chamber containing several Test Beds and a climate chamber in which battery cells can be tested by carrying out various Test procedures.

Test Bed: A Test bed consists of cycler for controlling the input and output current and voltage for battery cell and measurement equipment for measuring the parameters Current(A), Voltage(V), Climate chamber temperature(°C), Battery Cell Temperature(°C).

Climate Chamber: This is used to generate and maintain a specified climate, usually temperature and humidity for the duration of the test.

The Test Factory for Battery cells or Battery pack is fully automated by an innovative software technology thereby conducting Test strategy experiments to study the battery cell charge and discharge process in the laboratory setup simulated in a controlled climatic condition. There are different types of comprehensive tests conducted in an automated battery test system as per given automobile industry standards in order to characterize the battery performance. According to the summary of International standard IEC 62660 series, published under the general title Secondary lithium-ion cells for the propulsion of electric road vehicles and from the detailed documentation presented by BMW AG to safeguard lithium-ion cells for BEV, PHEV/HEV vehicle, there are many basic and lifetime tests (input/output test, OCV, Pulse power driving profile, Power pulse according to specification, Rated capacity, cyclic service life. some of the tests conducted are given in the *Figure 5.*



Overview of tests

General tests

- Pre-conditioning cycle
- Standard cycle
- Standard discharge
- Standard charge

Performance tests

- Energy and capacity at RT
- Energy and capacity at different temperature and discharge rates
- Energy efficiency at fast charging
- SOC loss at storage
- Cycle life

Reliability tests

- Dewing
- Thermal shock cycling
- Vibration
- Mechanical shock

Abuse tests

- Short circuit protection
- Overcharge protection
- Over discharge protection

Figure 8: Overview of tests conducted for a battery cell

Some of the Test strategies considered during this Master Thesis are:

- **Standard charge**-This test is performed to ensure standard charge procedure for Li-ion cell at a room temperature (25°C) in a climate chamber if not instructed differently.

  ➢ Procedure**:**

  1. Charge with 1C (PHEV/HEV) or 1/3C(BEV).
  2. Stop under following conditions:
     a. Specified SOC has been reached
     b. At reaching U (max, ch) the charging procedure is continued CV until the current reaches 20mA.

- **Standard discharge**- This test is performed to ensure standard discharge procedure for Li-ion cell at a room temperature (25°C) in a climate chamber if not instructed differently.

  - Procedure:

    1. Discharge with 1C (PHEV/HEV) or 1/3C(BEV).
    2. Stop under following conditions:
        a. Specified SOC has been reached.
        b. U (min, dch) has been reached.

- **Preconditioning/ incoming inspection-** This test is conducted in order to ensure the usability of the cell for further testing's. Furthermore, the conditioning of the cell guaranties that all cells are in comparable state for the further tests.

  - Procedure:

    1. Visual check for faults (Leakage, loose connectors or similar faults)
    2. Determination of the open circuit voltage (at delivery state)
    3. Conditioning: Capacity measurement

- **Capacity measurement-** This test describes the capacity measurement procedure to determine the reference value for the SOC adjustment at room temperature.

- Procedure:

1.	Temperature adjustment to 25°C in climate chamber
2.	Charge with SCH (Standard Charge procedure)
3.	Wait 30 min
4.	Discharge with SDCH (Standard Discharge)

Table 1: Capacity measurement procedure

- **Intermediate test-** During lifetime testing a short performance test is conducted with the cell in regular intervals. Through this aging of the cell is determined.

  - Procedure:

1.	Temperature adjustment to 25°C in climate chamber
2.	Charge with SCH (Standard Charge procedure)
3.	Wait 30 min
4.	Discharge with SDCH (Standard Discharge)
5.	Wait 30 min
6.	Repeat the step 2 to 5 twice in accordance to basic validation
7.	Charge with SCH (Standard Charge procedure)
8.	Wait 30 min

9.	Discharge with 1/3SDCH Discharge with SDCH (Standard Discharge)
10.	Wait 30 min
11.	Discharge with 3C for 30s

Table 2: Intermediate Test

- **Cyclic Aging-** In this test cyclic life time of a Li-ion cell is to be tested. The test ends if EOL criteria or the specified cycle number is reached.

  ➢ Procedure:

1.	Temperature adjustment to 25°C or instructed differently in climate chamber
2.	Charge with SCH (Standard Charge procedure)
3.	Wait 30 min
4.	Discharge with SDCH (Standard Discharge)

Table 3: Cyclic aging

- **Energy and capacity at room temperature-** This test measures device under test capacity in Ah (Ampere hour) at constant current discharge rates corresponding to the supplier rated capacity.

From above all tests the common test procedures carried according to the battery cell data available for accomplishing the aim of the thesis is given in the table below.

Step (cycle)	**Procedure**	**Ambient Temperature**
1.1	Standard Charge (1.5A)	24°C
1.2	Standard Discharge(-2A)	24°C
2.1	Standard Charge (1.5A)	24°C
2.2	Standard Discharge(-2A)	24°C
.	.	.
.	.	.
.	.	.
20.1	Standard Charge (1.5A)	24°C
20.2	Standard Discharge(-2A)	24°C

4 Review of battery modelling:

A common battery datasheet contains the fundamental information such as the battery chemistry, dimensions, capacity, internal impedance, nominal voltage, end of charge voltage and current, end of discharge voltage, charge specifications in terms of maximum current in CC mode, operating temperature range, storage temperature. This information represents the standard tests performed by the manufacturers, that are not representative of actual working conditions of battery [3]. The aim of the basic battery data sheet is to define the safety operating conditions a cell can operate. Additional information like cycle life in actual duty cycle and actual capacity according to the discharging strategy are necessary to further test batteries.

In order to ensure a safe and efficient operation, a proper battery model is essential in predicting the battery behavior under various operating conditions to avoid improper operations. The battery behavior under various operating conditions helps design on-board control and maintenance. Another important application of a battery model is to estimate battery states, such as state of charge and state of health which are not yet directly measurable [4].

Over the years, researchers have developed different battery models of different level of accuracy and complexity [5]. The battery models are mainly divided in to three groups:

1) White box models (e.g. electro chemical model)
2) Grey-box models (e.g. equivalent electric circuit model)
3) Black-box models (e.g. Neural network model/ Machine learning model)

4.1 Electrochemical models:

This model has the advantage of representing the electrochemical dynamics of a battery [6]. This type of battery model is represented by a set of couple partial differential equations (PDEs). These equations tell how the cell's potential is produced and effected by electrochemical reactions that took place inside the cell. This model is more accurate compared to other battery models as they explain the key behaviors of battery at microscopic scale based on chemical reactions occurring inside the battery. These models are complex and time consuming because they involve a system of coupled time-variant spatial partial differential equations. This type of solution requires days of simulation time, complex numerical algorithms, and battery-specific information that is difficult to obtain and develop a physical model [7].

4.2 Electric circuit-based models:

This model is useful and easy to implement specifically for an electrical designer [8]. In comparison with the electrochemical models, circuit-based models don't deal with the complicated electrochemical interactions at the cell level but simply the battery performances at the system level. This model simulates the behavior of the battery in different applications using voltage source, resistors and capacitors in different combination such as Thevenin-based model, impedance-based model and runtime- based models. One of the widely used model is Thevenin's equivalent electrical circuit model (RC Model), in this a single resistor serves to represent the internal resistance of the battery and a combination of parallel RC circuit to represent the transient behavior of the cell. Here the number of parallel RC combination can be decided by the designer up on validating the accuracy of the model. This could also change for every cell type and for different test strategies.

Below Figure 6 is an example of Thevenin based electric model:

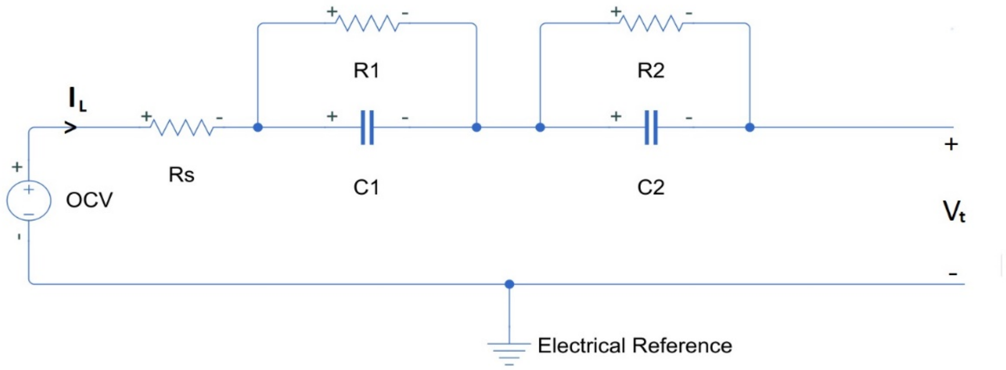


Figure 9: Equivalent circuit model of a Battery cell

OCV - Open circuit voltage of Cell

R_s - Represents the internal resistance of Cell

RC - Parallel RC networks represent the transient behavior of the Cell

I_L - Load Current

The terminal voltage of third order RC-Parallel circuit is given as follows.

$$V_t = OCV - V_{RC1} - V_{RC2} - I_L R_s$$

In order to develop an electric circuit-based models the circuit elements are to be initialized with a value at different states of battery charging & discharging process. For this a Hybrid Pulse power Characterization (HPPC) [9] test is conducted with a constant current pulse of a defined duration for different or same current rates for discharging or charging a battery at different ambient temperatures. In order to accomplish the parameterization of battery model mostly discharge pulse of a constant current input is chosen and the corresponding voltage out is considered to calculate the passive elements resistors, capacitor, open circuit voltage at a relaxed pulsed duration time. Later the pulses are considered at different states of discharge duration starting from 0 to 100% discharge till the cut-off voltage is reached [10].

For a given HPPC pulse input current, the single pulse output voltage response for a battery cell is given in Figure 7 . From each of these pulse outputs at different state of discharge corresponding parameters are calculated.

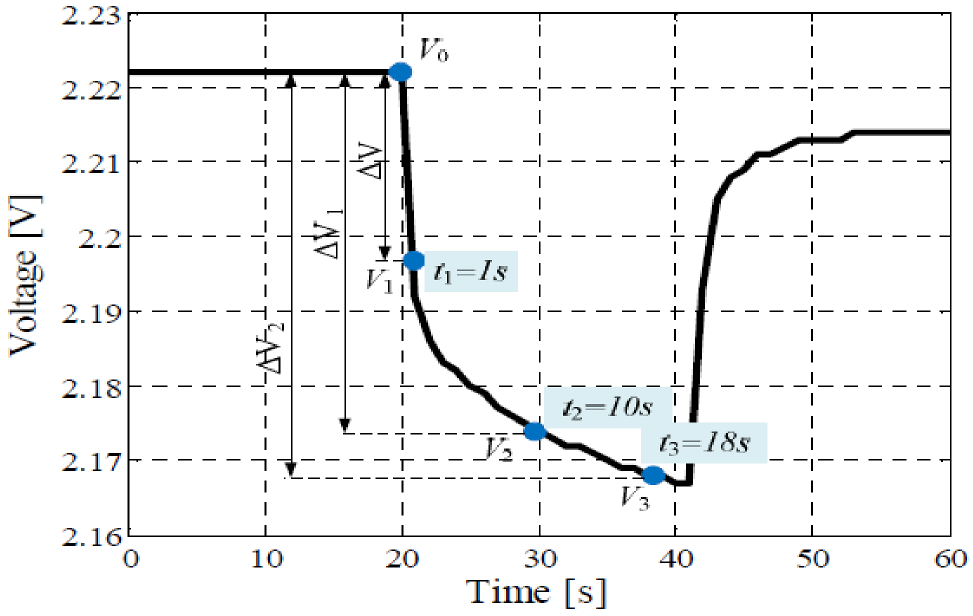


Figure 10: Single pulse voltage output of HPPC Test

$$R_s = \frac{\Delta V}{\Delta I} = \frac{V_0 - V_1}{I} \quad , \quad R_1 = \frac{\Delta V}{\Delta I} = \frac{V_1 - V_2}{I} \quad , \quad R_2 = \frac{\Delta V}{\Delta I} = \frac{V_2 - V_3}{I}$$

$$\tau_1 = \frac{t_2 - t_1}{\ln \frac{V_2(t_2)}{V_2(t_1)}} \quad , \quad \tau_2 = \frac{t_3 - t_2}{\ln \frac{V_2(t_3)}{V_2(t_2)}}$$

$$C_1 = \frac{\tau_1}{R_1} \quad , \quad C_1 = \frac{\tau_1}{R_1}$$

By calculating these parameters at each state of discharge and tabulating all the parameters using a look-up tables in MATLAB/Simulink develops a battery electrical circuit model which can be further used to estimate the states of battery at different conditions. The estimated output curves can be validated with the experiment results in laboratory and the new parameters values can be updated by using curve fitting algorithms or using machine learning algorithms.

By parameterization of RC model and thus determining the time duration to reach 100% SOC or 100% SOD will be able to estimate the time for other different input conditions and finding the accuracy by validating each time and updating the new model parameters by using optimization algorithms. But most of the battery test procedures are started with standard charge and discharge tests in order to estimate time for standard input current and voltage tests an extra pulse input test is needed to be carried for each battery type which consumes one extra cycle test of battery charge and discharge process. This type of test is used to determine the power and internal ohmic resistance for charge and discharge tests as well as the open circuit voltage as a function of state of charge and only applicable to limited operating conditions when parameters have identified [11].

As the main objective of the thesis is to predict only the execution time of the battery charge and discharge process at different operating conditions for different battery types at different test strategies, irrespective of finding the behavior of battery by parameterizing at different states of charging and discharging process, a data driven model with help of machine learning algorithms can accomplish the task to estimate the execution time.

4.3 Data-driven approach model:

Data-driven techniques generally learn from historical data of the system and then wisely suggest a decision through results. One of the advantages of this model is they can learn the behavior of the battery based on monitored data and thus do not demand battery chemical modeling and knowledge. The main part of the data-driven modelling is the unknown mapping between system's inputs and its outputs from the available data [12]. There are number of areas contributing to data driven modelling: data mining, knowledge discovery in databases, computational intelligence, machine learning, Intelligent data analysis, soft computing and pattern recognition.

Machine learning (ML) [13] is an area of computer science that was for a long time considered as a sub-area of artificial intelligence (AI). Machine learning concentrates on the theoretical foundations of learning a data [14]. The process of building a machine learning model is given below.

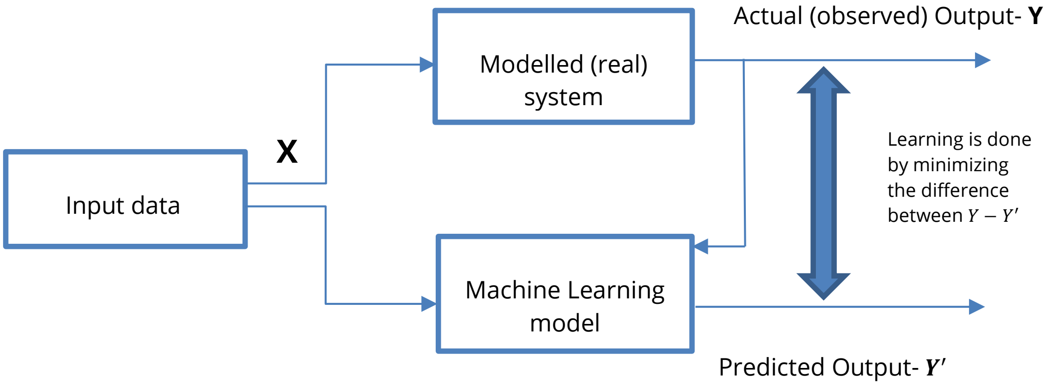

Actual (observed) Output- **Y**
Modelled (real) system
X
Input data
Learning is done by minimizing the difference between $Y - Y'$
Machine Learning model
Predicted Output- Y'

Figure 11 :A typical Machine Learning model

4.4 Regression model:

Regression analysis is a statistical methodology that is most often used for numeric prediction. Whereas classification models predict the categorial (discrete, unordered) labels, regression models predict the missing or unavailable numerical data value. In a machine learning model prediction, it’s important to understand the prediction errors. There are two type of prediction errors bias and variance.

Bias Error: Bias are the simplifying assumptions made by a model to make the target function easier to learn. Generally, linear algorithms have a high bias making them fast to learn and easier to understand but less flexible. In turn, they have low predictive performance on complex problems.

- Low Bias: Suggests less assumptions about the form of the target function.

- High-Bias: Suggests more assumptions about the form of the target function.

Variance Error: Variance is the amount that the estimate of the target function will change if different training data is used. Ideally, it should not change too much from one training dataset to the next, meaning that algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables.

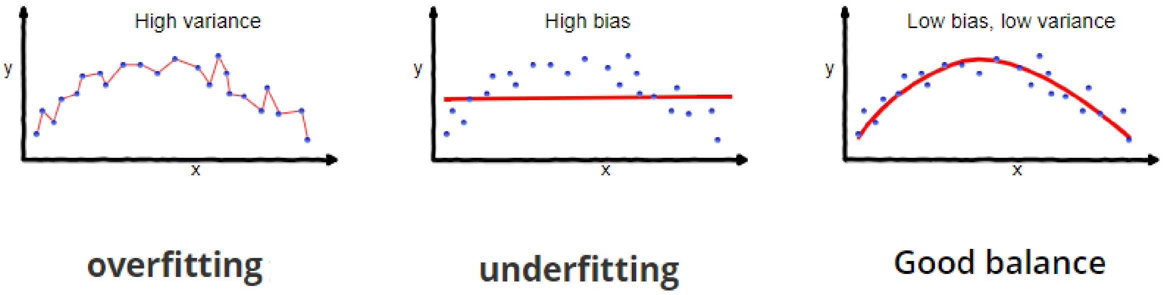
- Low Variance: Suggests small changes to estimate of the target function with changes to the training dataset.
- High Variance: Suggests large changes to estimate of the target function with changes to the training dataset.

Bias- Variance Trade-off: The goal of any machine learning algorithm is to achieve low bias and low variance and algorithm should achieve good prediction performance.

The relationship between bias and variance is

- Increasing the bias will decrease the variance (Underfitting of a model).
- Increasing the variance will decrease the bias (Overfitting of a model).

The parameterization of the machine learning algorithms is often a solution to balance out the bias and variance trade-off.



overfitting

underfitting

Good balance

Figure 12: Bias-Variance Trade-off

4.5 Data Utilized:

For this master thesis the Test bed from the “SYSTECS Informationssysteme GmbH” is not yet ready to conduct the battery cell experiments, measurements from NASA organization are utilized as a third-party data for battery modelling. The battery dataset from NASA [15] comprised of a set of Lithium ion battery cells with charging and discharging profiles at room temperature. This dataset of battery is applicable for the performance test analysis of battery. Below are the details about the battery charge and discharge process.

Battery type: Lithium-ion

Battery type name: ’B0005’

Rated Capacity: **2Ah**

Charging:

Charge method- **CC-CV**

Chamber temperature-**24°C**

Charging current-**1.5A**

Constant voltage- **4.2V**

Charge cutoff current-**20mA**

Discharging:

Discharge method-**CC**

Chamber temperature-**24°C**

Current load-**2A**

Discharge Cutoff Voltage-**2.7V**

With the above characteristics charging & discharging cycles are performed and the below parameters are noted for 20 cycles.

Charging:

Voltage measured: Battery terminal voltage(volts)

Current measured: Battery output Current (Amps)

Temperature measured: Battery temperature (degrees)

Time: Time vector for the cycle (secs)

Discharging:

Voltage measured: Battery terminal voltage(volts)

Current measured: Battery output Current (Amps)

Temperature measured: Battery temperature (degrees)

Time: Time vector for the cycle (secs)

4.6 Charging method:

From the documentation of battery testing procedures, the battery cell is being charged using constant current- constant voltage method (CC-CV) [16]. This method consists of three phases, the first phase, trickle-charge phase used to test whether the battery is functioning properly or if it is damaged. Generally, a very small of input current is applied to avoid excessive heating if the battery is damaged. In the second phase, constant-current phase charging current is increased to its full level and the battery voltage is observed and continued until it reaches to rated maximum level then the third phase, constant-voltage in which a constant voltage equal to rated maximum voltage of the battery cell is applied across the battery and the battery current is observed. In this phase, the cell determines how much current it can absorb to continue the charging process, as the current drops to the charge cutoff current the cell is considered fully charged and CV phase is stopped. This combination of CC-CV is a fast charging method compared to constant voltage, constant current methods [17]. This method has a self-regulating current in the constant voltage phase and does not cause the battery to overcharge.

Constant current-Constant voltage (CC-CV) Charge:

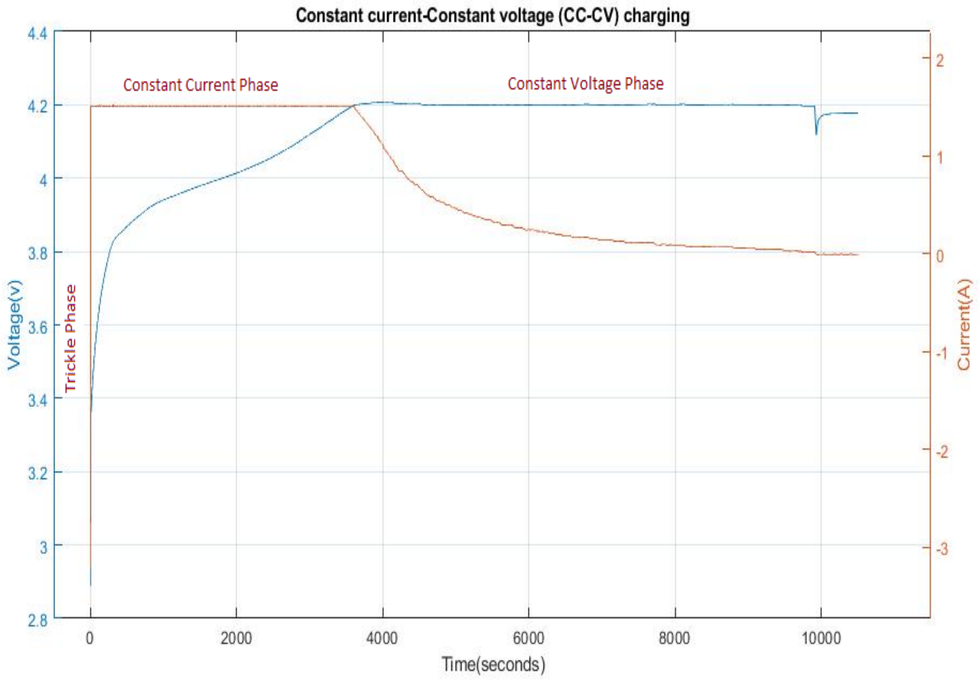


Figure 13: CC-CV charging curve of a Battery cell

In *Figure 10* details the typical charging of a CC-CV mode in which an input current of 1.5 amperes is applied to the battery till battery terminal voltage reaches to rated maximum voltage (here 4.2V) then a constant voltage input is applied till the battery current reaches to charge cutoff current (20mA). Once the charge is terminated, the battery voltage drops due to self-discharge, some chargers apply a brief topping charge to compensate for the small self. From the start of input charge current to end of battery current reaching its cut-off current is considered as the execution time for the Battery charging.

4.7 Discharging method:

For discharging a battery constant current method is used, where a negative input current is applied, and the battery cell voltage is observed till it reaches to the Discharge cutoff voltage

Discharging parameters vs Time:

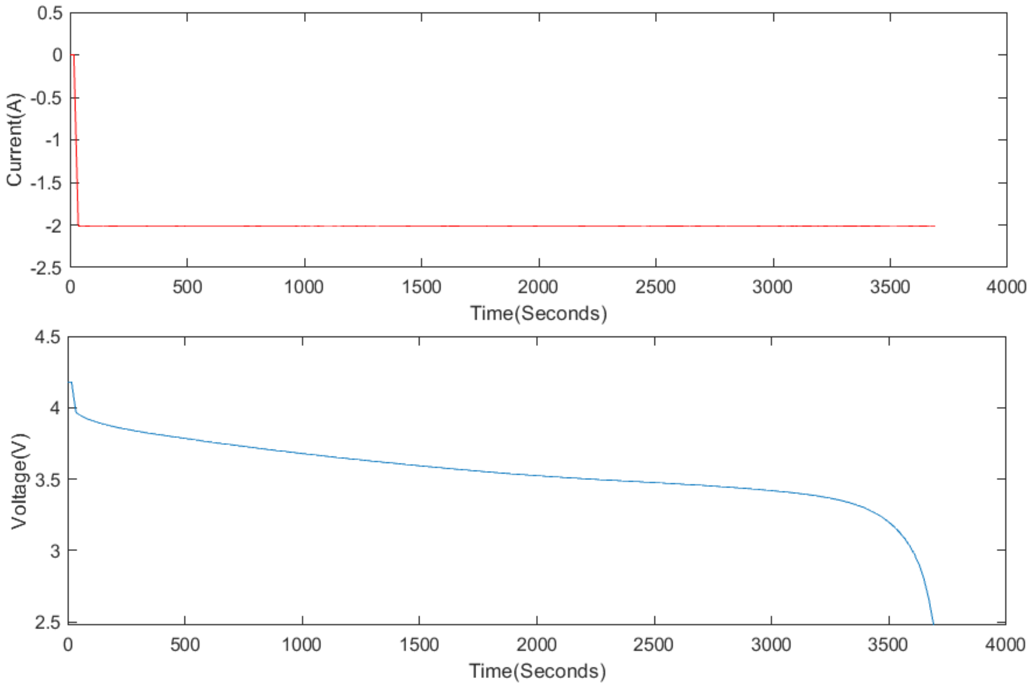


Figure 14: Initial discharge characteristics of Lithium-ion battery

In the above figure the discharging characteristics of the NASA dataset with a constant load current of -2 A is applied till the battery terminal voltage reaches the cut-off voltage of 2.7V.

5 Machine learning Battery Model Development:

General machine learning model development process [18] involves the following steps:

1. Data pre-processing
2. Feature engineering
3. Feature extraction
4. Feature selection
5. Algorithm selection
6. Hyperparameter optimization

5.1 Data pre-processing:

Before a data-driven battery model to be build the input data to be trained for the machine learning model needs to be relevant, cleaned, normalize and free from outliers. As in the condition of battery data the readings are taken from a variety of sensors which lead to measurement errors. In order to train the model data must be preprocessed i.e. missing values must be replaced by zero, mean, mode or any other value of each input feature of the model according to its requirement. Later the data is to be normalized if necessary and reducing the data in order to achieve a less dimensionality dataset so the machine learning model trains faster.

Before removing outliers, according to the battery charging and discharging characteristics. From the *Figure 9* it is evident that the current charge starts from negative input current and increased to the required input charging current and after the charge terminates the current input is again maintained to 0A and kept at standby mode. The last stage of 0A input current given us the open circuit voltage information of the battery after every charging process termination. As the main goal is to train the machine learning model with the duration time period between the starting time of the charging and end time of the charging, the other data set irrelevant for the charging process must be eliminated. So, from the battery dataset, the sample size from current input 1.5A to 20 mA is considered and remaining data is to be eliminated. This step is also applied to battery discharge dataset where the sample size starting from load current -2A to battery discharge cutoff voltage is considered.

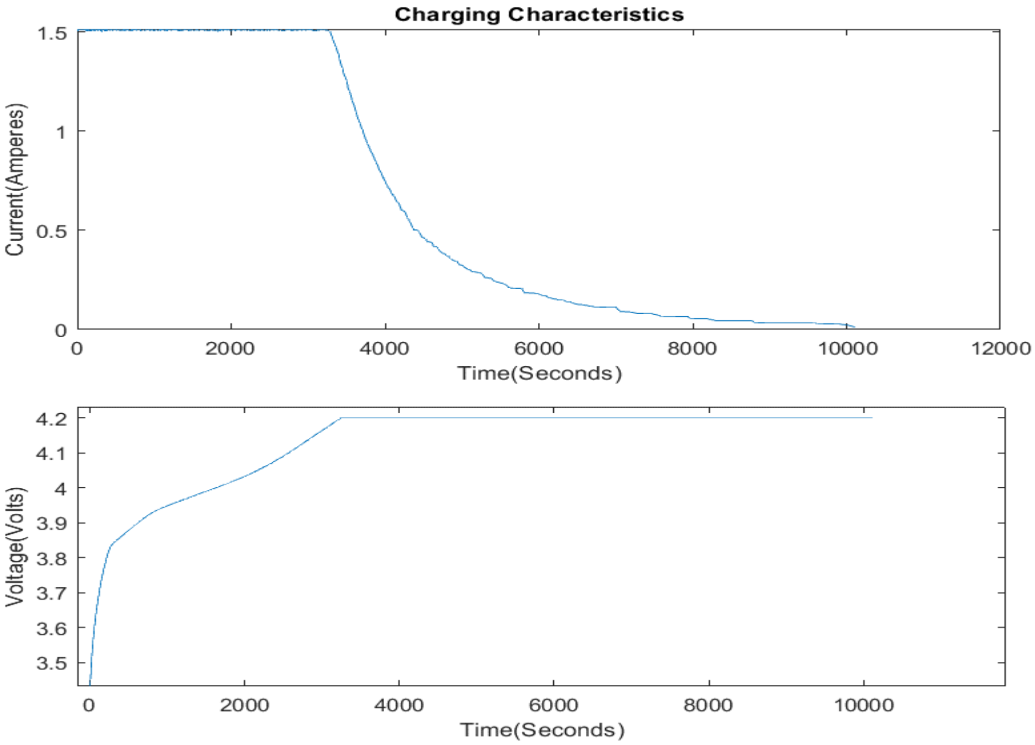


Figure 15:Charge characteristics after data cleaning

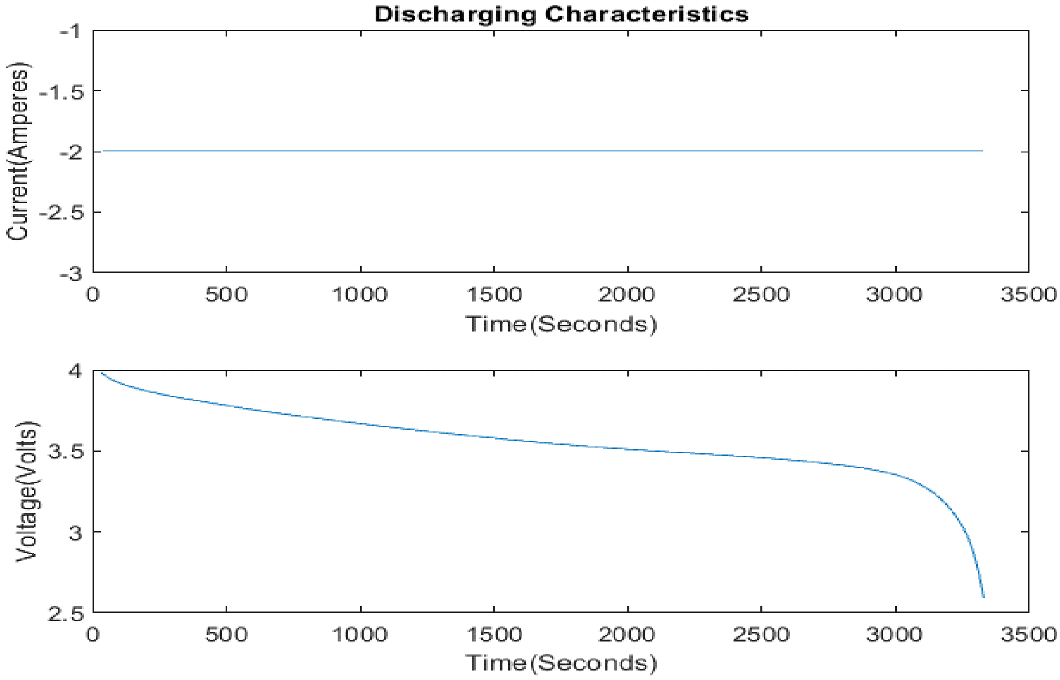


Figure 16:Discharge characteristics after data cleaning

5.1.1 Data binning:

In the charging process the battery output voltage after reaching constant voltage phase should be maintained at 4.2V but due to the measurement errors the battery output voltage has a non-linearity readings as shown in Figure 13.In order to have a best fit for the estimated variable. This phase of data needs to be grouped to 4.2V or maximum rated voltage as specified by the supplier. For this data binning is is applied to the constant voltage phase where the battery voltage after reaching the maximum charging voltage is grouped to only a constant voltage value.

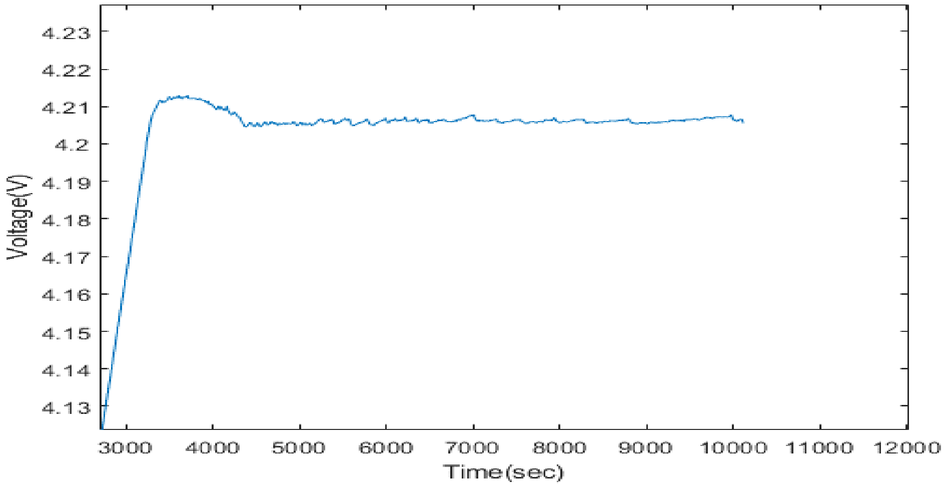


Figure 17:Constant voltage phase at charging before binning

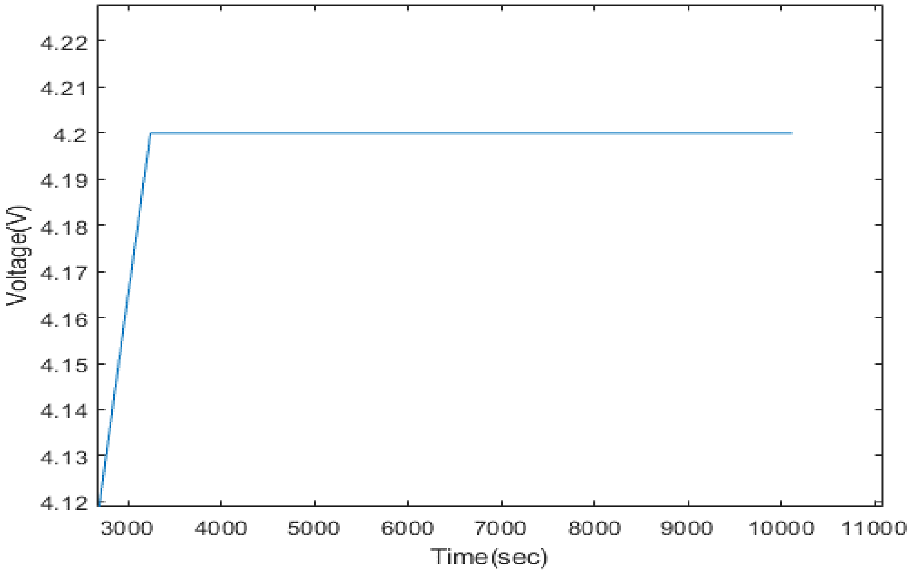


Figure 18: Constant voltage phase after removing the non-linearity

5.1.2 Feature selection:

In order to select the relevant features from the test bed that are required for the run time estimation of the battery, feature selection needs to be processed. As per the given documentation the industrial parameters measured and stored from a Battery test bed are 1.) Cell Current (A) 2.) Cell Voltage (V) 3.) Cell Temperature (°C) 4.) Climate Chamber Temperature (°C). Industrial parameters are the signals read out from the corresponding sensors, it is also important to note the cycle number of the battery cell to determine the current and further states of battery. In order to find out whether the features have high or low influence on the output variable. For this Pearson correlation coefficient method is utilized for each feature with respect to the time.

The Pearson correlation coefficient [19] is used to measure the strength of a linear association between two variables (x, y), where if the coefficient value r = 1 means a perfect positive correlation and the value r=-1 means a perfect negative correlation. The formula for calculating is given below.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}$$

From the data set available calculated the correlation coefficient matrix for the charging and discharging data set of single battery type at one cycle and represented it in a heat map as shown in *Figure 15, Figure 16*

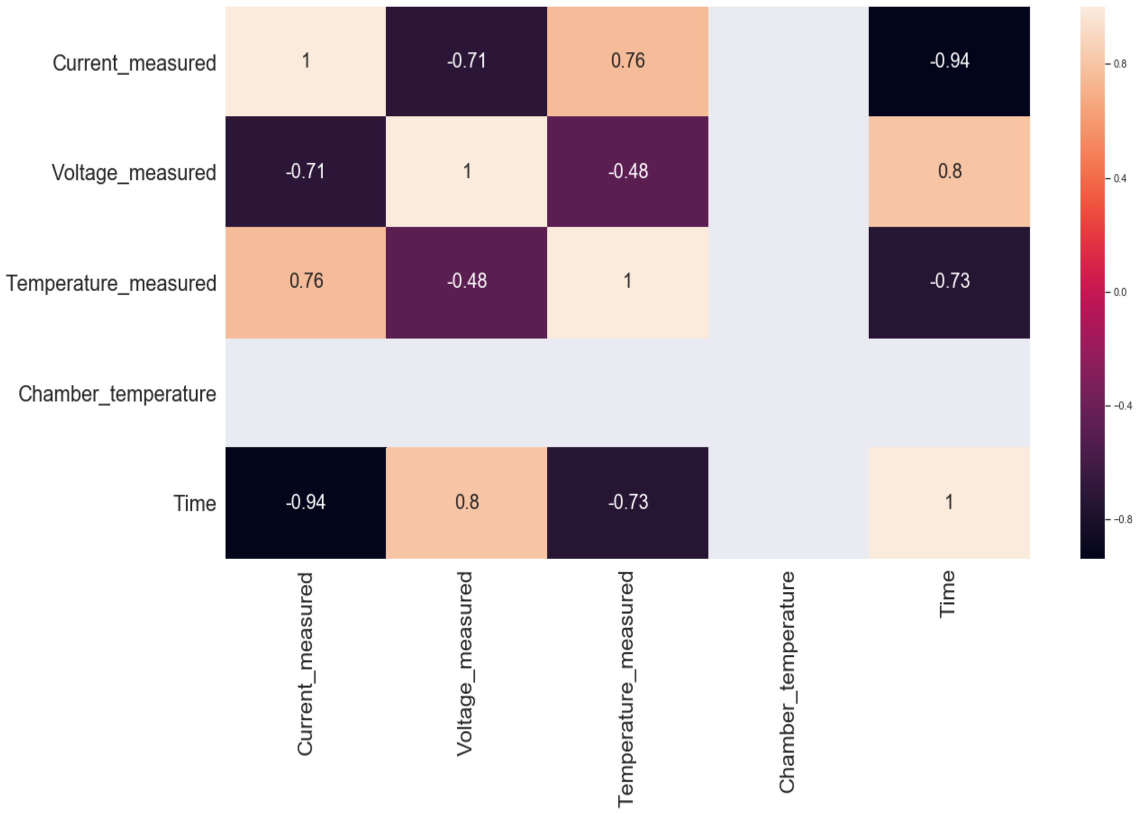


Figure 19: Heatmap for Correlation matrix of Charging data

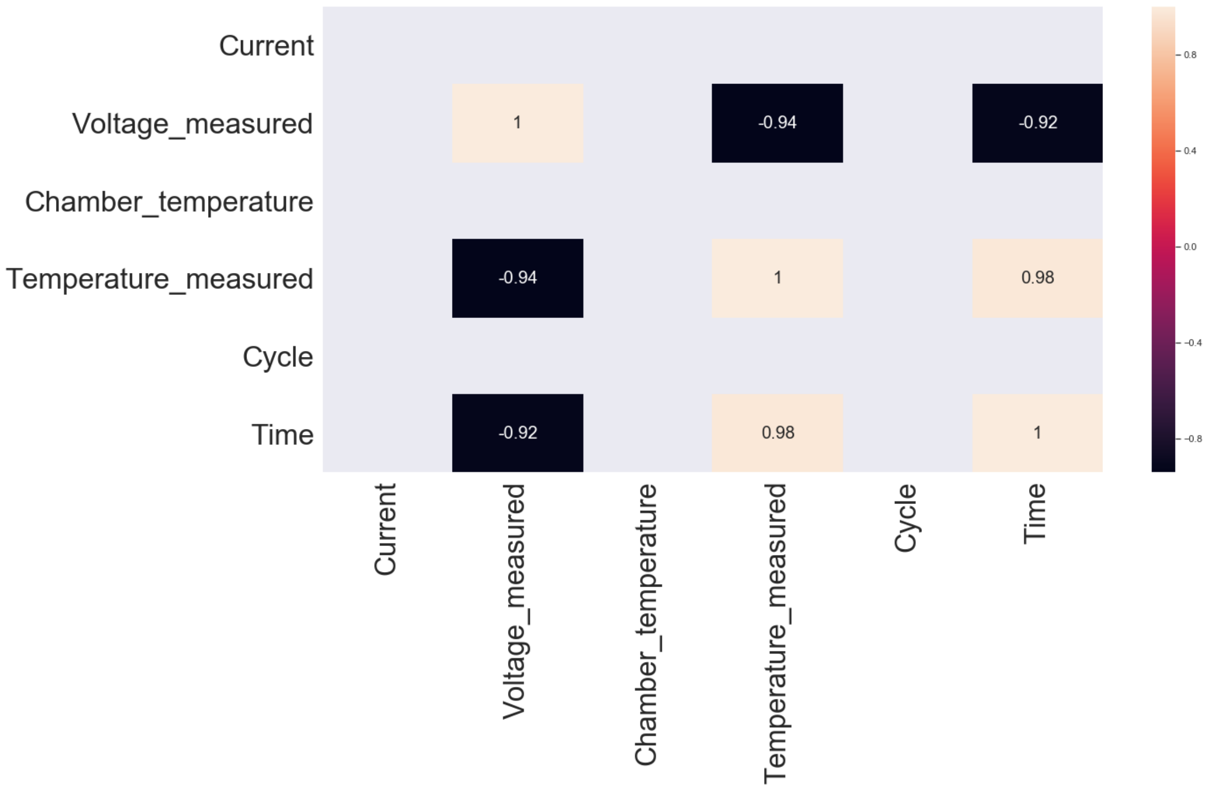


Figure 20:Heatmap for correlation matrix of Discharging data

From the above correlation coefficient matrix the time variable correlated with respect to time is given in the last column and here the ambient temperature and cycle number are constant for a single cycle during the entire charging and discharging process so these two variable have no correlation with respect to time at a single cycle of charge and discharge process because the values are constant during the entire single test. For discharging the current load is constant and is not varied from the start of discharge time so current load has no high correlation with discharge time variation during the process.

From the two-correlation matrix the required highly correlated feature vector to predict time for charging and discharging process is given by

$$Time_{Charge} = [Current\ measured, Battery\ voltage, Battery\ temperature]$$

$$Time_{Discharge} = [\ Battery\ voltage, Battery\ temperature]$$

In order to develop a battery machine learning model for both charge and discharge run time we can combine the feature vector as given below.

$$Time_{Battery} = [Current\ measured, Battery\ voltage, Battery\ temperature]$$

As the battery data set available for the analysis has only a constant ambient temperature. In order to have a test plan for charging and discharging at different chamber temperature the chamber temperature feature must be included in the data set to be trained. For most of the test strategies like cyclic aging, energy and capacity test runs for several cycles which is useful to determine the capacity fade in the battery after a certain battery charge and discharge cycles, the cycles number is also an important feature to estimate the time at different. The feature vector after including these feature variables is given as

$$Time_{Battery} = \begin{bmatrix} Current\ measured, Battery\ voltage, Battery\ temperature, \\ Chamber\ temperature, Cycle\ number \end{bmatrix}$$

From the correlation matrix of charge and discharging process run time has a high correlation with the battery temperature, but the battery cell temperature dependent on the variables Chamber temperature, current charge/discharge, Aging effects [20]. This parameter is used as for operational condition monitoring of cell temperature in order not to exceed manufacturer mentioned cell temperature in the data sheet. Battery cell temperature is an unpredictable during the process of charging and discharging, adding including this feature may lead a machine learning model to bias or variance effect. As we already have the features like current charge/discharge, chamber temperature that are required to monitor the end of charging and discharging of a battery cell, we can eliminate the battery temperature feature from the test bed in order to estimate the run time of battery process. As we are developing a model for the both charge and discharge process, included one more variable 'Charging current' which classifies the charge and discharge data set from the training dataset. As there are many battery cells from different manufacturer's tested in a Test Factory, Therefore it is a possible of two battery cells of same charging parameters being tested at once, in order to predict the execution time for individual battery cell it is mandatory to classify the battery cell testing data while training the model, for this included a categorial variable "Battery type". From the all above conclusions the final battery feature vector set for runtime prediction is given as below.

$$Time_{Battery} = \begin{bmatrix} \text{Charging current} \\ \text{Current measured} \\ Battery\ voltage \\ Chamber\ temperature \\ Cycle\ number \\ Battery\ type \end{bmatrix}$$

A typical example of feature vector for charging time prediction for two cells are:

$Time_{UuT1}$= {1.5A, 20mA, 4.2V, 25°C, 2, Panasonic NCR 18650PF}

$Time_{UuT2}$= {1.5A, 20mA, 4.2V, 25°C, 2, LG Chem ICR 18650S3}

5.1.3 Feature Engineering:

Feature engineering in the data preprocessing is the process of using domain knowledge of the data clean the raw data, it increases the predictive power of machine learning algorithm [21]. One of the important features for prediction of execution time is the battery cell current. From the Figure 17 it is evident that from the exponentially decreasing phase contains noises also called as outliers which deviates from the other observations. If the outliers not filtered or removed may lead to machine learning model to learn these outliers/noises and have poor prediction for output variable. As explained in the section charging method the battery current is decreased in constant voltage phase to the charging cut-off current given by the manufacturer. The current readings which are not in a decreasing function are unwanted disturbance in the signal. The first step is to detect the unwanted disturbance data points, once the outliers are detected and decide to filter them. For the available data found out that lowness filters is efficient in smoothing the data but the current range used here is from 1.5 to 20 mA but if this range is variable and different type of noises included and main difficulty in choosing the window size in the filter algorithm although there are some heuristic methods to estimate the window size for a given data, with the present available data cannot conclude to a particular filtering method for current signal noise removal and also smoothing algorithms deviates the original current cut-off point instant where charging should be terminated. In order to find a universal algorithm to find out the solution for the machine learning model to train with the noise free data removed the outliers from the given data and trained the machine learning algorithm to estimate the charging time of battery cell. This algorithm shows a best predictor to estimate the charging time and removing the outlier data time instant from the dataset Figure 18 has an advantage of dimensionality reduction leads to training the model faster .

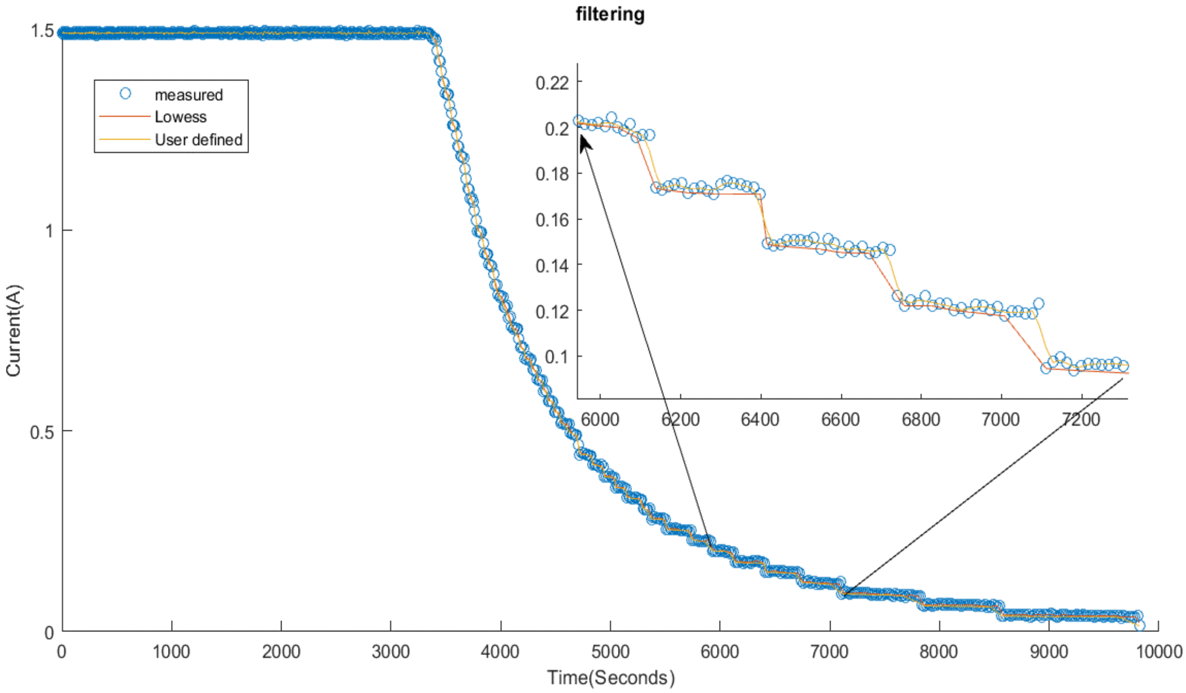
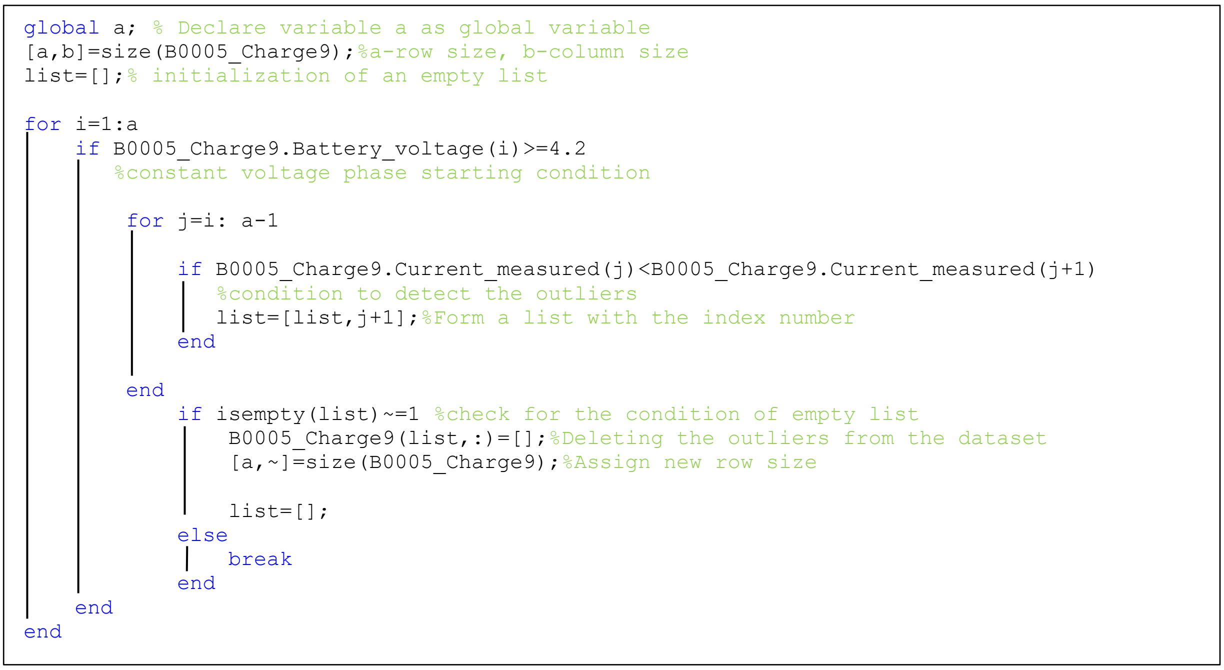


Figure 21: Noises in the battery cell current

The algorithm to detect the outliers in battery current measured and eliminating them is given as below



```
global a; % Declare variable a as global variable
[a,b]=size(B0005_Charge9);%a-row size, b-column size
list=[];% initialization of an empty list

for i=1:a
    if B0005_Charge9.Battery_voltage(i)>=4.2
       %constant voltage phase starting condition

        for j=i: a-1

            if B0005_Charge9.Current_measured(j)<B0005_Charge9.Current_measured(j+1)
               %condition to detect the outliers
               list=[list,j+1];%Form a list with the index number
            end

        end
            if isempty(list)~=1 %check for the condition of empty list
                B0005_Charge9(list,:)=[];%Deleting the outliers from the dataset
                [a,~]=size(B0005_Charge9);%Assign new row size

                list=[];
            else
                break
            end
    end
end
```

Figure 22 :User defined algorithm to eliminate outliers

5.2 Algorithm selection:

5.2.1 Support Vector Regression:

Support vector machine (SVM) has first introduced by Vapnik. There are two main categories for support vector machines: support vector classification (SVC) and support vector regression (SVR). A version of an SVM for regression has been proposed in 1997 by Vapnik, Steven Golowich, and Alex Smola which is called support vector regression (SVR). The model produced by support vector classification only depends on a subset of the training data, because the cost function for building the model does not care about the training points that lie beyond the margin. Analogously the model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close (with in the threshold ε) to the model prediction.

In SVR the main goal is to find a function $f(x)$ that deviates from y_n by a value no greater than ε for each training point x_i and at the same time is as flat as possible.

If we have a set of training data where x_n is a multivariate set of N observations with observed response values y_n. The linear function is given as:

$$f(x) = x\omega + b$$

Here flatness means small ω value, it is required to minimize the Euclidean norm i.e. $||\omega||^2$. Formally this can be written as a convex optimization problem by requiring

$$J(\beta)=\frac{1}{2}\beta'\beta)$$

$$\text{Subject to } \begin{cases} y_i - \omega x_i - b \leq \varepsilon \\ \omega x_i + b - y_i \leq \varepsilon \end{cases}$$

It is also possible that no such function $f(x)$ exists to satisfy these constraints for all points. For this slack variable (ξ_n, ξ_n^*) are introduced for each point. This is like the "soft margin" in SVM classification because the slack variables allow regression errors to exist up to the value of ξ_n *and* ξ_n^*. Yet still satisfy the required conditions. By including these variables, the objective function is given as

$$J(\beta)=\frac{1}{2}\beta'\beta + C\sum_{n=1}^{N}(\xi_n + \xi_n^*),$$

$$\text{Subject to}\begin{cases} y_i - wx_i - b \leq \varepsilon + \xi_i \\ \omega x_i + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

The constant C > 0 determines the tradeoff between the flatness of f and the amount up to which deviations larger than ε are tolerated. The linear ε- intensive loss function ignores errors that are within ε distance of the observed value by treating them as equal to zero. The loss is measured based on the distance between observed value y and the ε boundary. This is given by.

$$L_\varepsilon = \begin{cases} 0 & if\ |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon & otherwise \end{cases}$$

The optimization problem is computationally simpler to solve in Lagrange dual formulation. This can be achieved by constructing a Lagrangian function from the primal function by introducing nonnegative multipliers α_n and α_n^* for each observation x_n. This leads to dual formula, where we minimize

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(a_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i x_j + \varepsilon\sum_{i=1}^{N}(\alpha_i + a_i^*) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i)$$

$$\text{Subject to constraints}\begin{cases} \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)x_n \\ 0 \leq \alpha_n \leq C \\ 0 \leq \alpha_n^* \leq C \end{cases}$$

The β parameter can be completely described as a linear combination of the training observations using the equation.

$$\beta = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*) x_n$$

The function used to predict new values depends only on the support vectors.

$$f(x) = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*)(x_i x) + b$$

The Karush-kuhn-Tucker (KKT) complementarily conditions are optimization constraints required to obtain optimal solutions. For linear SVM regression. The conditions are

$$\begin{cases} \alpha_n(\varepsilon + \xi_n - y_n + x_n\beta + b) = 0 \\ \alpha_n^*(\varepsilon + \xi_n^* + y_n - x_n\beta - b) = 0 \\ \xi_n(c - \alpha_n) = 0 \\ \xi_n^*(c - \alpha_n^*) = 0 \end{cases}$$

These conditions indicate that all observations strictly inside the epsilon tube have Langrange multipliers $\alpha_n = 0$ and $\alpha_n^* = 0$. If either α_n or α_n^* is not zero, then the corresponding observation is called a support vector.

The property Alpha of a trained SVM model stores the difference between Langrange multipliers of support vectors, $\alpha_n - \alpha_n^*$. The properties Support Vectors and bias store x_n and b.

SV algorithm can be made nonlinear by simply preprocessing the training patterns x_i, by a map ᶲ$\phi: X \rightarrow 3$, in to some feature space "**3**" and then applying the standard SV regression algorithm. The dual formula for nonlinear SVM regression replaces the inner product of the predictors $(x_n x)$ with the corresponding to the kernel function $\mathrm{K}(x_i, x)$. the kernel function is defined as a linear dot product of the nonlinear mapping, i.e.,

$$\mathrm{K}(x_i, x) = \varphi(x_i)\varphi(x)$$

Nonlinear SVM regression finds the coefficients that minimize

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)k\big(x_i, x_j\big) + \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*)$$

$$\text{Subject to}\begin{cases} \sum_{n=1}^{N}(\alpha_n - \alpha_n^*) = 0 \\ 0 \leq \alpha_n \leq c \\ 0 \leq \alpha_n^* \leq c \end{cases}$$

The function used to predict new values is equal to

$$f(x) = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)k(x_i, x) + b$$

The KKT complementarity conditions are

$$\begin{cases} \alpha_n\big(\varepsilon + \xi_n - y_n + f(x)\big) = 0 \\ \alpha_n^*\big(\varepsilon + \xi_n^* + y_n - f(x)\big) = 0 \\ \xi_n(c - \alpha_n) = 0 \\ \xi_n^*(c - \alpha_n^*) = 0 \end{cases}$$

5.2.2 Decision tree:

Decision tree is a top to down tree like structure where learning is accomplished by splitting the training data set in to subsets from root node(beginning) down to leaf node. The leaf node gives the numeric responses.

Common terms used in Decision trees:

1. Root node: It represents entire population or sample, and this further gets divided in to two or more homogeneous sets.
2. Splitting: It is a process of dividing a node in to two or more sub-nodes.
3. Decision node: When sub-node splits in to further sub-nodes, then it is called a decision node.
4. Leaf/Terminal node: Nodes do not split is called leaf or Terminal node.
5. Pruning: When we remove sub-nodes of a decision node, this process is called pruning.
6. Branch/ sub-Tree: A sub section of entire tree is called branch or sub-tree.
7. Parent and child node: A node, which is divided in to sub-nodes are called parent node of sub-nodes whereas sub-nodes are the child of parent node.

5.2.3 Decision Tree Algorithm:

There are many specific decision tree algorithms like ID3(Iterative Dichotomiser 3), CART etc., but as we are using battery training data set which comprises of both categorical (strings) variables, continuous (numeric) variables, CART (Classification and Regression Tree) algorithm is a best practice to use for a training dataset containing both categorical and continuous variables.

5.2.4 Classification and Regression Tree (CART):

CART grows an overly large tree using forward selection. At each step, finds the best spilt to build a tree. The fundamental idea is to select each split of a subset so that the data in each of the descendant subsets are “purer” than the data in the parent subset. Suppose we have ‘m’ number of features and n observations each feature can be a numeric variable or an ordered factor (categorical variable) the best split is the one with greater decrease in impurity. For a categorical feature (Classification) input impurity is selected by various algorithms like Deviance, Gini index, Information gain. For a continuous input feature (Regression) residual sum of square is widely used in all regression problems.

For building a Regression tree:

In a CART algorithm the leaf node (Terminal node) represents a numeric value if the prediction is a continuous variable. In contrast, for a Classification problem the leaf node represents true or false in their leaves. As in this data set the prediction is continuous variable (Time) the measure of impurity at each split is accomplished by finding a region with least residual sum of square error (RSS) or Sum of square residuals error (SSR) for each input feature.

$$\text{RSS}=\{R_m\}_{m=1}^{M}\frac{1}{N}\sum_{i\in R}(y_i-\bar{y}_m)^2$$

The initial step in building a tree is finding a node that splits the entire dataset homogenously. This is done by finding sum of square residual for each data point corresponding to remaining data set in the training data set and predicting the output. The first data point which gives the least sum of square residual and its corresponding feature is called the root node and the data point is called the cut point s to divide the dataset in to two regions. Unlike SVR, CART algorithm classifies the categorical input and further splits the root node Figure 23.

A cut-off point is selected by following basis

$$min[rss(y_i|x_{ik}<s)+rss(y_i|x_{ik}>s)]$$

y_i-predicted output

x_{ik}-each data point i for each feature k

s-cut-off point

This procedure is continued in choosing decision nodes which are further split in to sub nodes till the minimum leaf node is reached. In decision tree we can control the maximum number of splits and meaning number of leaf node to build an optimized regression tree to predict the output accurately.

Decision tree diagram:

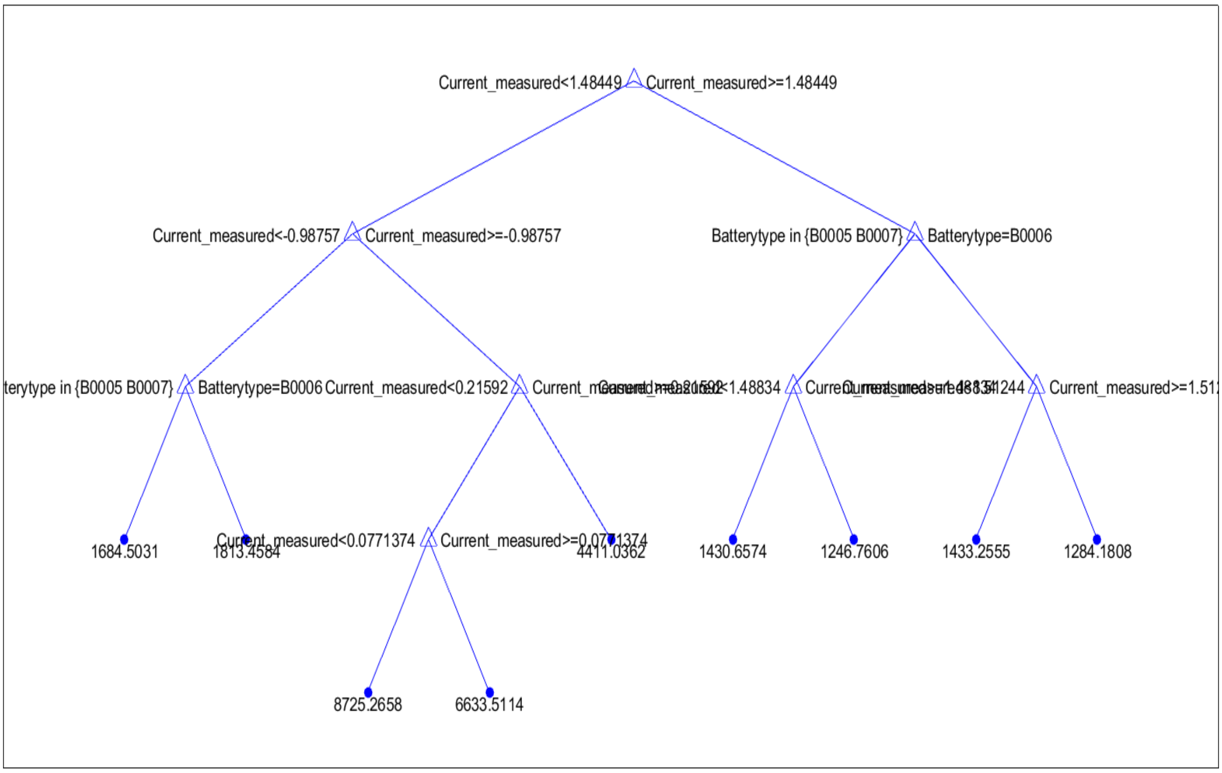


Figure 23: A Typical example of Decision Tree with both continuous and categorical inputs

5.2.5 Ensembles:

The ensemble learning is used to develop an accurate machine learning model with the help of ensembles methods. The main principle behind it is combining prediction from multiple machine learning algorithms together to make more accurate predictions than an individual model, thus increasing the accuracy of the model. The main cause of difference in actual and predicted values are noise, variance and bias. Ensembles help to reduce the factors (expect noise, which is irreducible error). Using techniques like Bagging and Boosting helps to decrease the Bias-variance trade-off problem and increases the robustness of the model. Combinations of multiple regression trees decrease variance, especially in the case of unstable regression model, and may produce a more reliable regression model than a single regression model.

5.2.6 Bagging:

Bagging or Bootstrapped aggregating is one of the first cases of an ensemble of decision trees. It is also the most intuitive, simple method and performs very well. Diversity in Bagging is obtained by using bootstrapped replicas of the original training set: different training data sets are randomly drawn with data replica with the use of the standard approach. Thus, each tree can be defined by a different set of variables, nodes and leaves. Finally, their predictions are combined by from the each formed with randomly sub divided datasets. The steps involved in building a Bagging Decision Tree are:

1. Create random sub-sample of dataset.
2. Train a CART model on each sample.
3. Given a new dataset, calculate the average prediction from each model.

5.2.7 Boosting:

Boosting algorithm utilize weighted averages to make weak learners in to stronger learners. Unlike bagging that had each model run independently and then aggregate the output, boosting algorithms seek to improve the prediction power by training a sequence of weak models, each compensating the weaknesses of its predecessors. One of the widely used boosting algorithm is Gradient boosting.

5.2.8 Gradient boosting decision tree:

The Gradient is the version of boosting algorithm supported in many of programming language machine learning libraries. Below are the steps involved in building a gradient boosting regression tree.

For a given input data be trained $\{(x_i, y_i)\}_{i=1}^{n}$ *(n- number of data points)* and a differentiable Loss function L$\big(y_i, F(x)\big)$. Loss function is to evaluate how well a model can predict the output. The loss function that is mostly used in Regression problem with gradient boost is $\frac{1}{2}(Observed - Predicted)^2$.The steps to build a Gradient Boosting Decision tree are:

Step 1: Initialize the model with a constant value:

$$F_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, \gamma)$$

Step 2: For m= 1 to M number of Trees:

a. Derivative of loss function with respect to predicted value

- $r_{im} = -\left[\frac{\partial L\left(y_i, F(x_i)\right)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$

- $r_{,i,m}$- *r* is the residual, *i* the data point number, *m* is the tree

b. Fit a regression tree to the r_{im} values and create terminal regions R_{jm}, for j=1…j_m – m is the index number of tree and j is the index for each leaf in a tree.

c. For j=1…j_m compute $\gamma_{jm} = \underset{\gamma}{argmin} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$- this step is to determine the output values for each leaf. If two residuals ended up in a leaf (terminal node), it is unclear what its output value should be, so for each leaf in a tree the "γ_{jm}" is calculated thus given a single output value for each leaf in a tree.

d. Update $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm}\, I\big(x \in R_{jm}\big)$

- This is the final predicted value of each sample data point and the second term summation is to add up the output values, $\gamma_{j,m}$'s for all leaves, $R_{j,m}$, that a sample x, can be found. The Greek character 'ν'-nu is the learning rate ranging between 1 and 0. This learning rate can be used to control each tree effect on the final prediction of output this helps to reduce the number of tress required to form and improves the accuracy.

Finally, a Gradient boosting decision tree model is built for a given training data set. Here in the step 2.b it should be noted that a regression tree is built by predicting the residuals rather than the output values and then followed by adding the previously predicted output value for the same sample. This process can be continued till the residual reaches to minimum or zero. This sums up the formation of a boosting algorithm where it builds a decision tree model from sequential weak learners to a strong learner to predict the output accurately.

5.3 Cross Validation & Hyperparameter Tuning:

5.3.1 Hyperparameter Tuning:

Whether it's a SVM, Decision tree, Ensemble model, the parameters like learning rate, number of leaves, maximum number of splits, number of trees to be formed (Ensemble modelling) and kernel function are to be set before training a model, these parameter set is called Hyperparameter. These hyperparameters are to be finalized for a given data set in order to predict the output accurately. Various strategies are used to find the optimal hyperparameters for a given training data set. In this thesis mainly two strategies are used:

1) Grid search
2) Random search

Grid search: Grid search is a traditional way to perform hyperparameter optimization. It works by searching exhaustively through a specified subset of hyperparameters. The benefit of grid search is it finds the accurate optimal combination of parameters, but main drawback is that model is tested with every combination of hyperparameter values within given grid division which is time consuming and computationally expensive.

Random search: Random search differs from grid search mainly in that it searches the specified subset of hyperparameters randomly instead of exhaustively. The major benefit is decrease in the processing time. But it is possible that Random search will not find as accurate of result as Grid search

For this a new method called random grid sweep is mentioned in [22] . This method uses only a subset of all possible combinations, selected randomly in an entire grid. The difference between random grid and the random search mode is that the latter chooses the parameter randomly within the specified range, while the former uses only the exact values defined in the algorithm module. In this thesis utilized the Random grid sweep method to find a best hyperparameters set.

5.3.2 Cross-Validation:

There is always necessity to validate the machine learning model. The performance of model cannot be finalized with training data. In order to know the model performance for an entirely new dataset. For this the training dataset is partitioned in the following way:

k-Fold Cross validation:

In this method the data is divided in to k subsets as shown in Figure 2, such that each time one of the k subset is used as test set and the other k-1 subsets are used to train the model and test data is used for prediction. During the cross validation step the optimal hyperparameters also can be selected where at each iteration of validation the hyperparameter tuning method is implemented and calculated the error estimation is averaged over all k trials and the parameter set which gives less error is considered as best Hyperparameter set for the model.

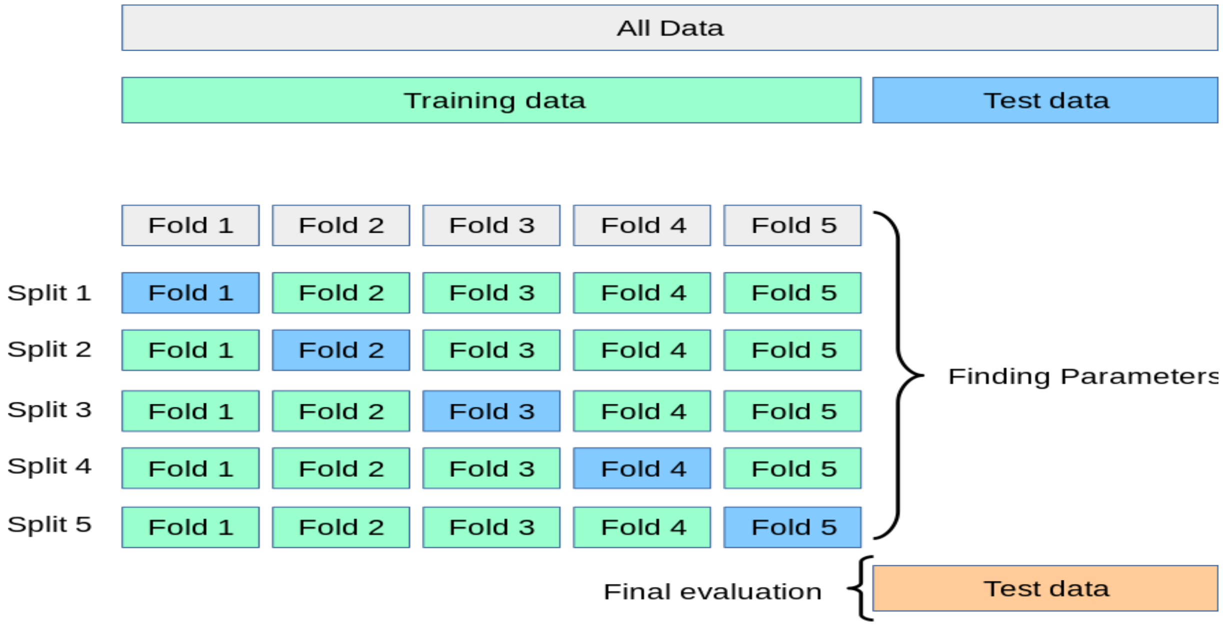


Figure 24: Cross-Validation & Hyperparameter Tuning

6 Results & Discussion:

Performance Metrics:

A machine learning model can be evaluated by calculating the performance metric, and further improvements can be made to achieve a desirable accuracy. The evaluation metrics explain the performance of a model, the metrics utilized are given below:

Coefficient of Determination ($\boldsymbol{R^2}$) :

In regression analysis learning to access a model or to know the proportion of variance in the dependent variable that is predictable from independent variable coefficient of determination is calculated as below

$$R^2 = 1 - \frac{\sum(y - y')^2}{\sum(y - \bar{y})^2}$$

$$= \qquad 1 - \frac{SS_{res}}{SS_{tot}}$$

Here y- measured variable , SS_{res} – Residual sum of squares

$\bar{y}$ - mean of measured variable , SS_{tot} – Total sum of squares

y'- estimated output variable

- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.

Root mean square error (RMSE):

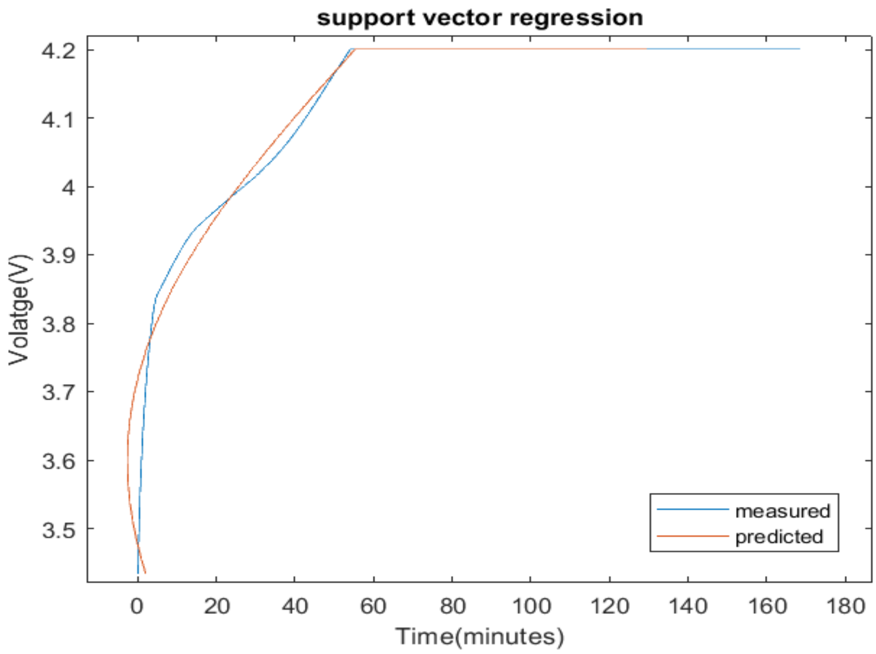
It represents the sample standard deviation of the differences between predicted values and observed values (called residuals).

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - y_j'\right)^2}$$

- A lower value close to 0 means that dependent variable predicted without error from the independent variable.
- A higher value of RMSE means that dependent variable predicted with larger variance from the independent variable.

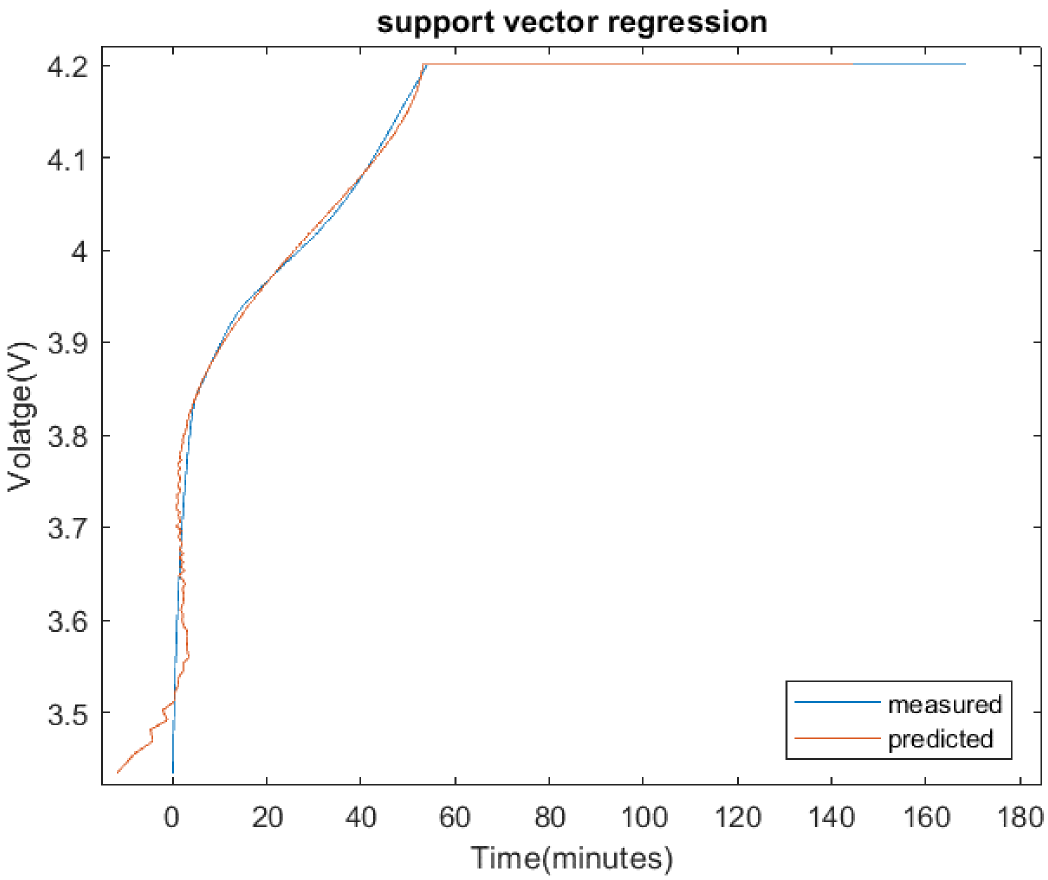
Support vector regression:

Support vector regression algorithm is implemented for a single cycle battery charging data. As per the algorithm “gaussian” kernel function is selected because the output has a nonlinear behavior with the input dataset, linear function cannot be used, even though a polynomial function can be utilized but choosing the degree of the polynomial varies for each time a new data is being added, By these conclusions gaussian kernel function is the best choice for nonlinear mapping of input features to a linear feature space.



- Coefficient of determination R^2 is 0.9750
- RMSE is 31.4327
- Measured charging time:168.5805
- Predicted charging time: 129.4737

From the above results SVR algorithm for estimating charging time is not accurate and in order to increase the accuracy hyperparameter tuning is carried out and the implemented the new best optimized parameters and trained the algorithm again and estimated the charging time and the best parameter set for SVR algorithm is



- Coefficient of determination R^2 :0.9910
- RMSE:26.6486
- Measured charging time:168.5805
- Predicted charging time:144.5439

Even though after implementing the optimized parameters residual of measured and predicted charging time is large. The reasons for estimating the charging time with less accuracy are,

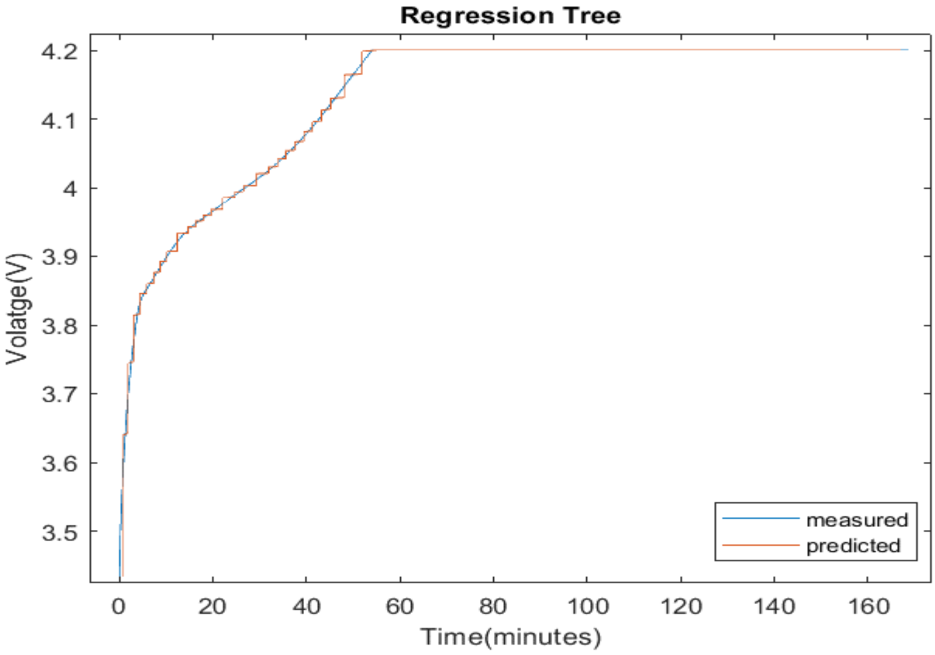
1. The algorithm has support vector datapoints upon which future estimation is totally dependent. While considering only support vector data points and

ignoring other data points could lead to less accuracy in prediction of execution time for a battery charging and discharging process.

2. Increasing the support vector data points also leads to large residual error for a new dataset. The input feature set contains both continuous and categorial variables which is a both regression and classification problem. For this all the categorial variables are encoding to dummy variables which are either integers or binary numbers. After encoding the categorial inputs the model predicts with less residual error but for a new dataset it produces a large residual error. This is due to the reason that after encoding the categorical feature, these variables are treated as regression problem instead a classification problem where a new battery type feature is not considered as a categorical variable.

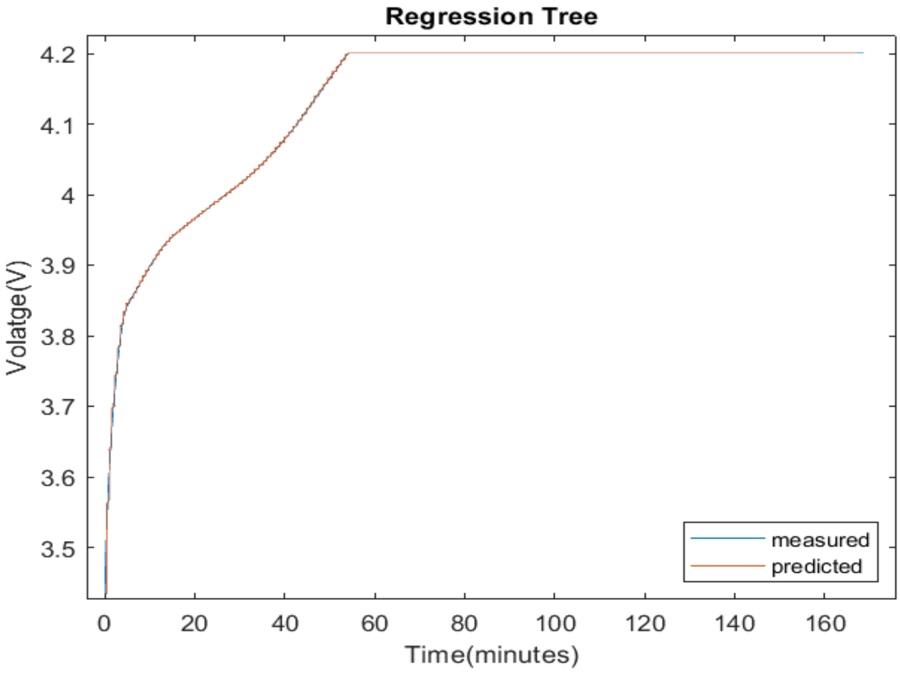
Regression Tree:

Unlike the SVR, regression tree classifies the categorical parameters in a combination set of continuous and categorical feature as shown in *Figure 23* . Before tuning hyperparameters the default parameters are selected for the regression tree: minimum leaf size '1', Maximum splits '50' and trained the algorithm with the single cycle data set.



- Coefficient of determination R^2:0.9993
- RMSE:57.2916
- Measured charging time:168.5805
- Predicted time:166.9118

Decision tree gives a better coefficient of determination for the trained data but RMSE value is large so that the variance of the model prediction is high or overfitting. In order to have less error tuning of hyperparameters is required to reduce the bias-variance tradeoff for the model. Hyperparameter tuning is done with random grid search method for 60 iteration with a 5-fold cross validation.

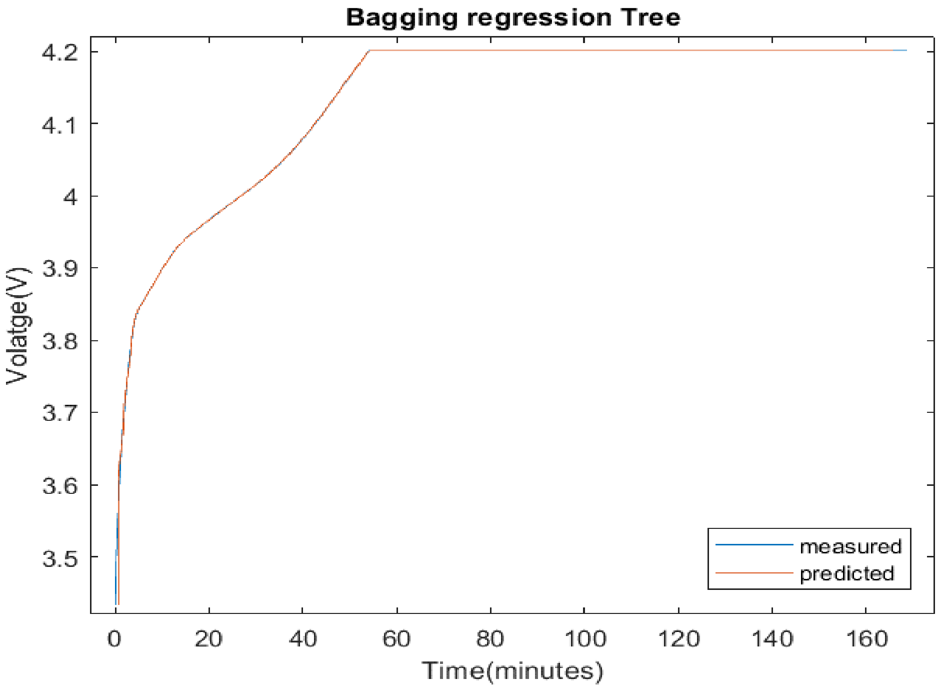


- Coefficient of determination R^2:0.9995
- RMSE:46.5870
- Measured charging time:168.5805
- Predicted time:166.9118

From the results even after training with optimized parameters the RMSE value is still huge and the predicted time is the same. For this in order to reduce the error, implementations of ensemble of trees is required.

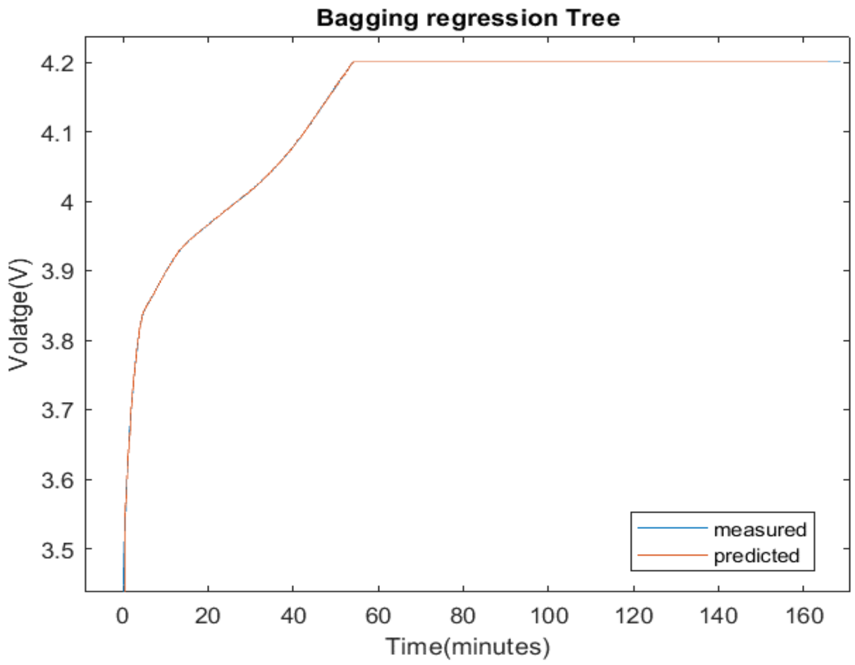
Bagging Decision Tree:

Bagging ensemble implementation of decision tree is used to reduce the variance in the model prediction. Initially default parameters are given to train the algorithm minimum leaf size 1, maximum number of splits 50, number of learning cycles 100.



- Coefficient of determination R^2:0.9999
- RMSE:19.6963
- Measured charging time:168.5805
- Predicted time:165.7147

Without optimal hyperparameters the RMSE is 19.6963 and the residual charging time is also more. Using random grid search trained again the model with the optimized hyperparameters.

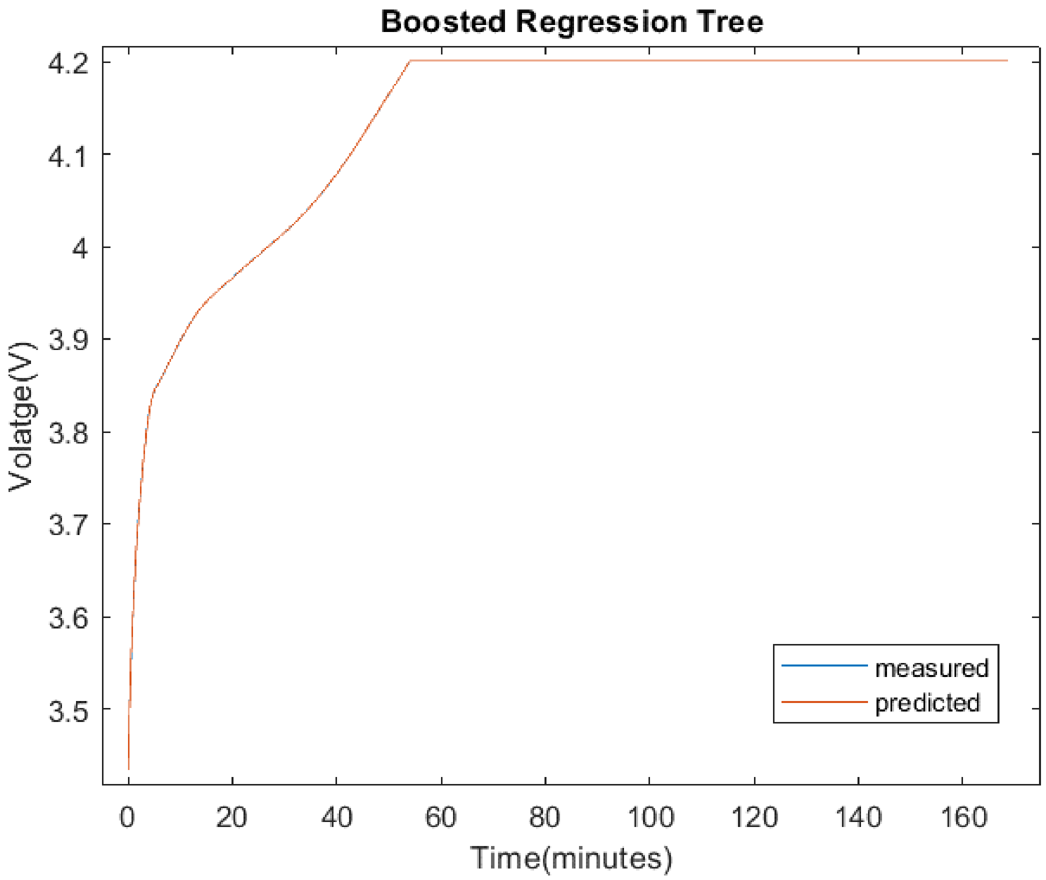


- Coefficient of determination R^2:0.9999
- RMSE:17.5120
- Measured charging time:168.5805
- Predicted time:165.6865

The optimized parameters are minimum leaf size 1, maximum number of splits 694, number of learning cycles 24. RMSE error is not varying much and the execution time of charging is also not predicting accurately. As from the bagging ensemble algorithm, it forms several numbers of trees for the same data set but at the end the prediction is done with aggregate of all the trees formed which means the error is aggregated instead of minimization while forming the trees.

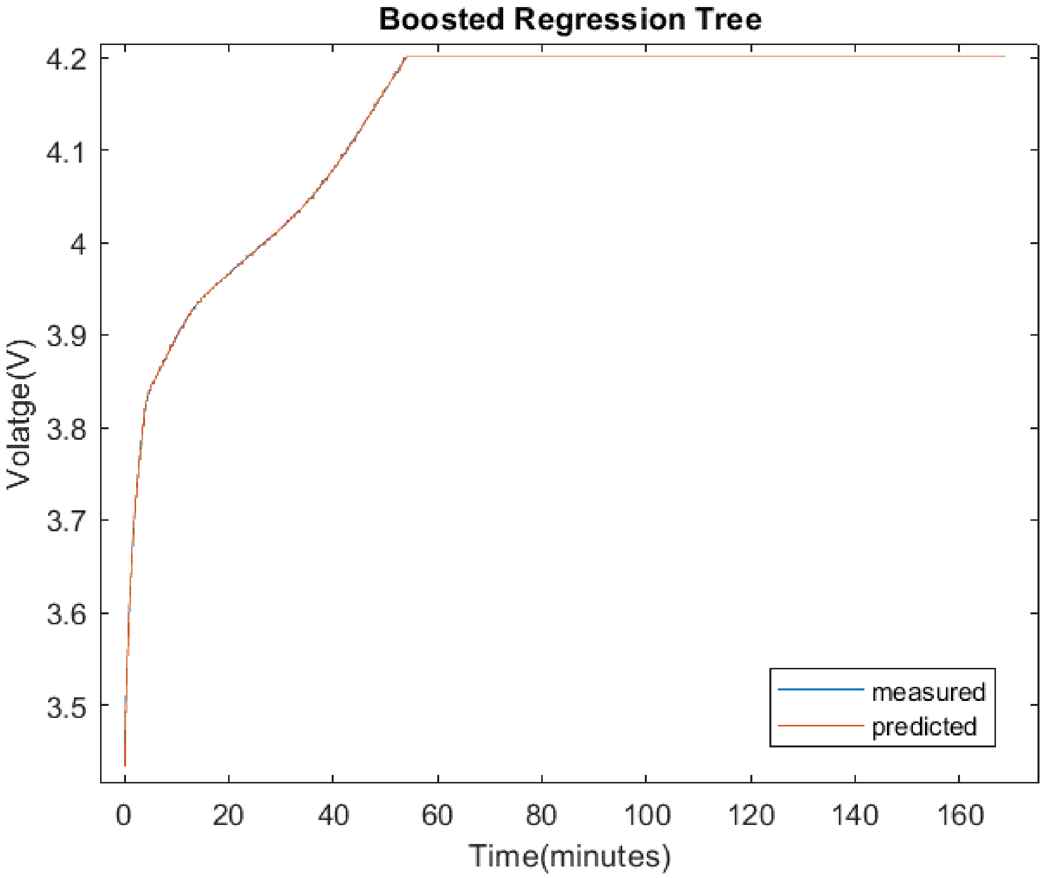
Boosted Decision Tree:

As boosted decision tree algorithm builds decision tree with minimizing the error from the previous decision tree, a boosted ensembles of decision tree with default hyperparameter values are applied and trained the algorithm.



- Coefficient of determination R^2:1.0
- RMSE:0.0621
- Measured charging time:168.5805
- Predicted time:168.5723

Boosted decision tree given a best prediction of execution time and less RMSE error with default parameters values and coefficient of determination is also 1 which is 100% prediction. As to make the process complete tuning of parameters is necessary.



- Coefficient of determination R^2:1.0
- RMSE:0.0014
- Measured charging time:168.5805
- Predicted time:168.3794

Tuning hyperparameters using random grid search gives an optimal prediction of charging time with very less RMSE error. From this it is evident that the best algorithm to predict the execution time of cell charging and discharging is Boosted decision tree.

Now trained the algorithm with “8 cycles” of three different types of cell charging and discharging processes dataset to find whether the Boosted decision regression tree predicts the output accurately. The following gives an overview of the three types of cell charging and discharging operating conditions.

- Battery Types: B0005, B0006, B0007
- Charge method: CC-CV
- Constant Voltage:4.2V
- Charge Cutoff Current:20 mA
- Charge Current: 1.5A

- Discharge method: CC
- Discharge current(load): -2A
- Discharge cut off voltage:2.7V, 2.5V, 2.2V
- Rated capacity: 2Ah

Below is the table for top 10 combinations of parameter set for 60 iterations of Random grid sweep method.

Table 4:Top 10 Parameter combination set with Random grid search method

Rank	**NumLearningCycles**	**LearnRate**	**MinLeafSize**	**MaxNumSplits**
1	56	0.196568758	19	253
2	123	0.106335898	4	5514
3	128	0.209834698	65	4984
4	241	0.246280386	59	41
5	198	0.149230429	71	102
6	12	0.358455747	27	3180
7	231	0.891362195	172	87
8	240	0.696944883	1	7
9	21	0.812778019	41	18
10	47	0.139883635	37	10

From the Table 4 rank 1 parameter set gives a less RMSE value, these parameter set are implemented for training 8 cycles of charging & discharging processes and predicted the 9th cycle execution time.

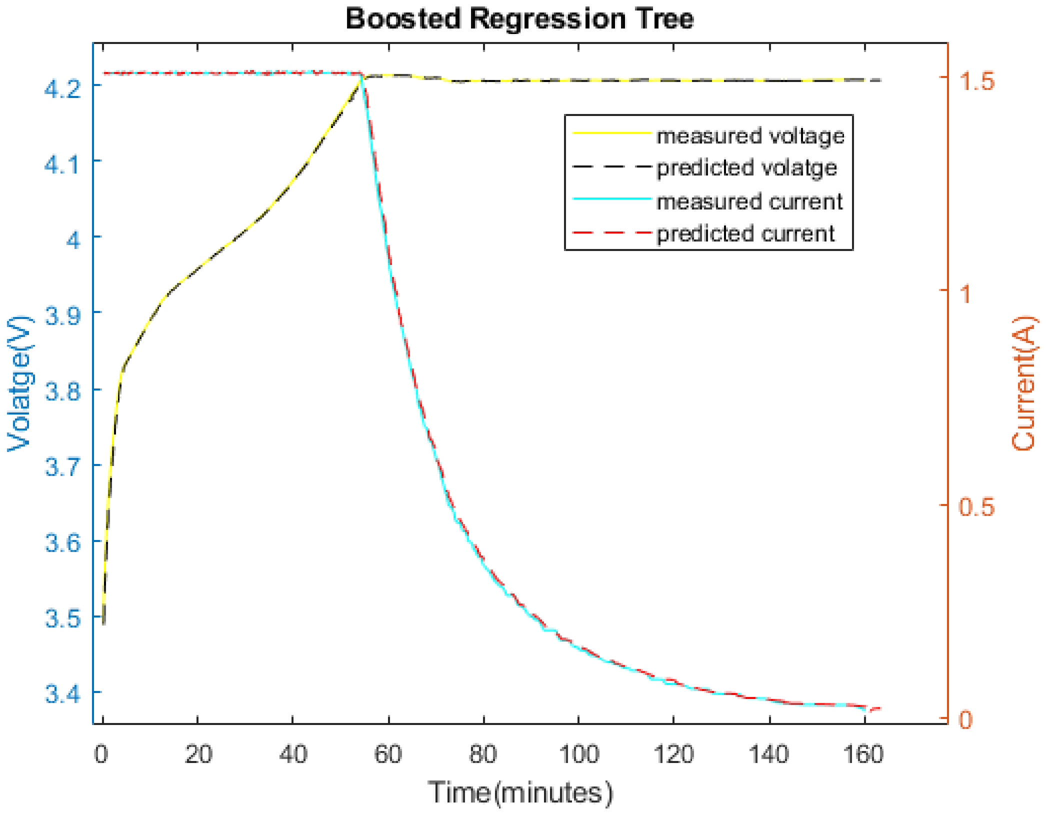


Figure 25: Charging Time prediction (9th cycle) for three different types of cell (8cycles of charging & discharging process)

- Coefficient of determination R^2:0.9997
- RMSE:19.7925
- Measured charging time:160.0891(minutes)
- Predicted time:161.1850 (minutes)

From the results of prediction of charging time for 9th cycle has a residual error of 1.0959. Training with larger dataset also algorithm predicts better and the hyperparameter tuning method “Random grid sweep” given a better parameter set.

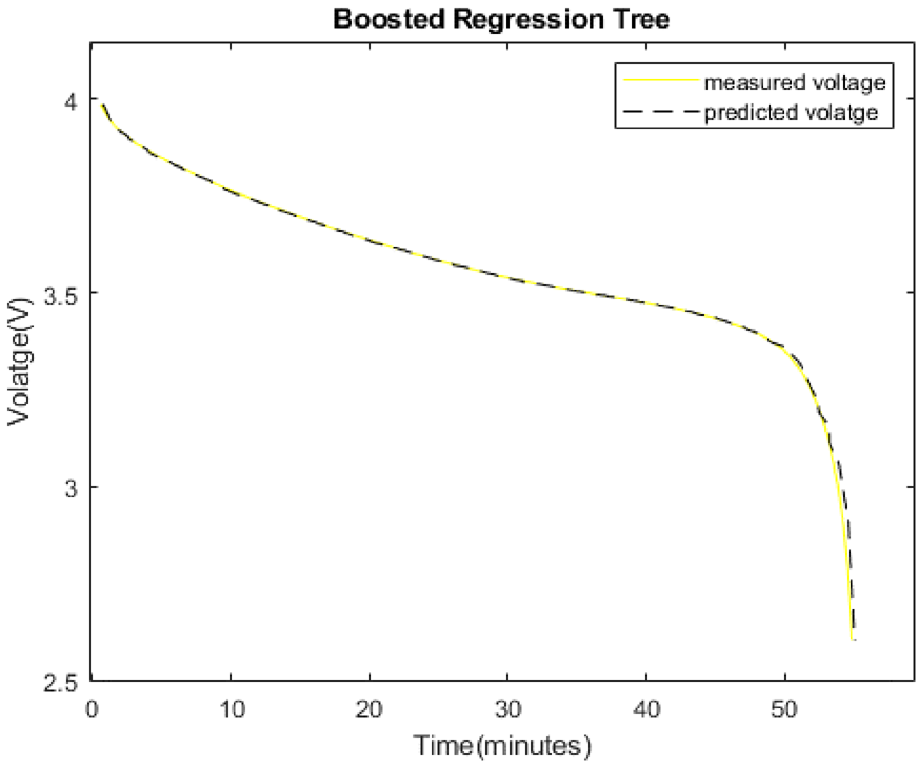


Figure 26: Discharging Time prediction (9th cycle) for three different types of cell (8 cycles of charging & discharging process)

- Coefficient of determination R^2:0.9999
- RMSE:0.7536
- Measured charging time:54.8315 (minutes)
- Predicted time:55.0596 (minutes)

Discharge time prediction of cell is even more accurate than charging time as the discharging process (CC discharge phase) under goes only in CC phase which is not complex where as charging of cell is done in CC-CV phases.

7 Conclusion:

The main target is to find the best appropriate Battery modelling method as a digital twin to predict the execution time of different types of cell charging and discharging process for different test strategies in Battery test bed. As per the available Battery cell test bed data, the developed Battery model predicts execution time for different types of cells from different manufacturers for different test strategies *(See section 3.0)*, but the chamber temperature used in the dataset is only for room temperature. According to the boosted decision tree functionality and its behavior in decision making for root split for different cycles, it is also expected that the predicts time for different chamber temperatures is also possible. By predicting the discharge time at different cycles, capacity fade of the battery can also be calculated as follows:

Predicted discharge time at cycle 9 is 55.0596(minutes)

Estimated capacity=Discharge current (A)× Predicted Time (h)

$$=2\times \frac{55.0596}{60}$$

$$=1.835 \text{ Ah}$$

Measured capacity= 1.824 Ah, Rated Capacity =2Ah

After “9” cycles of charge and discharge process estimated capacity of the cell is 8.25% and the measured capacity fade of the cell is 8.8 %. From the results it is also possible to estimate the capacity fade in a cell by predicted execution time of discharge of cell for the corresponding cycles.

Limitation: The training of data for all the cell is implemented starting from cycle number “2”. Because before testing the cell in a test bed, they are stored in warehouse. During this period self-discharge of cell take place and at the beginning of the test the state of the charge of the battery is not known.

8 References

[1] D. Linden, Linden's Handbook of Batteries, McGraw-Hill.

[2] Y. Gogus, Energy Storage Systems-Volume II, EOLSS Publications, September 30, 2009.

[3] "Battery University," 2003. [Online]. Available: https://batteryuniversity.com/learn/article/difficulties_with_testing_batteries.

[4] M. A. H. M. S. H. L. A. P. J. K. DICKSON N. T. HOW, "State of Charge Estimation for Lithium-Ion Batteries Using Model-Based and Data-Driven Methods: A Review," IEEE, Malaysia, 2019 .

[5] F. X. H. Fida Saidani, "Lithium-ion battery models: a comparative study and a model-based powerline communication," Advances in Radio Science, Stuttgart, 21 September 2017.

[6] L. R. F. Allen J. Bard, ELECTROCHEMICAL METHODS FUNDAMENTALS and APPLICATIONS, JOHN WILEY & SONS, INC., 2001.

[7] C. L. S. M. Z. Y. Zhang, "Battery Modelling Methods for Electric Vehicles - A Review," IEEE, France, 2014.

[8] E. S. a. S. K. K. Seyed Saeed Madani, "An Electrical Equivalent Circuit Model of a Lithium Titanate Oxide Battery," www.mdpi.com, Denmark, 2019.

[9] R. X. a. J. F. Hongwen He, "Evaluation of Lithium-Ion Battery Equivalent Circuit Models for State of Charge Estimation by an Experimental Approach," MPDI, China, 2011 .

[10] A. J. A. P. Y. K. N. R. N. I. Low Wen Yao, "Modeling of Lithium-Ion Battery Using MATLAB/SIMULINK," ResearchGate, Malaysia, 2013.

[11] S. R. ,. K. P. M. I. YOHWAN CHOI, "Machine Learning-Based Lithium-Ion Battery Capacity Estimation Exploiting Multi-Channel," IEEE, South Korea, June 5, 2019.

[12] P. K. a. N. Yodo, "A Data-Driven Predictive Prognostic Model for Lithium-Ion Batteries based on a Deep Learning Algorithm," Energies, Fargo, USA, 18 February 2019.

[13] A. Ng, "Coursear.org," Stanford, 1 April 2019. [Online]. Available: https://de.coursera.org/learn/machine-learning. [Accessed 1 april 2019].

[14] R. T. F. Trevor Hastie, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer Series in Statistics, 21. April 2017.

[15] "National Aeronautics and space administration," [Online]. Available: https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/.

[16] J.-H. L. I.-Y. H. T.-K. L. C.-Y. W. Jung-Song Moon, "An Efficient Battery Charging Algorithm based on State-of-Charge Estimation for Electric Vehicle," IEEE , 2011.

[17] "Electropaedia, Battery and Energy Technologies," [Online]. Available: https://www.mpoweruk.com/chargers.htm.

[18] V. Roman, "Towards DataScience," 23 December 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af.

[19] R. S. V. R. C. T. Ujera, "Analyzing correlation coefficient using software metrics," IEEE, Tirunelveli, India, 2017.

[20] Z. C. X. F. S. O. X. L. J. C. C. J. Anna Tomaszewska, "Lithium-ion battery fast charging: A review," ScienceDirect, Shanghai, PR China, 2009.

[21] S. D. S. J. Przemysław Przybyszewski, "Use of domain knowledge and feature engineering in helping AI to play Hearthstone," IEEE, Warsaw, Poland, 2017.

[22] P. S. Thomas K Abraham, Hands-On Machine Learning with Azure: Build powerful models with cognitive machine learning and artificial intelligence, Packt, OCtober 31, 2018.

Self-Declaration

I declare that I have developed and written the enclosed Master Thesis completely by myself and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The Master Thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Date and Location Signature