CHALMERS
UNIVERSITY OF TECHNOLOGY

Data-Efficient Hybrid Sampling Method for Battery Health Estimation and Prediction

A comparative study of sampling methods for lithium-ion batteries using machine learning

Master's thesis in High-Performance Computer Systems &
Sustainable Electric Power engineering and Electromobility

ADAM LINDGREN
ALICE NORDKVIST

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2024
www.chalmers.se

Master's thesis 2024

Data-Efficient Hybrid Sampling Method for Battery Health Estimation and Prediction

A comparative study of sampling methods for lithium-ion batteries using machine learning

ADAM LINDGREN
ALICE NORDKVIST

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of Systems and Control
Chalmers University of Technology
Gothenburg, Sweden 2024

Data-Efficient Hybrid Sampling Method for Battery Health Estimation and Prediction
A comparative study of sampling methods for lithium-ion batteries using machine learning
ADAM LINDGREN
ALICE NORDKVIST

Supervisors:
Evelina Wikner, Electrical Engineering
Pedro Petersen Moura Trancoso, Computer Science Engineering
Zhang Huang, Volvo Group

Examiner:
Torsten Wik, Electrical Engineering

Master's Thesis 2024
Department of Electrical Engineering
Division of Systems and Control
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: An artistic interpretation of the battery data decreasing by using a better sampling method. OpenAI. (2024), ChatGPT.

Typeset in LaTeX
Gothenburg, Sweden 2024

Data-Efficient Hybrid Sampling Method for Battery Health Estimation and Prediction
A comparative study of sampling methods for lithium-ion batteries using machine learning
ADAM LINDGREN
ALICE NORDKVIST
Department of Electrical Engineering
Chalmers University of Technology

Abstract

The rapid market adoption of electric vehicles (EVs) has been greatly driven by advances in lithium-ion batteries over the past years. As batteries are used under various load profiles in EVs, their capacity will decrease over time due to complex interactions of multiple degradation mechanisms. This necessitates advanced battery state-of-health (SOH) estimation and prediction. Existing methods rely on periodically sampled data for battery SOH estimation and prediction. However, with the ever-increasing measurements of battery data and complexity of algorithms, the limited hardware resources in the onboard battery management system (BMS) are strained. This makes the deployment of existing methods based on periodically sampled data challenging. More so, the research on the topic is limited and few sampling methods have been investigated, only validated on battery ageing datasets under static protocol cycling.

The purpose of this thesis work is to develop data-efficient sampling methods for battery SOH estimation and prediction. The performance of two sampling methods has been evaluated on an open-source dataset for realistic EV driving profiles, and benchmarked to the periodic sampling method. Battery SOH estimations and predictions were produced using a Gaussian Process regression (GPR) as it provides a principled approach to handling uncertainties. To optimize the model, 29 features were manually extracted and 4 different feature sets were created based on the feature's correlation with capacity.

The experimental results showed that by using a 3-feature set with a hybrid of an event-based and periodic sampling method, a more accurate SOH prediction could be achieved while using significantly less data. Specifically, the proposed sampling method and feature set reduced RMSE by 45.28% and the required data used for inferring the model by 99.4%. To achieve even better results, improved features for the machine learning model should be investigated. The results of the work show that the method seems promising to alleviate the limited hardware resources of the BMS.

Keywords: Lithium-ion batteries, Machine learning, State of health, battery health prediction, Sampling methods, Event-based sampling, Electric vehicles, Energy storage system.

Acknowledgements

This work has been conducted together with Volvo GTT in Gothenburg, Sweden together with their BMS Software and control team. Without the help from everyone on the team, this work would not have been possible and we would like to thank you for your tips, wisdom and guidance. We would especially like to thank Olle Friberg, the team's manager, for a welcoming and pleasant collaboration.

We would like to extend our appreciation to our supervisors Evelina Wikner and Pedro Petersen Moura Trancoso. Your guidance has been pivotal to the success of our thesis. We would also like to thank Torsten Wik for his time and for making the project possible. Finally, we are incredibly thankful for everything our last supervisor, Huang Zhang, has done for us. Your time, expertise and enthusiasm have been instrumental in reaching our goals.

Adam Lindgren & Alice Nordkvist, Gothenburg, June 2024

List of Acronyms

Below is the list of terms that have been used throughout this thesis listed in alphabetical order:

Acronyms

BESS	Battery Energy Storage System
BMS	Battery Management System
CC	Constant Current
CV	Constant Voltage
DoD	Depth of Discharge
EOL	End Of Life
EV	Electric Vehicle
FFT	Fast Fourier Transform
GP	Gaussian Process
GPR	Gaussian Process Regression
ICA	Incremental Capacity Analysis
LFP	Lithium iron phosphate
LIB	Lithium-Ion Battery
LOF	Local Outlier Factor
LTO	Lithium-Titanate Oxide
ML	Machine Learning
NMC	Nickel-Manganese Cobalt
RMSE	Root Mean Square Error
RPT	Reference Performance Test
RUL	Remaining Useful Life
SOC	State Of Charge
SOH	State Of Health
UDDS	Urban Dynamometer Driving Schedule

Nomenclature

Below is the nomenclature of parameters, and variables that have been used throughout this thesis.

$\mathbb{E}$	Expected value
f	Feature
FX	Feature X
i	Current
k	Covariance function
m	Mean
P	Probability
Q	Capacity
R	Resistance
t	Time
v	Voltage
Δ	Difference
μ	Vector of means
Σ	Covariance matrix
σ	Standard deviation

Contents

List of Figures

List of Tables

1

Introduction

With the escalating concerns regarding climate change; energy efficiency and sustainability have become more relevant than ever before. This can certainly be seen in electrical vehicles (EVs), which have seen a remarkable growth [1],[2]. Given that climate change is an ongoing issue, and that the need for vehicles is expected to keep increasing during the upcoming years [3], the number of EVs will likely continue to grow rapidly. To ensure a sustainable future, it is therefore essential to further develop electrical propulsion systems and energy storages.

Lithium-ion batteries (LIBs) have emerged as the primary choice for energy storage solutions across a wide range of applications, such as EVs and stationary energy storage systems [4],[5]. However, as a result of the complex interplay of different physical and electrochemical mechanisms, the performance of LIBs degrades over time [6]. To ensure safe and optimal usage of LIBs determining State-of-Health (SOH) is an essential state in a health-aware battery management system (BMS).

While LIBs are one of the best options for chemically storing electric energy they are also known for being sensitive to their operational conditions and environment, such as temperature. This is why a BMS is needed to help protect the cell, while also monitoring their health and operation. The BMS gathers data using a multitude of sensors that are used to, for example, protect against overcharging, overuse, and short-circuiting [7]. Most importantly, these sensors can also be used to determine the battery's present and future health. The problem is that predicting battery health accurately with limited degradation data is challenging because of the LIB degradation being nonlinear and in a very complex way influenced by operating conditions and manufacturing variations [8].

While a lot of work has been put into creating comprehensive models for predicting the SOH [9],[10], recent research has focused more on using machine learning techniques [11], [12]. However, far more limited research has been put towards how to sample the batteries' metrics effectively. Sampling of the data is an essential step as it determines what data is available to infer the battery's health. While this step is essential, sophisticated methods are yet to be developed for Battery Energy Storage Systems (BESSs). This thesis work investigates how the capacity loss can be predicted with more efficiently sampled battery data while retaining the battery health estimation and prediction accuracy.

1.1 Previous Work

BMS has become a hot topic of research but that has only partially extended towards the gathering of the data from the batteries. One relevant study by Yan *et al.* [13] uses a form of event-based sampling, which they refer to as Lebesgue sampling, to diagnose and prognose LIBs. They show that the traditional periodic sampling can be swapped for an event-based approach, "execution only when necessary" to keep performance while lowering the computational needs of the BMS. Their Lebesgue sampling monitors the capacity and when a certain threshold has been reached a new sample is taken, e.g. when the capacity has fallen 0.1 Ah since the previous measurement. On the other hand, their approach requires separate models for diagnosis and prognosis, requiring additional development resources to implement.

The work done by Yan *et al.* was followed up in 2022 when Niu *et al.* [14] implemented a Lebesgue-based strategy for sampling but to further improve the efficiency and accuracy they used a deep belief network together with a sequential Monte Carlo method. This shows that ML models are apt for use in health prediction applications.

1.2 Research Objective

To summarize, the main question this thesis addresses is the following:

How can battery management system's energy efficiency be improved by using more efficient sampling methods for health estimation in electric vehicles?

To reach the main goal of the thesis, the following subquestions had to be answered:

- How can a baseline ML model be created using periodic sampling?
- How can new sampling methods be developed that lower the computational complexity and data amount needed?
- How can the sampling methods and the ML model be optimized by using aging-relevant features as inputs?

1.3 Limitations

To limit the scope of the thesis an open-source dataset was used [15] and no new data was collected. The cells in the dataset are cylindrical lithium batteries with a positive electrode of Nickel Manganese Cobalt (NMC) and a negative electrode of Silicon-Graphite [16]. The temperature for all tests was reported to be a constant ambient temperature of 23 °C.

The project was limited to using one model for the predictions and time was not spent comparing with other potential models. As the same model was used for all sampling methods, comparisons could directly be made between the different sampling methods.

The goal was to develop and validate new techniques for sampling battery data. The proposed techniques will therefore only be implemented in software and, specifically, the results are gathered using Python.

The work was limited to looking at batteries on cell level, where more details are available and cells act independently from each other. Batteries on a pack level will therefore not be included.

In battery literature, the conventional definition for end of life (EOL) of a battery is 80% of the initial nominal capacity or 200% of the initial nominal internal resistance [12], [17]. The battery is then assumed to be unqualified for further use and has therefore reached its EOL. Because of this, the focus is on only the batteries before their EOL.

Another factor for BESS is the power output of the batteries. While this is highly relevant for EVs, and we encourage future work to be put into this, it is not something that was explored.

1.4 Ethical Aspects of Consideration

With the research objectives in mind, it is always relevant to showcase what aspects may be beneficial and/or detrimental to society and the environment. A more efficient way of predicting the SOH of the battery opens up the possibility to better control its working conditions which extends the battery's lifetime. There may also be some negative ethical and environmental aspects as a result of this work. An example of a negative consequence is that faults in the program can lead to unnecessary degradation of the battery which, in contrast to previous arguments, would lead to faster reaching EOL.

An economic aspect relevant to the thesis and the relatively nascent market of EVs is that prices for EVs are currently high meaning that the market is primarily accessible for wealthier people and organisations. This implies that this work will mainly benefit these individuals and institutions. On the other hand, it is anticipated that, in the long term, it could contribute to reduced prices for EVs and other BESSs applications fostering accessibility across a broader spectrum of society.

For the work, a constant ambient temperature and pressure will be assumed. This can be seen as an ethical discussion since it means that the results of the work may not be fully accurate to every location depending on the environmental conditions. For example, the requirements to make an optimal method in the context of Sweden and its cold climate may differ from another country with a warmer environment.

Lastly, to limit the project within reasonable bounds, aspects regarding the manufacturing of the battery were outside this work's scope. However, there are well-known problematic aspects, such as human rights abuse, related to the extraction of the materials used in battery production, such as cobalt, that should be mentioned when working with related subjects [18].

1.5 Structure of the Paper

The rest of this paper is structured as follows: Section 2 introduces the background of the project and clarifies some theory needed to understand the methodology. Section 3 describes the process of the work; what and how it was done. Section 4 presents and discusses the results of the features and the prediction model. Section 5 summarizes and concludes the work and ends by bringing forth potential future work.

2
Background

2.1 Lithium-ion batteries

LIBs are rechargeable batteries that can take on many shapes and materials. They are commonly used in a large range of different applications such as portable electronics, power tools, and are the primary choice for energy storage in electric vehicles because of their relatively high energy density, fast charging and their low self-discharge rate [4], [19].

A battery is built up of three main parts: a negative electrode, a positive electrode and an electrolyte. A separator is placed between the electrodes to avoid contact that would cause a short circuit in the battery. Some common materials of LIBs are Nickel-Manganese Cobalt (NMC), Lithium Iron Phosphate (LFP), graphite and Lithium-Titanate Oxide (LTO). The materials used make for different behaviours, advantages and disadvantages.

The anode and cathode store lithium and lithium ions are transported between them through the electrolyte during charge or discharge. When the battery is being discharged, the lithium-ion particles move from the negative to the positive electrode. Electrochemical oxidation and reduction reactions take place at the electrodes, redox reactions. On the negative electrode oxidation take place, releasing an electron that moves thought the external circuit to the positive electrode where reduction take place. When the battery is being charged, the lithium ions and the electrons are moving in the opposite direction, which charges the battery. The maximum capacity of the battery depends on the amount of active electrode material and its potential to store lithium. Different materials have different electrochemical potentials. The cell voltage is the difference between the two electrode materials electrochemical potentials and will therefore depend on the battery chemistry [20].

The rest of the section will talk about the most important aspects of LIBs for this work, which are the capacity, how it is determined, and the SOH; what it is and how it is defined.

2.1.1 Battery Capacity

The battery capacity is the main metric required for estimating the health of a battery. Capacity is a metric that specifies the amount of energy, or specifically,

the amount of electric charge stored in a battery. To accurately establish its value is therefore a top priority when creating a model to determine the SOH. Coulomb counting is a common way to determine a battery's capacity. It is a method that works by integrating the current supplied to the battery during charging. A battery's capacity can therefore be explained by

$$Q_{bat} = \int_0^T i(t)\ dt, \tag{2.1}$$

where Q_{bat} denotes the battery capacity, i is the charging current, and T is the time required for the battery to be charged from 0% to 100% [19].

2.1.2 State of Health

Since batteries undergo degradation during their lifetime, both due to calendar-ageing and usage, the prediction of their health is necessary to make sure to use the battery properly. An accurate SOH prediction opens up the possibility of ensuring safe operation, extending the lifetime and minimizing costs of maintenance for the battery [21].

There is no standardised way of estimating the SOH for a LIB, different organisations use different methods to define the current state of the battery [11]. The general purpose of defining and estimateing the SOH of the battery is to better control it and predict the RUL until it reaches its EOL. The definition of EOL for a LIB may differ for different applications. The broad way to define it for EV applications is when it can no longer reach the requirements for typical usage. This is usually calculated using the capacity and internal resistance as key parameters, where EOL is reached if either the capacity has dropped below 80% of its initial capacity or the internal resistance has reached a value twice its initial value [12].

The reason for the degradation in the battery and why these key parameters are affected depends on several factors such as temperature, intensity of usage, and variability during manufacturing, which makes the predictions hard to perform [17]. The degradation can be defined using the following equations:

$$SOH = \frac{R_{EOL} - R_{present}}{R_{EOL} - R_{initial}} \tag{2.2}$$

$$SOH = \frac{Q_{max}}{Q_{initial}} \cdot 100 \tag{2.3}$$

where R_{EOL}, $R_{present}$ and $R_{initial}$ is the internal resistance of the battery at EOL, now and at the initial state respectively [12]. Q_{max} is the maximum capacity at the current cycle and $Q_{initial}$ is the initial capacity of the battery.

2.2 Prediction model

When predicting the SOH for a battery, there are several approaches to make the estimation. The three general approaches used are analytical model-based methods, empirical ageing maps, and data-driven models [17], [22].

The analytical model-based method is built up by a "physics-based" battery model usually using the equivalent circuit model to predict the deviation. Other analytical model-based methods are using a electrochemical models or continuum approaches using porous-electrode theory. The model uses parameters such as voltage, current, and temperature to simulate battery behaviour under various operating conditions. By comparing these simulations with actual battery performance data, deviations can be analyzed and used to estimate the battery's SOH [17], [22].

The idea of empirical ageing maps is to model the capacity fade as a function of time. The method then uses an ageing function or a look-up table of previous parameter relations for a certain operating condition to predict the future or present capacity. Recent research has divided it into two functions, one for calendar ageing and one for cycle aging, to model each component more accurately [17].

To predict capacity using data-driven approaches means making a model directly from the data instead of trying to physically derive the behavior of batteries. This model can either use raw data or features as input. Extracting features requires an additional step in pre-processing where relevant parameters are extracted from the data. The advantage of using a feature-based data-driven approach is that the model can be less complex and require less training data [17]. ML models are a commonly used data-driven approach based on a program performing specific tasks without being specifically programmed to do so. Using this, a model can be created that predicts the battery's capacity without using knowledge about the internal chemical reactions or behavior of the battery.

There are numerous models and architectures that exist in the field of ML depending on what the model should be used for. A Gaussian Process Regression (GPR) has been shown to be a suitable ML model for predicting capacity. It offers flexibility in its use cases and is commonly used for regression [23]. GPR is a supervised, non-parametric ML model that in contrast to simple regression methods, such as linear regression, outputs probabilistic functions to give a more accurate representation of the prediction and its' uncertainty.

2.2.1 Gaussian Processes

The details of a Gaussian Process (GP) can further be explained by its' underlying math. The output for any finite input points $\mathbf{X} = \mathbf{x_1}, \mathbf{x_2}, .., \mathbf{x_N}$ is a Gaussian probability distribution $P_\mathbf{f} = p(f(\mathbf{x_1}), f(\mathbf{x_2}), .., f(\mathbf{x_n}))$ with a mean $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ that determine its confidence interval. Therefore, the GP itself is

a collection of random variables with a joint Gaussian distribution as described by

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{2.4}$$

where $\mathbf{x}$ is the input vector, $m(\mathbf{x})$ is the mean of the input, $k(\mathbf{x}, \mathbf{x}')$ is a covariance function between the input vector and the previous vector and $\mathcal{GP}$ refer to it being a Gaussian Process [23],[21].

The mean and covariance function, respectively, are denoted by

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{2.5}$$

and

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))^T], \tag{2.6}$$

where $\mathbb{E}$ is the expected value and $f(\mathbf{x})$ is a sample from the GP for inputs $\mathbf{x}$. This is the definition for a single vector but to fully explain the probability distribution of all points $\mathbf{X}$, the following equation is used:

$$P_{\mathbf{f}}(\mathbf{X}) = \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x_1}) \\ m(\mathbf{x_2}) \\ \vdots \\ m(\mathbf{x_n}) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x_1}, \mathbf{x_1}) & k(\mathbf{x_1}, \mathbf{x_2}) & .. & k(\mathbf{x_1}, \mathbf{x_n}) \\ k(\mathbf{x_2}, \mathbf{x_1}) & k(\mathbf{x_2}, \mathbf{x_2}) & .. & k(\mathbf{x_2}, \mathbf{x_n}) \\ \vdots & \vdots & \vdots & \vdots \\ k(\mathbf{x_n}, \mathbf{x_1}) & k(\mathbf{x_n}, \mathbf{x_2}) & .. & k(\mathbf{x_n}, \mathbf{x_n}) \end{bmatrix}\right), \tag{2.7}$$

where the collection of mean values $m(\mathbf{x_1}), m(\mathbf{x_2}), .., m(\mathbf{x_n})$ can be denoted as μ and the matrix of covariance functions $k(\mathbf{x_i}, \mathbf{x_j})$ as Σ, henceforth refereed to as the covariance matrix, such that

$$P_{\mathbf{f}}(\mathbf{X}) = \mathcal{N}\left(\mu, \Sigma\right), \tag{2.8}$$

which is a more concise definition.

The covariance matrix, Σ, defines both the shape of the distribution as well as the characteristics of what is being predicted. Setting it up correctly is therefore paramount for GPs to behave similarly to what they are trying to model. By defining the covariance matrix in specific ways it becomes possible to model scenarios including complex nonlinear systems such as LIBs [23].

Designing the covariance matrix is done by using kernels of the GP. Here, kernel is simply a different word used for covariance function, and its general purpose is to return a similarity measure between function values at two inputs x and x', i.e., cov(f($\mathbf{x}$), f($\mathbf{x}'$)) = k($\mathbf{x}, \mathbf{x}'$). There are a number of different kernels commonly used, three of which will be highlighted here.

The Radial Basis Function (**RBF**) is the most standard kernel currently used and is also known as the Squared Exponential (SE) kernel. The kernel function is defined as:

$$k_{RBF}(x - x') = \exp\left(-\frac{|x - x'|^2}{2l^2}\right), \tag{2.9}$$

where the parameter l is the lengthscale which defines how the function changes along the horizontal axis. A small value of l means that the function oscillate faster and more dramatically. An underlying assumption for RBF is that the function it models is smooth and infinitely differentiable, something limiting its' use case in real-world processes. The output from a GP is built up by an infinite number of samples and produces a Gaussian distribution at each point x. A handful of samples from an RBF kernel along with its corresponding distribution are shown in Figure 2.1. The visualisations are made with an online tool found at [24].

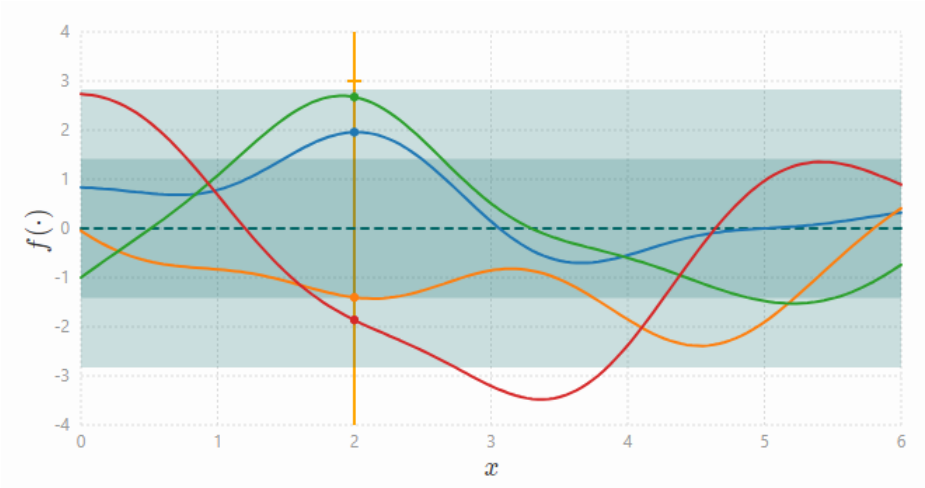

Figure 2.1: Visualisation of an RBF kernel's prior distribution used for GP regression, with four coloured sample functions. The dotted line represents the mean, with its' first and second standard deviation being coloured dark and light teal respectively.

Another commonly used kernel is the **Matérn** kernel which is a class of functions that are defined as

$$k_{Matérn}(x - x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}(x - x')}{l} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}(x - x')}{l} \right), \tag{2.10}$$

where ν is the smoothness factor, l is again the lengthscale, Γ is the gamma function and K_{ν} is a modified Bessel function. It is k times differentiable, where k $< \nu$. The kernel is usually used with $\nu = p + 1/2$, p being a natural number, making it significantly simpler. For ML two commonly used values are $\nu = 3/2$ & $\nu = 5/2$, i.e.

$$k_{Matérn_{\nu=3/2}} = \left(1 + \frac{\sqrt{3}(x - x')}{l} \right) \exp \left(-\frac{\sqrt{3}(x - x')}{l} \right) \tag{2.11}$$

$$k_{Matérn_{\nu=5/2}} = \left(1 + \frac{\sqrt{5}(x - x')}{l} + \frac{5(x - x')^2}{3l^2} \right) \exp \left(-\frac{\sqrt{5}(x - x')}{l} \right) \tag{2.12}$$

Contrary to the RBF, the Matérn kernel has a varying amount of smoothness allowing it to better model a larger variety of scenarios. Note that when $\nu \to \infty$ the

Matérn kernel becomes identical to the RBF kernel. A handful of samples from a Matérn kernel along with its corresponding distribution are shown in Figure 2.2.

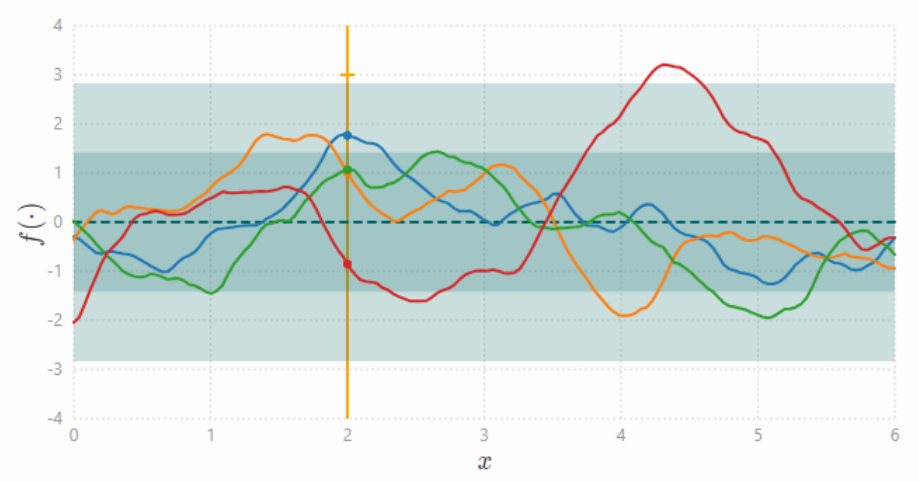

Figure 2.2: Visualisation of a Matérn kernel's prior distribution, used for GP regression, with $v = 3/2$ and four coloured sample functions. The dotted line represents the mean, with its' first and second standard deviation being coloured dark and light teal respectively.

The last kernel to mention is the **Linear** kernel. It is in practice equivalent to the classic Bayesian linear regression and is given by

$$k_{Lin}(x - x') = \sigma_b^2 + (x - c)(x' - c) \tag{2.13}$$

where c is the offset that determines at what x-coordinate the posterior lines will go through, and σ_b^2 determine where the function will cross the x-axis. An example of how a linear kernel distribution can look like can be seen in Figure 2.3.

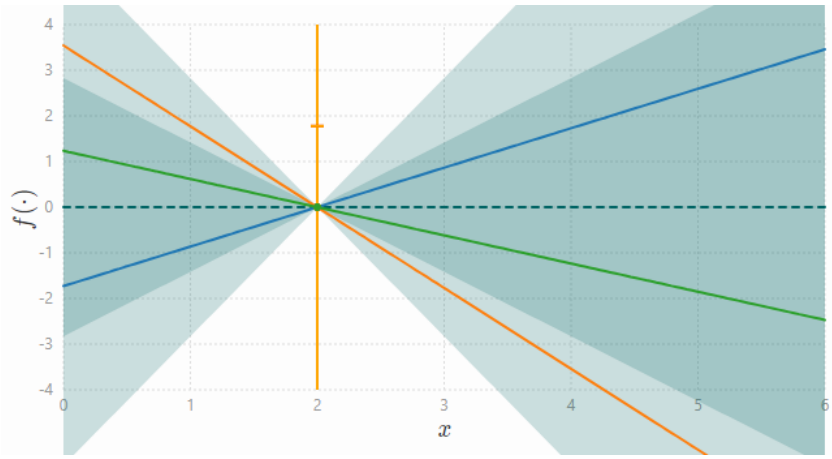

Figure 2.3: Visualisation of a Linear kernel's prior distribution used for GP regression, with three coloured sample functions and parameter $\sigma_b^2 = 2$. The dotted line represents the mean, with its' first and second standard deviation being coloured dark and light teal respectively.

The kernels can be used stand-alone or be combined to create a composite kernel that best fits the situation where the GP will be implemented. When trained, the hyperparameters of this composite kernel are optimized to the training data. The next step is to let the trained model observe the test data. With a data point, the probability collapses into a known value as seen in the example of Figure 2.4.

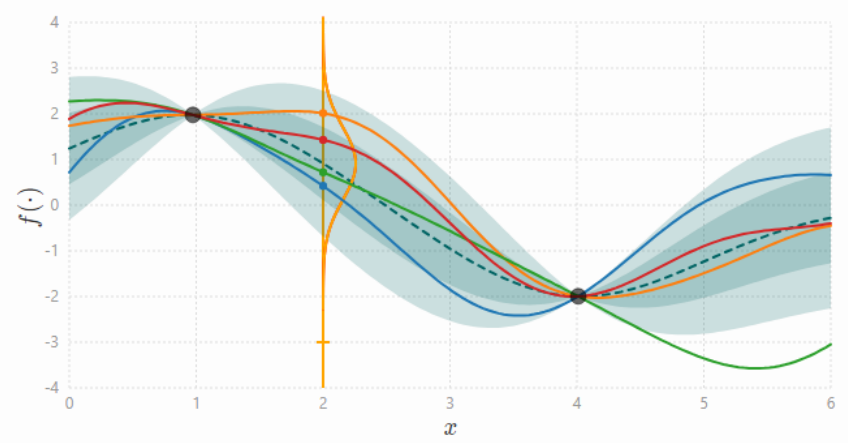

Figure 2.4: Visualisation of the posterior distribution of the RBF kernel with two data points being observed. The dotted line represents the mean, with its' first and second standard deviation being coloured dark and light teal respectively.

These two parts: optimizing a kernel and having observed data points to collapse the distribution, make it possible to train a model capable of approximating the behaviour of LIBs.

2.3 Sampling methods for battery health estimation and prediction

Different sampling methods can be used for collecting information from batteries. Two that offer distinct strengths and advantages are Periodic and Event-based. Periodic sampling is the more straightforward and traditional approach where data is collected with a predetermined period [13]. As shown by Gu *et al.* [25], a higher sampling frequency for this method results in better health diagnosis at the cost of more data to handle. Event-based sampling, also known as Lebesgue-based, is the other alternative and has been shown to have high potential, first by its pioneers Åström and Bernhardsson [26] and recently by Yan *et al.* [13] and Niu *et al.* [14]. It is an approach that can be described as a reactive way of collecting data, only doing so when the measurement has changed enough to justify it. This ensures that sampling is only done when the data has been changed and is relevant for new processing [13].

Three different sampling methods, one periodic and two different event-based sampling methods were explored and are explained in this section.

2.3.1 Periodic sampling method

The periodic sampling method is based on data being extracted with periodic time intervals regardless of any events that may happen within the cell. The intervals at which the data is sampled are kept constant throughout the whole life of the cell as illustrated in Figure 2.5.

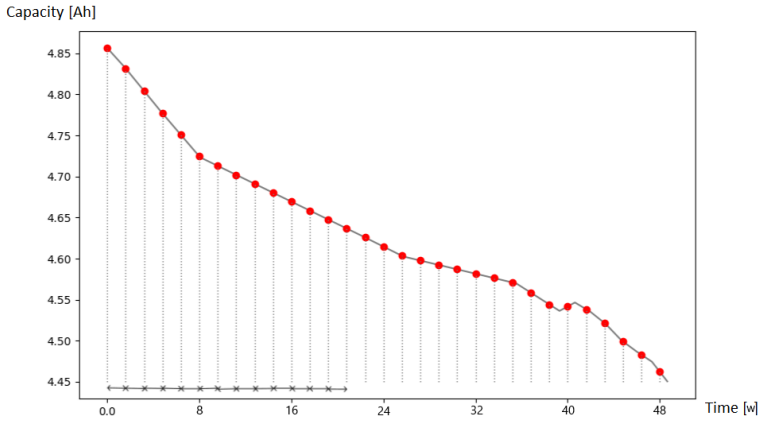

Figure 2.5: Visual representation of a periodic sampling with a period of 1.6 weeks and samples represented by the red dots.

2.3.2 Event-based sampling method 1

For the first event-based sampling method, data will be sampled as the capacity of the cell has fallen by a certain amount, ΔQ, as illustrated in Figure 2.6. The value of ΔQ can be decided depending on the purpose and can either be a fixed absolute value or a percentage of the initial capacity value. By choosing it to be a fixed value, ΔQ would take on the same value for all cells. If ΔQ is set to a percentage of the initial value, ΔQ is kept constant throughout the whole life of the cell but would vary between different cells, because of variations in their initial capacity. Using the second defined ΔQ, the number of samples taken from the data during a cell's life can be fixed since the EOL per definition will occur at 80% of the initial capacity.

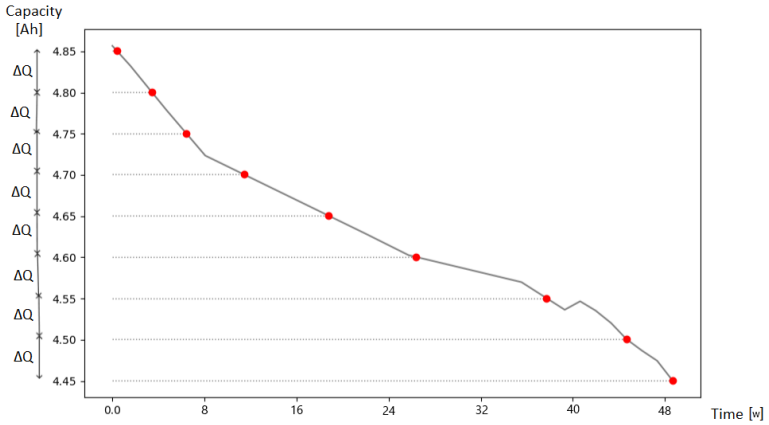

Figure 2.6: Visual representation of the first event-based sampling method with a ΔQ of 0.05 Ah and red dots representing the samples.

2.3.3 Event-based sampling method 2

The other event-based method is sampling the data when the capacity is degrading faster than a certain threshold value. Fixed time intervals are derived from the capacity curve. Within each time interval, the capacity fade is determined and if the decrease is larger than the threshold value, the sampling is triggered. This means that the time intervals, ΔT, will be kept constant for the whole life of the cell while ΔQ is a parameter calculated from the capacity and will therefore vary within each time interval. This sampling method is illustrated in Figure 2.7, showing the fixed time intervals and the different ΔQ. Using this sampling method, data is collected when the degradation of the capacity is quick and when important events may occur. The number of samples cannot be decided for this method since it is dependent on the behavior of the capacity fade of the specific cell.

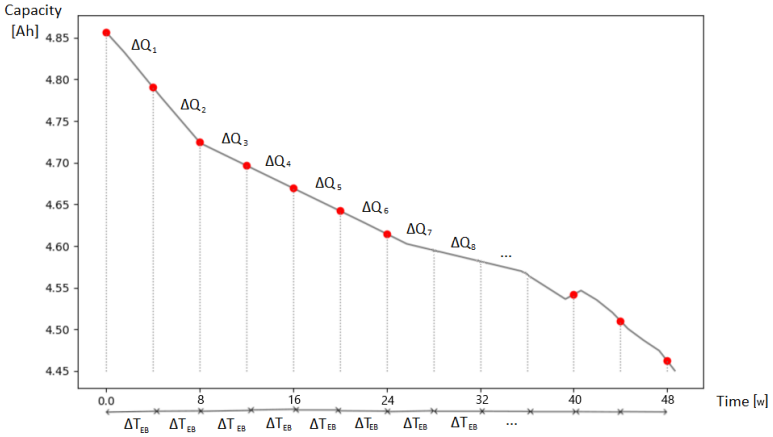

Figure 2.7: Visual representation of the second event-based sampling method with a period of 4 weeks and the red dots representing the samples where ΔQ_i has reached the threshold.

2.4 Dataset

The dataset used in this work was generated by Pozzato *et al.* [15] at Stanford Energy Control Laboratory, Stanford University. The dataset consists of experimental cycling and diagnostic data for ten NMC cells, LG Chem's INR21700-M50T [16], that were tested over 28 months in a temperature-controlled environment of 23°C. It was chosen because it uniquely replicates a typical EV discharge profile by following the Urban Dynamometer Driving Schedule (UDDS), making the dataset a more realistic representation of how batteries would be aged in a real-life scenario.

The data is divided into 14 phases, with a reference performance test (RPT) conducted at the end of each phase. Before ageing begun, a starting RPT was performed to characterize the battery's initial condition. The RPT includes tests to determine the following: capacity through a capacity test, high frequency resistance through a hybrid pulse power characterization test, and impedance through electrochemical impedance spectroscopy testing.

The load cycles consists of 6 steps. The first 4 steps are for charging up to 100% State Of Charge (SOC), where step 1 and 3 are constant current charging and step 2 and 4 are constant voltage charging as seen in Figure 2.8. The last 2 steps are for discharging, where step 6 follows the UDDS protocol and shows the dynamic nature of discharge that UDDS entails. Discharging is finished when SOC reaches 20%. The data for current, time during phase, charge capacity, discharge capacity, step_index and open circuit voltage calculated from SOC was collected with a frequency hovering around 10 Hz.

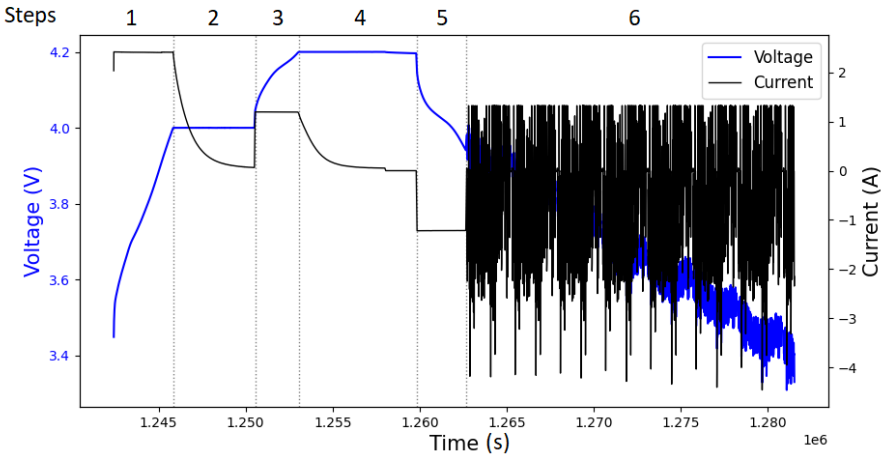

Figure 2.8: The 6 steps of an arbitrary cycle for current (black) and voltage (blue) from the ageing battery data set. Note the location of 0 A on the right y-axis.

The cells were charged with different C-rates during the second step of each cycle. The C-rates for each cell are represented in Table 2.1 together with the names of the ten cells. All other steps were cycled identically between the cells.

Table 2.1: Names and C-rates for the cells

Cell name	W3	W4	W5	W7	W8	W9	W10	G1	V4	V5
C-rate for step 2	3C	C/4	C/2	C/4	C/2	1C	3C	3C	C/4	1C

3
Methodology

The work is divided into two main phases, first establishing a baseline using periodic sampling, and secondly testing the proposed sampling methods. A visualisation of the ML pipeline can be seen in Figure 3.1. By implementing new sampling methods the data available for the system changes, which requires the features to be adapted appropriately. By also designing the features correctly, the model can operate more efficiently making the features important for the system.

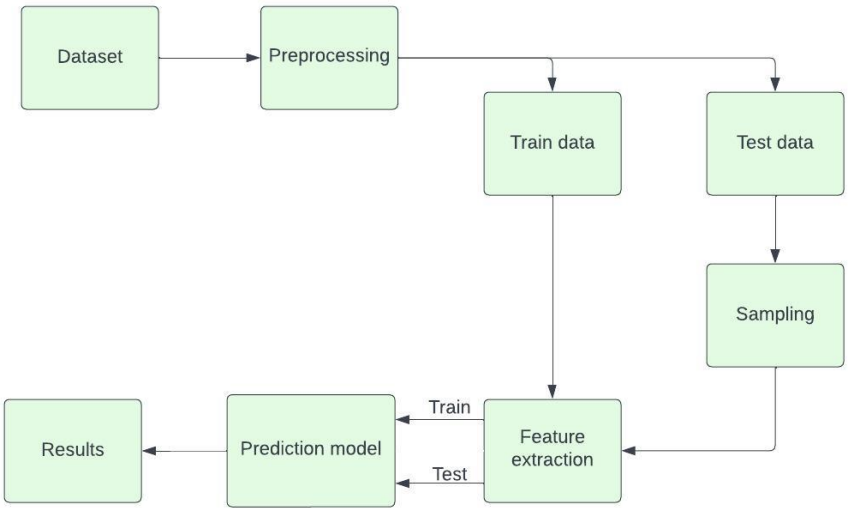

Figure 3.1: Flowchart of the proposed ML pipeline for capacity prediction.

3.1 Experimental design and pre-processing

The experimental results are gathered on a system running an Intel Core i7-11850H with Windows Version 10.0.19045 and Python 3.11.7. The libraries used in Python can be found in Table 3.1.

Table 3.1: Python libraries used for implementation.

Library	Version
Numpy	1.26.3
Pandas	2.2.0
SciKit-learn	1.4.0
MatPlotLib	3.8.2
SciPy	1.12.0
Seaborn	0.13.2

To make the ML models as efficient as possible, a preprocessing phase to shape the data with the right parameters was added. The preprocessing was initially planned to include filtering of the data and regression to reduce the dependency of its sampling frequency.

For filtering, mainly two different methods were investigated to make it as suitable as possible, window-based filtering and local outlier factor (LOF). When running those filters on the data a couple of issues were detected. Using the window-based filtering, some of the periodic peaks in the data were removed which would change the characteristics of the data. For the LOF, one could also see that useful data were removed while the noise remained. Since none of the methods resulted in a fully suitable solution and the data seemed to not include much noise or outliers, the decision was made not to filter the data in the preprocessing. This way the ML model had to handle all the data and filter these things out based on the overall shape of the data.

After filtering, we aimed to perform regression analysis on the raw data to replicate the behaviour. The regression would give a function of time for the different parameters that would not be dependent on the sampling frequency of the measurements of the raw data. To make the regression, a couple of different methods were investigated including simple moving average, spline regression, gaussian process regression, random forest, and Fast Fourier transform (FFT). None of the mentioned methods were able to provide a good-looking regression of the large data set used, except FFT. The regression resulted in a function describing the behaviour of the different parameters. However, because of the complex shape of the data, the function contained a large amount of components which caused problems in handling it for further computation. Because of the computation problems, the regression was not used further in the project.

As a result, the final preprocessing for the data was a removal of duplicate values and re-structuring the file format. When structuring the data, it was converted from separate MATLAB files for each phase into csv files, one for each cell including all phases with a downsampling factor of 50. While current, voltage, and step index could be directly extracted, time needed to be converted into absolute time. This is because after each RPT is completed the time measurements are reset, meaning that each phase counts time from 0. To convert to absolute time, the last time value of each phase is stored away to be used as the starting time of the following phase. Additionally, two new variables were introduced to simplify future calculations. The first variable introduced was the start time for the current cycle which could be used as a cycle count. The second variable introduced was capacity. The most reliable data points available for the capacity were obtained from the RPTs at the start and end of each phase. To get values during the phases, the capacity was linearly interpolated between the RPTs. The interpolated capacity curves for the cells can be seen in Figure 3.2.

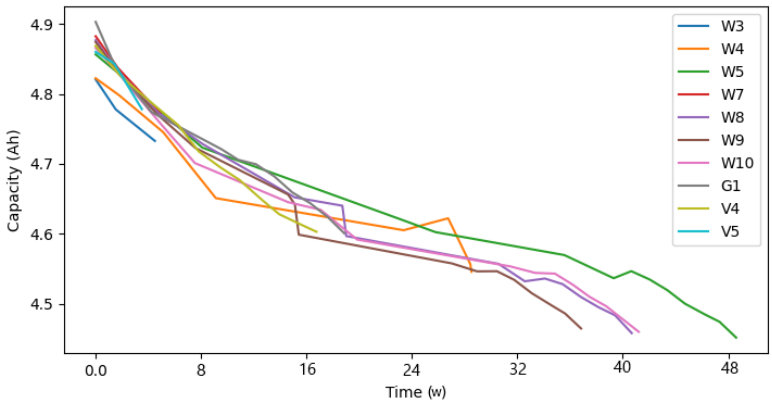

Figure 3.2: Absolute interpolated capacity for all cells from the ageing battery dataset.

When using a data-driven model, one crucial design decision is how to split the data into test data and training data. A label was added to each data point called 'cell marker' to identify which cell a certain data point was taken from. This marker allows the data to be split by cell, enabling training and testing by specific cells for the ML. When looking at the interpolated capacity for all cells (see Figure 3.2), it can be seen that not all cells go through all 14 phases. It can also be seen that capacity starts to diverge after 16 weeks. Because of its overall predictable behaviour and that it went through all phases, cell W8 was chosen to be the designated test cell. The model and pipeline are designed to work for any cell but the results for this work are based on this cell.

3.2 Sampling methods for battery health estimation and prediction

The project will compare capacity predictions using different sampling methods. In this section the implementation of the three sampling methods that were used are explained, with the first being the periodic sampling used as the baseline. The sampling is performed only on the test data since training the ML model is done using data from all cycles.

3.2.1 Periodic sampling method

Periodic sampling was implemented to establish a baseline for the different sampling methods. The main idea behind implementing the periodic sampling is a simulation of reading measurements from a vehicle's sensors by downsampling the dataset's measurements. As the dataset has a measurement frequency of 10 Hz a downsampling by e.g. 600, results in a simulated sampling frequency of 1/60 Hz = 1 min. Therefore setting the downsampling factor of the data directly amounts to setting the time period of the sampling.

Downsampling is not only done to act as periodic sampling, but it is also necessary to reduce the size of the dataset due to computational limitations. Depending on the features used for the prediction model, the downsampling factor needed to be adjusted according to the size of those features. For example, using all the time data from the time series takes more memory and storage than a narrow feature that is more directly correlated with capacity. The downsampling factor was therefore set as low as possible while still being able to fit the data in memory when training the model. To accomplish this the downsampling factor was incrementally increased until the data could be successfully allocated. For the periodic sampling, the smallest downsampling found was a factor of 5000 which corresponds to a simulated sampling frequency of 8.3 min and was used for the baseline results. By using the established time period, the measurements from current, voltage, step index, and time were extracted for each phase.

3.2.2 Event-based sampling method 1

For the first event-based sampling method, data were extracted each time the cell capacity from an RPT decreased by 0.6% from its previous capacity. The method is illustrated in Figure 2.6 (note that the value of ΔQ in the figure is an example and does not represent 0.6% of the initial capacity). To accomplish the sampling for this criteria, the capacity curve was investigated. Whenever a decrease of $Q_{init} \cdot 0.006 \cdot x$ Ah was observed, the time for the event was saved into an array. x was specified to be an integer > 0 and the criteria could only be triggered once for each value of x for one cell.

For each time the sampling was triggered, the cycle for this event was identified. The data for the whole cycle was then extracted and considered the event-based sampled data. To obtain information on the cell's initial state, the second cycle was also extracted regardless of any capacity fade. The number of events triggering a sample for each cell is specified in Table 3.2.

Table 3.2: Number of samples for each cell using the first event-based sampling method

Cell name	W3	W4	W5	W7	W8	W9	W10	G1	V4	V5
Number of samples	3	9	13	4	14	13	13	9	9	3

3.2.3 Event-based sampling method 2 / Hybrid

Besides the first event-based sampling method previously described, a second event-based sampling method was investigated. This method was based on the change in capacity within fixed time intervals for the cell. The basic idea of the method builds on an interpolation of the capacity values from the RPTs of the cells. Using the interpolated graph of the capacity over time the capacity change was investigated for fixed time intervals by calculating the percentage difference within each specific interval, as illustrated in Figure 2.7. The number of samples for different values of the time intervals (ΔT) and the capacity change (ΔQ) were tested and are presented in Figure 3.3.

The first time interval investigated was $\Delta T = 13.1h$ which corresponds to the average time of one cycle based on all cells. The spread of samples for different values of ΔQ using this time interval is illustrated in Figure 3.3a. As can be seen, the cells had widely different values of ΔQ to reach the same number of samples. The ΔQ that triggers the samples is measured in percentage. The only clustering of the number of samples can be seen with ΔQ around 0.040% - 0.0415% where 5 cells gather with about 20 samples each. For the rest of ΔQ the cells are more spread out. This spread of values complicated the choice of ΔQ since no value would give a reasonable number of samples for all cells.

To solve this problem, samples were also investigated for $\Delta T = 18.3h$, corresponding to the average time for one cycle based on the cells going through all 15 RPTs. The number of samples for different ΔQ using $\Delta T = 18.3h$ is illustrated in Figure 3.3b. As can be seen in the figure, this value of ΔT resulted in a slightly more compact clustering than $\Delta T = 13.1h$ with 6 cells but still with a quite high number of samples around 17.

Finally, the number of samples was also investigated for $\Delta T = 24h$, representing one day. This distribution is illustrated in Figure 3.3c. As can be seen in the figure, the clustering for this ΔT appears at a lower number of samples, around 12.

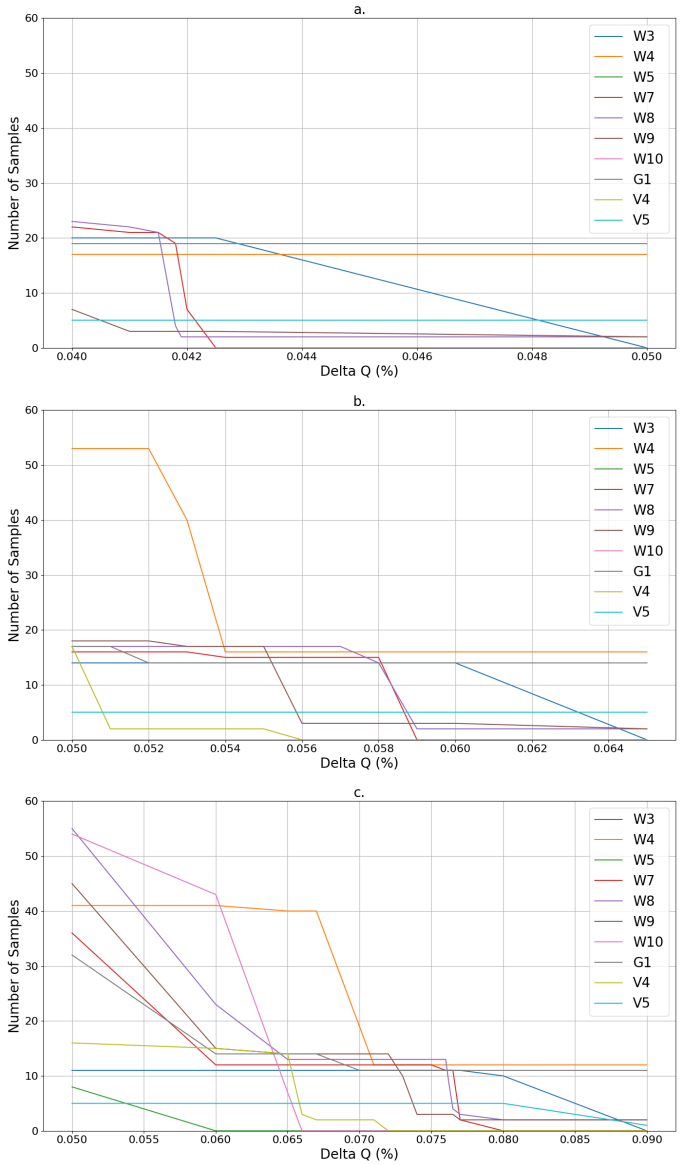

Figure 3.3: a. Number of samples for different ΔQ using ΔT=13.1h b. Number of samples for different ΔQ using $\Delta T = 18.3h$ c. Number of samples for different ΔQ using $\Delta T = 24h$.

Based on these results, the fixed time intervals were chosen to $\Delta T = 24h$ together with a capacity change threshold value of 0.0762 %. These parameters result in the number of samples for each cell that is shown in Table 3.3.

Table 3.3: Number of samples using the chosen fixed time interval and capacity change threshold, only using the second event-based sampling method

Cell name	W3	W4	W5	W7	W8	W9	W10	G1	V4	V5
Number of samples	11	12	0	11	9	3	0	11	0	5

As can be seen in Table 3.3, some cells were never triggered. This means that the data of these cells wouldn't be used for validation of the sampling method and that an estimation of the capacity degradation of these cells wouldn't be possible. Because of this, the sampling method was made hybrid by adding a sparse periodic sampling to ensure data for all cells were captured.

Turning the sampling method hybrid would also be positive since it makes it useful for a wider range of behaviors of the batteries. If the threshold of the event-based sampling were to never be met, battery degradation could continue until EOL without any samples being triggered. Adding a sparse periodic sampling ensures that the sampling was triggered at some point even if the event-based threshold was never met.

If a trigger by the capacity degradation didn't occur, the degradation of the battery was slow which meant that the periodic sampling could be sparse without losing any critical data. Different values for the periodic sampling of each cell were investigated and can be seen in Figure 3.4. As can be seen in the figure, there is no obvious clustering where the cells have a similar number of samples for the same value of T. Because of this, an arbitrary value of T was selected where all cells had a reasonable number of samples. The interval of the periodic sampling was set to $T = 3.5 \cdot 10^6$s or $T = 5.5$ weeks. If the sampling was triggered by capacity degradation, the countdown for the periodic part was reset, meaning that it wasn't executed if not needed. In addition to the cycles extracted based on the sampling criteria, the second cycle of the cell was extracted to get information on the cell's initial state. The new number of samples using the hybrid sampling method is shown in Table 3.4.

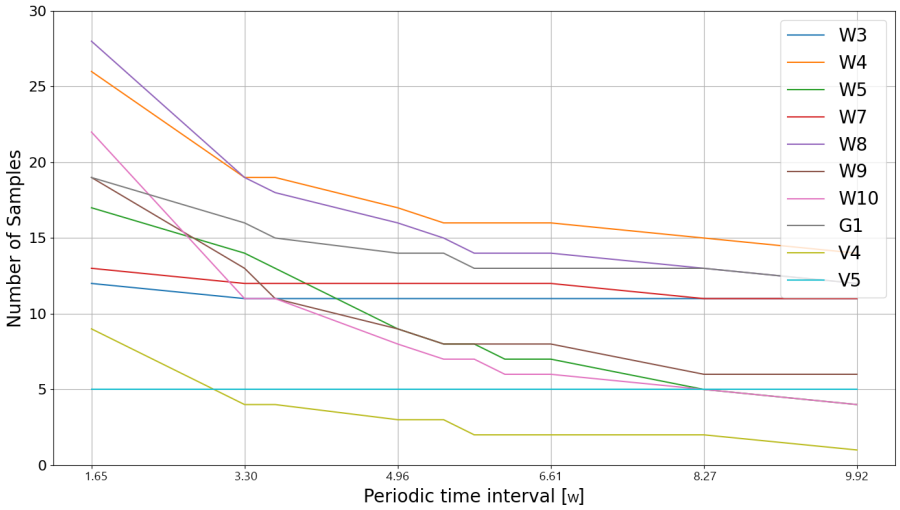

Figure 3.4: Number of samples for different periodic sampling times using T = 24h

Table 3.4: Number of samples for each cell using the hybrid sampling method

Cell name	W3	W4	W5	W7	W8	W9	W10	G1	V4	V5
Number of samples	11	16	8	12	15	9	7	13	2	5

From now on this sampling method will be called the **hybrid** sampling method and the purely event-based method in Section 3.2.2 is now referred to as the **event-based** method.

3.3 Feature engineering

To train a GPR model with accurate SOH prediction, appropriate features need to be extracted and selected as inputs. A feature is a measurable property showing a correlation to the model's output and depend on the measurements available in the dataset. In this case, features were chosen to show a strong correlation to the capacity of the battery. Since the hybrid and event-based sampling methods extracted full cycles when triggered, the features was chosen to only depend on one cycle to be applicable to all three sampling methods.

To find appropriate features that would give the model a clear indication of the capacity degradation of the battery, several features were extracted and investigated. Since not all features would contribute to efficient training and testing of the model, a feature selection was executed which will be explained further in Section 3.3.6. A brief introduction to each feature investigated will follow as well as how that feature was implemented.

3.3.1 Characteristics of voltage curve during charging mode

Four features from the voltage curve were extracted during charging. The voltage curve together with its' corresponding step index for one whole cycle is illustrated in Figure 3.5. The features selected from the data had step indices 9 and 10 to match the desired durations of the cycles, constant current (CC) charging and constant voltage (CV) charging mode. These features have been previously implemented with a GPR for SOH estimation by Yang *et al.* [27] using a dynamic drive cycle and are proven to give high accuracy and robustness to the model . As the battery ages, the characteristics of the charging curves will change, making them useful indicators of the SOH. The details of the four features extracted from the charging curves are explained below and illustrated in Figure 3.6. For information about how the voltage curves behave over time for all cells, see Appendix A.

- Feature 1 (F1): Time of CC mode. As the number of cycles increases, the charging time in CC mode is expected to decrease because of a decreased capacity and increased impedance. The increased impedance gives rise to higher polarisation and can be observed in Figure 3.6 where one can see a shorter CC mode duration the later the phases are. This phenomenon is significant for this specific cell but has varying significance for the different cells in the data set. Because of this, the time of CC mode is extracted as a feature.

- Feature 2 (F2): Time of CV mode. As time spent in CC mode decreases, the corresponding time in CV will increase to fully charge the battery. The time for the battery to be fully charged in CV mode is a result of the ageing of the battery and can be seen as an inverse indicator of the battery's SOH. Because of this, the time of CV mode is extracted as a feature.

- Feature 3 (F3): The slope of the voltage curve at the end of CC charging mode. As can be seen in Figure 3.6, the end slope of the CC charging mode seems to be stable for each individual phase. This means the slope can be extracted through $F3 = \frac{dV}{dt} \approx \frac{\Delta V}{\Delta t}$. For a fixed sample interval, this means that F3 will be directly proportional to ΔV and, being extracted from the CC mode, also directly proportional to $\frac{\Delta V}{\Delta t}$. This way F3 can be an indicator of the internal resistance, which is closely associated with the SOH. The slope is based on the 5 last data points for the CC charging mode. Because of this, the end slope of CC mode is extracted as a feature.

- Feature 4 (F4): The vertical slope at the corner of the CC charging curve. Figure 3.6 shows how the voltage increases fast at the beginning of the CC charging phase to at a certain point start increasing slower. The shape of this corner changes as the battery ages. For later cycles, the curve tangent slope seems to increase, which means the perpendicular slope will have a decreasing trend related to the battery aging which can be used as an SOH indicator in the model. Because of this, the perpendicular slope at the corner of the voltage curve is extracted as a feature.

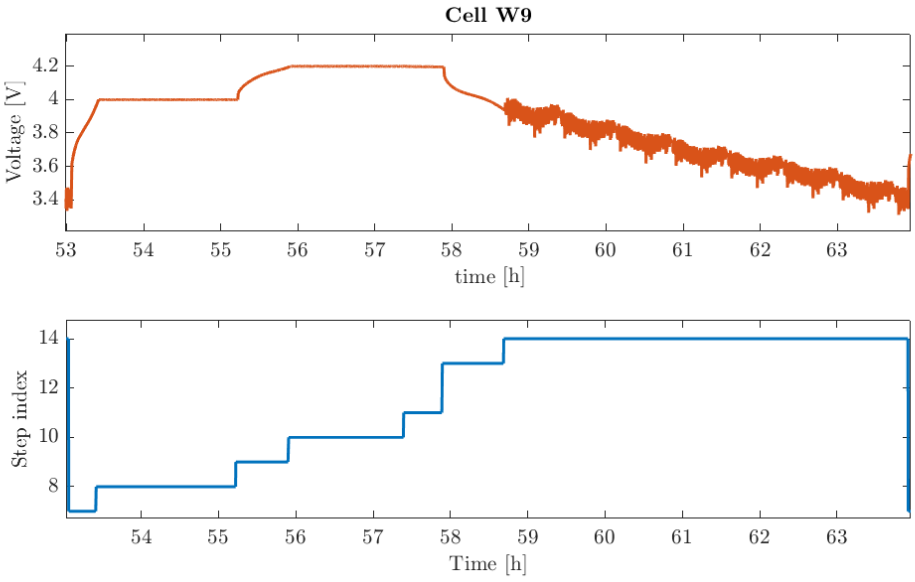

Figure 3.5: Voltage curve (orange) and step index(blue) for one cycle of cell W9

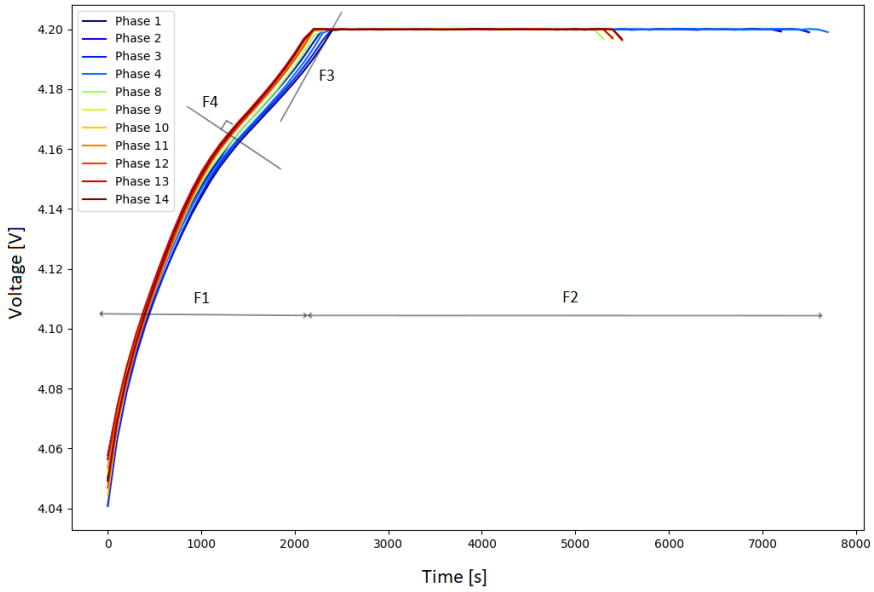

Figure 3.6: Ageing voltage curve for CC and CV mode of cell W9 with features F1-F4 illustrated

3.3.2 Depth of discharge

To quantize the charge level in a battery one can express it with the SOC level, meaning that the SOC level is 100% when the battery is fully charged. The charge level can also be explained using the term depth of discharge (DoD). The DoD is the opposite of the SOC meaning that the cell has 100% DoD if fully discharged. According to [28], the definition of the DoD is

$$\text{DoD} = \frac{Q_{current}}{Q_{init}} \cdot 100\% \tag{3.1}$$

where $Q_{current}$ is the current capacity and Q_{init} is the inital capacity of that cycle calculated using equation 2.1.

The relation between DoD and the health of a LIB has been investigated by Khizbullin *et al.* [29]. They conlcluded that DoD has a strong relation to the degradation of the battery since it indicates capacity depletion in relation to the nominal capacity. They show in their work that going from 100% to 50% DoD during usage, the number of cycles performed before EOL is reached could increase by 220% which shows a strong relation between the SOH and DoD of the battery.

During usage of the battery, it will undergo degradation and the internal structure will start to change. According to Chang *et al.* [30] the degradation is caused by factors such as the thickening of the solid electrolyte interface layer, lithium plating and dendrite growth on the anode surface. Varying the DoD can accelerate these factors, causing the battery to reach EOL at an earlier stage.

The maximum DoD of a cycle will be considered a feature for the ML model and will be referred to as Feature 5 (F5). The initial capacity was approximated to the nominal capacity value of 4.85 Ah [16]. Further, the battery was assumed to be fully charged for all cycles and no faults were taken into consideration.

3.3.3 Time

Another feature extracted from the dataset, further referred to as Feature 6 (F6), was time. During the time the cells undergoes the cycling and RPT's, their capacity degrades. Since the batteries in the dataset was constantly cycled, the time in this case will be strongly correlated to the energy throughput. Because of this, the relation between time and capacity was investigated in the same way as the previously described features, namely one value was extracted for each cycle performed. The value chosen to be extracted was the starting time for the cycle.

3.3.4 Histogram-based voltage and current windows

According to Greenbank and Howey [31], using histogram-based features indicating the time spent within certain voltage- and current intervals shows good results when estimating the SOH for a LIB using a GPR. Because of this, these features were investigated in this work.

The different voltage- and current windows chosen to be investigated were spanning between different thresholds represented by percentiles of the voltage. The thresholds were chosen using Figure 3.7 and are shown in Table 3.5. The voltage and current for the whole cycles during the test were taken into consideration when extracting the features meaning that data were extracted from both the charging and the discharging phase.

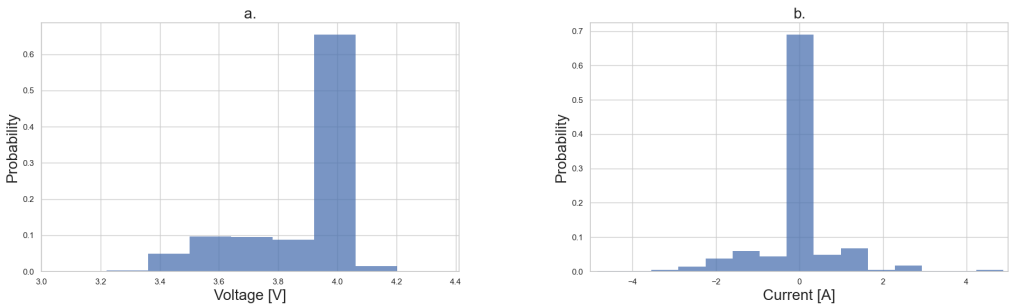

Figure 3.7: a. Probability distribution of time spent within voltage windows for all cells b. Probability distribution of time spent within current windows for all cells

Table 3.5: Voltage- and current limits for the used percentiles

Percentile	**Voltage [V]**	**Current [A]**
1^{st}	3.20	-2.62
5^{th}	3.51	-1.66
30^{th}	3.91	0.00
90^{th}	4.00	0.89
99^{th}	4.20	2.51

All combinations of these thresholds were investigated where the value of the features was represented by the time spent within the specified range. The features extracted are given in Table 3.6 and 3.7.

Table 3.6: Features extracted from Voltage ranges

Feature Name	**Percentile limits**	**Voltage Range [V]**
$F7$	1^{st} to 5^{th}	$3.20 < \mathrm{V} < 3.51$
$F8$	1^{st} to 30^{th}	$3.20 < \mathrm{V} < 3.91$
$F9$	1^{st} to 90^{th}	$3.20 < \mathrm{V} < 4.00$
$F10$	1^{st} to 99^{th}	$3.20 < \mathrm{V} < 4.20$
$F11$	5^{st} to 30^{th}	$3.51 < \mathrm{V} < 3.91$
$F12$	5^{st} to 90^{th}	$3.51 < \mathrm{V} < 4.00$
$F13$	5^{st} to 99^{th}	$3.51 < \mathrm{V} < 4.20$
$F14$	30^{st} to 90^{th}	$3.91 < \mathrm{V} < 4.00$
$F15$	30^{st} to 99^{th}	$3.91 < \mathrm{V} < 4.20$
$F16$	90^{st} to 99^{th}	$4.00 < \mathrm{V} < 4.20$

Table 3.7: Features extracted from Current ranges

Feature Name	**Percentile limits**	**Current Range [A]**
$F17$	1^{st} to 5^{th}	$-2.62 < \mathrm{A} < -1.66$
$F18$	1^{st} to 30^{th}	$-2.62 < \mathrm{A} < 0.00$
$F19$	1^{st} to 90^{th}	$-2.62 < \mathrm{A} < 0.89$
$F20$	1^{st} to 99^{th}	$-2.62 < \mathrm{A} < 2.51$
$F21$	5^{st} to 30^{th}	$-1.66 < \mathrm{A} < 0.00$
$F22$	5^{st} to 90^{th}	$-1.66 < \mathrm{A} < 0.89$
$F23$	5^{st} to 99^{th}	$-1.66 < \mathrm{A} < 2.51$
$F24$	30^{st} to 90^{th}	$0.00 < \mathrm{A} < 0.89$
$F25$	30^{st} to 99^{th}	$0.00 < \mathrm{A} < 2.51$
$F26$	90^{st} to 99^{th}	$0.89 < \mathrm{A} < 2.51$

3.3.5 Incremental Capacity Analysis

Incremental Capacity Analysis (ICA) is a technique that can be used to evaluate a battery's characteristics and has shown to be a promising feature that captures a battery's capacity loss [11],[32]. An incremental capacity curve is calculated per cycle by,

$$IC = \frac{dQ}{dV}, \tag{3.2}$$

where Q is the capacity and V is the voltage. ICA requires a constant current, which limits the curve to either be derived from step 1 or step 3 of the cycle (as seen in Figure 2.8). Step 3 was chosen due to having the same c-rate between cells, compared to step 1 where the rate varies between C/4 and 3C depending on the cell. During this chosen step, the voltage goes from 4 V to 4.2 V at a constant current of C/4 which limits the ICA to this voltage range and one peak. Using ICA on the RPTs, which enables analysis for the entire operational voltage range of the cell, more peaks were detected. From this, it was concluded that the peak in the range of 4-4.2V was the most significant. Due to the nature of ICA's definition it will enhance the noise in the collected data. To reduce the noise and make sure that the features could be extracted a Savitzky–Golay filter was implemented before features were extracted [33]. In the ICA in the range of 4 - 4.2V (step 3), three features were extracted: peak height (F27), peak position (F28), and peak area (F29), and are illustrated in 3.8. The peak area was chosen to include the 30% highest values.

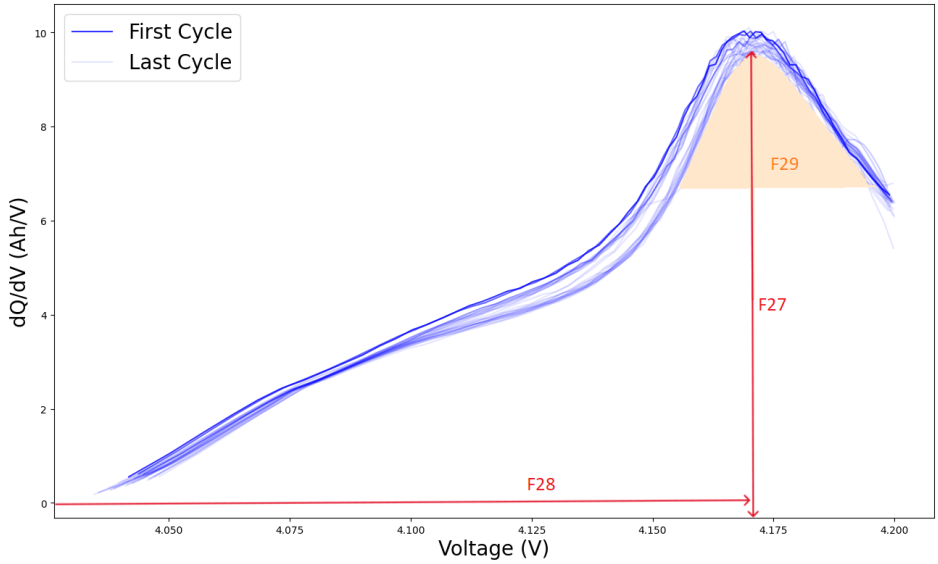

Figure 3.8: IC curve from cell W9 after being filtered with features F27-F29 illustrated

3.3.6 Feature selection

In order to make the model efficient, all features may not be necessary for an accurate prediction. A model based on less features will have better computational efficiency than one with many features. Hence, a selection of the extracted features was performed.

There are many options when analyzing the correlation of two features. In this case, it was done using the absolute value of Pearson's correlation coefficient

$$S_{i,j}(f_i, f_j) = \left\| \frac{\text{cov}(f_i, f_j)}{\sigma(f_i)\sigma(f_j)} \right\| \tag{3.3}$$

where f_i and f_j represent the features and σ is the standard deviation [31]. The correlations of the features were then visualized in correlation matrices.

When the correlations between the different features were visualized, the selection could be executed. The feature having the strongest correlation to the capacity were identified and marked as a strong feature. The other features having a stronger correlation than 0.85 to this feature were then removed in order to make sure not to feed the model with redundant data. After the removal, the feature having the second strongest correlation to the capacity was identified and the same procedure was performed on this one. This was done until no features had a correlation stronger than 0.85 to any other and all remaining features were numbered according to their correlation to the capacity.

Since there initially was a large number of features, the features were divided into different matrices where the histogram-based voltage and current windows features were compared separately to each other. After making a selection of these features, they were integrated into a single matrix including all remaining features. Based on this matrix a final selection was done.

For future reference, a summary of all extracted features is documented in Table 3.8 together with their notation.

Table 3.8: Names of features extracted for future reference

Feature	Feature name
Time of CC mode	F1
Time of CV mode	F2
slope of voltage curve at the end of CC mode	F3
Perpendicular slope at the corner of voltage curve (CC mode)	F4
DoD	F5
Time of cycle start	F6
Histogram-based voltage windows	F7-F16
Histogram-based current windows	F17-F26
Peak from ICA	F27
Position from ICA	F28
Area from ICA	F29

3.4 ML Model

The ML model chosen for this project was a GP model. It is commonly used for supervised learning tasks such as regression and gives a clear confidence score that makes the output more representative of the real world. Examples of its use are Richardson *et al.* [21] to accurately predict long-term capacity fade for LIBs and Yang *et al.* [27] that successfully used a GPR for SOH estimation with a battery under dynamic load.

Before the now extracted features could be used for ML they needed to be further preprocessed. Firstly the capacity was normalised to each cell's initial value so that all cells started at 100%. The next step was normalising the features of the model using Min-max normalisation to cover the range of 0-1. The last step was to sort the data according to time. With the data prepared it was divided into a test set containing the data for one cell, W8 in our case, and a training set containing the remaining cells. Lastly, both of the sets were divided into inputs, $\mathbf{X}$, and the corresponding output, $\mathbf{y}$.

Next, the GPR model needed to be set up. The kernel chosen was a Matérn kernel combined with a white noise kernel to account for small inaccuracies in the data set. This was chosen to get the general downward trend together with the less smoothed-out results compared to the commonly used RBF kernel. The hyperparameter for the Matérn kernel was set to 1 for l with bounds of 0.01 and 100 and ν was set to 2.5 to still have some smoothness. Finally, the GP was configured to not normalize the output as this was done manually to get the final results in percentages of capacity fade.

With everything set up, the next step was to run through all the cases. Firstly, the model was trained once with time and then once for each feature set using the training data. This was to get one model per feature set that understood the relation between the feature set and the capacity. The cases tested are shown in Table 3.9 and the features included in each feture set is presented in Table 4.1 in Chapter 4. By only changing one parameter at a time, each change could be validated by the new prediction. The next step was using the trained model with the test data for each sampling method to gather all the prediction results. As a GPR outputs a normalized value, the predictions had their normalizations reversed to get the percentage capacity fade from the initial value. Finally, each scenario had its RMSE and coverage probability calculated, i.e. how many of the test point lie withing the 95% confidence interval.

Table 3.9: The cases' combinations of features set and methods

Case	**Sampling method**	**Feature set**
1 / Baseline	Periodic	Time
2	Periodic	1
3	Periodic	2
4	Periodic	3
5	Periodic	4
6	Event-based	Time
7	Event-based	1
8	Event-based	2
9	Event-based	3
10	Event-based	4
11	Hybrid	Time
12	Hybrid	1
13	Hybrid	2
14	Hybrid	3
15	Hybrid	4

4
Results

Using the methodology described in detail in Section 3, the feature selection and prediction outputs will be detailed here. They will be analyzed and discussed individually, where the most notable and/or best results are highlighted. Due to a lot of the results not being directly relevant to the overall conclusion, they will be documented in appendices.

4.1 Feature selection

For the feature selection, the features were selected in different steps because of the large amount of features in the initial set. For the first step, only the correlation between the capacity and the histogram-based features was investigated. These correlation matrices can be found in Appendix B (Figure B.1 and B.2). A large amount of these features were strongly correlated to each other resulting in some of them being removed in the feature selection process, as described in Section 3.3. After this removal of strongly correlated features, the remaining features were included in a correlation matrix with all other features to compare their correlation to the capacity. This correlation matrix can also be found in Appendix B (Figure B.3).
From the matrix, the features were numbered based on their correlation to the capacity. These features together with their correlations are shown in Figure 4.1. The features showing the strongest correlation were selected for the feature sets used to train the ML model as shown in Table 4.1.

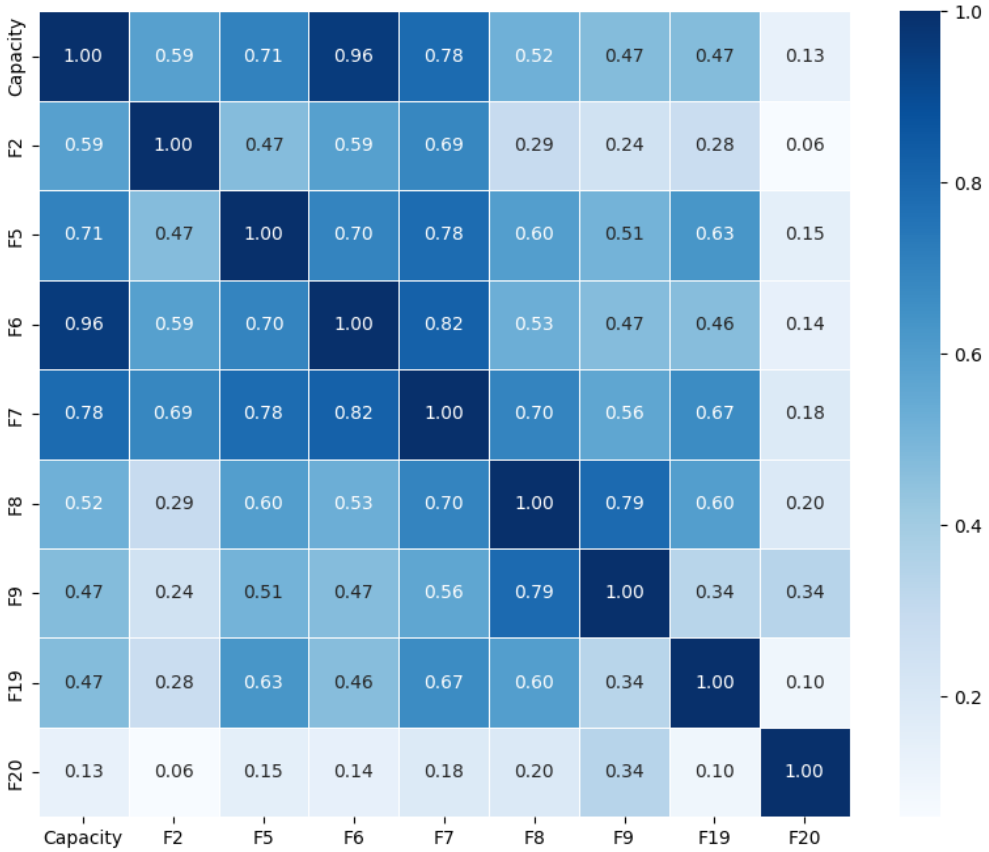

Figure 4.1: Correlation matrix for the strongest correlated features using the training data. Values created the basis for feature selection.

Table 4.1: Feature sets used to train the ML model

Feature set	Features
1	F6, F7, F5
2	F6, F7, F5, F2, F8, F20
3	F6, F7, F5, F2, F8, F20, F9, F19
4 [1]	F6

To validate the selected features for the training data, the correlation to the capacity was also visualized for the test data using the periodic sampling method. This is shown in Figure 4.2.

[1]Because the other feature sets showed unexpected results when running the model, a fourth feature set was added. Feature set 4 was only based on the most correlated feature, F6

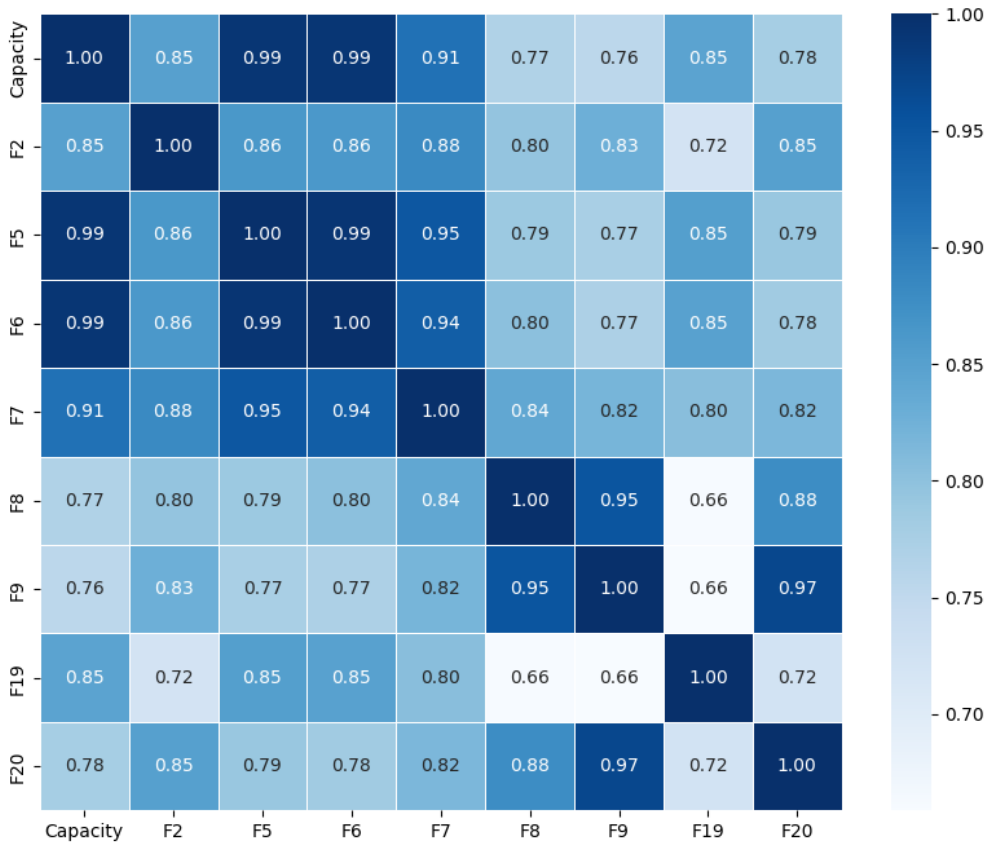

Figure 4.2: Correlation matrix for the strongest correlated features using the test data.

4.2 Prediction

In the following section, the performance of the model using the different sampling methods will be showcased. The plotted predictions of the best-performing case for each sampling method will be shown together with some notable results. For plots of all cases defined in Table 3.9, see Appendix C.

4.2.1 Periodic sampling method

When running the developed model for the different cases using the periodic sampling method, it resulted in the amount of data and validation metrics shown in Table 4.2. The first setup investigated was the baseline, based on the periodic sampling using all time-data to train and test the model. The plotted baseline prediction can be seen in Figure 4.3. Comparing the performance of the cases using different feature sets with the baseline, Case 5 was considered the best-performing setup of the model using the periodic sampling method. The plotted prediction of Case 5 can be seen in Figure 4.4. As seen in Table 4.2, the amount of data used for training and testing the model is significantly reduced for Case 5 compared to the baseline.

Table 4.2: Result metrics of the periodic cases

Case	Train [kB]	Test [kB]	RMSE (%)	Coverage Probability (%)
1 / Baseline	21639.9	203.93	0.54057	98.50
2	121.599	24.152	0.49568	49.70
3	216.963	43.051	0.55419	68.64
4	281.397	55.760	0.68103	60.36
5	73.725	14.670	0.27326	97.04

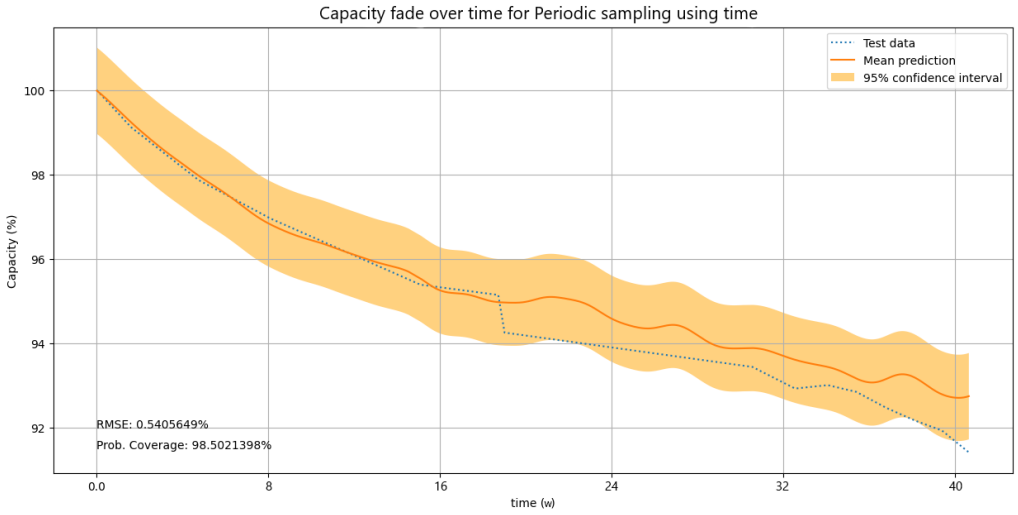

Figure 4.3: Prediction of capacity loss from initial capacity for the baseline case.

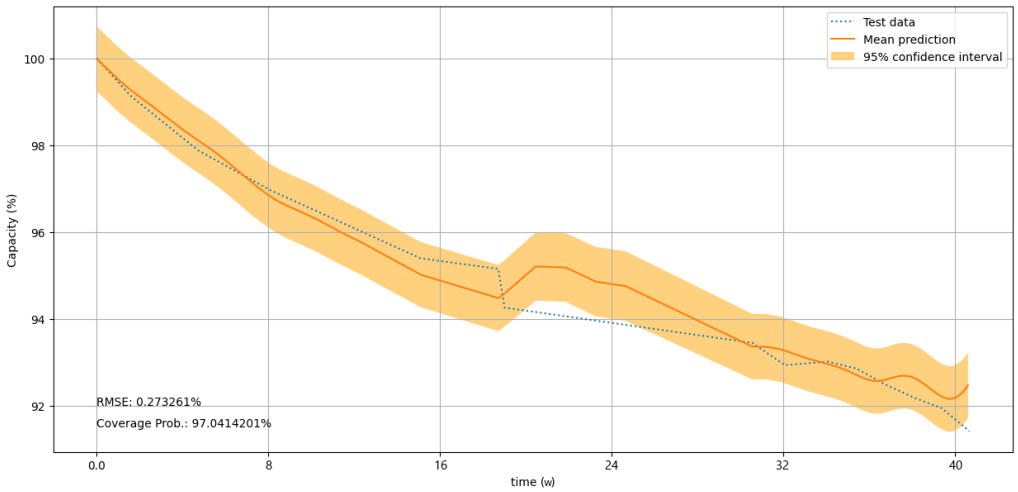

Figure 4.4: Prediction of capacity loss from initial capacity for Case 5.

The periodic sampling method has the advantage of a higher resolution for all the feature sets compared to the event-based and hybrid sampling methods since all cycles are included. The higher resolution increases the probability of a clearer trend

of the feature's behaviour but doesn't necessarily mean a more accurate or reliable result. A significant drop compared to the baseline was observed when looking at coverage probability for Cases 2-4 that uses Feature sets 1-3. The most significant case was for Case 2 with a coverage probability of only 49.70% compared to the 98.50% for the baseline. Only Case 5 stands out with a coverage probability equivalent to the baselines.

Looking at the RMSE of the cases presented in Table 4.2, Cases 2-4 have an RMSE comparable to the baseline while Case 5 has an RMSE 49.45% lower than the baseline. Case 5 also has a large reduction of the data, which would make it the best option for the periodic sampling method. Case 4 is an outlier in that it resulted in an increase in RMSE of 25.98%, showing that this feature set isn't suitable for the periodic sampling method.

4.2.2 Event-based sampling method

For the Event-based sampling method, cases using equivalent features as in the periodic case were tested. The results are shown in Table 4.3. For this sampling method, Case 10 was considered the best-performing setup, and the prediction of this case can be seen in Figure 4.5. Also in this case a remarkable reduction of the train and test data can be seen compared to the baseline presented in Section 4.2.1.

Table 4.3: Result metrics of the event-based cases

Case	Train [kB]	Test [kB]	RMSE (%)	Coverage Probability (%)
6	21639.9	122.3	0.45789	99.04
7	121.599	1.177	0.59783	66.67
8	216.963	2.009	0.73423	86.67
9	281.397	2.640	0.75245	86.67
10	73.725	0.739	0.39294	93.33

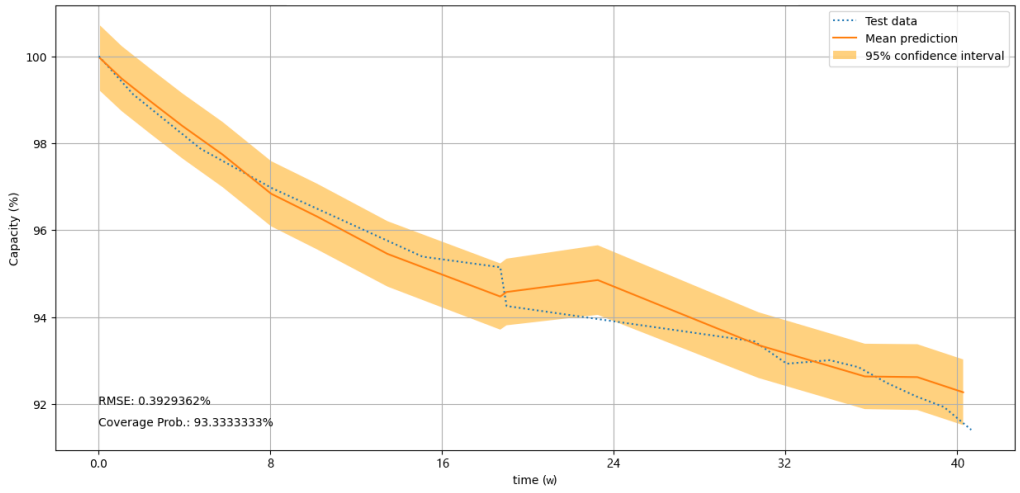

Figure 4.5: Prediction of capacity loss from initial capacity for Case 10.

Looking at these results, a large variation in the RMSE can be noted for the different feature sets. For Case 6, using time as input for the ML model, an RMSE decrease of 15.29% compared to the baseline was noted. This result validates the use of the sampling method, showing that the sampling method itself gives as accurate estimations of the SOH as the method used today. This is without using features to improve the model.

The RMSE for Cases 8 and 9 had a sharp increase of 35.0% and 36.6% respectively, which shows the weakness of the features used. An increase, although not as sharp, is also noted for Case 7 of 10.6% relative to the baseline. The last case, Case 10, was the only case using features that resulted in a better RMSE value than the baseline, with a decrease of 27.3%. Because of this, Case 10 can be considered the best result using the event-based sampling method.

For Case 9, the confidence interval of the prediction took on a shape deviating from the shape of the previously investigated cases. The prediction plot of Case 9 is therefore shown in Figure 4.6. Notably, Case 8 had a result similar to the result of Case 9. These cases had a large coverage probability and a high RMSE compared to the other cases observed. This can partly be explained by the fact that the cycles during the middle of the dataset are significantly more drawn out than those at the start and the end, something that remains unresolved. Detailed information about the dataset isn't yet available and our analysis of the data yielded no explanation for the behaviour. On the other hand, this increase in uncertainty should then also be noticeable when using Feature set 4 such as Case 10 in Figure 4.5 but is nowhere to be found. It is therefore not certain that this drawn-out middle section is the reason behind the large uncertainty. The large RMSE values for these cases hint that Feature sets 2 and 3 don't give as good performance as their correlations would suggest, at least not for the event-based method.

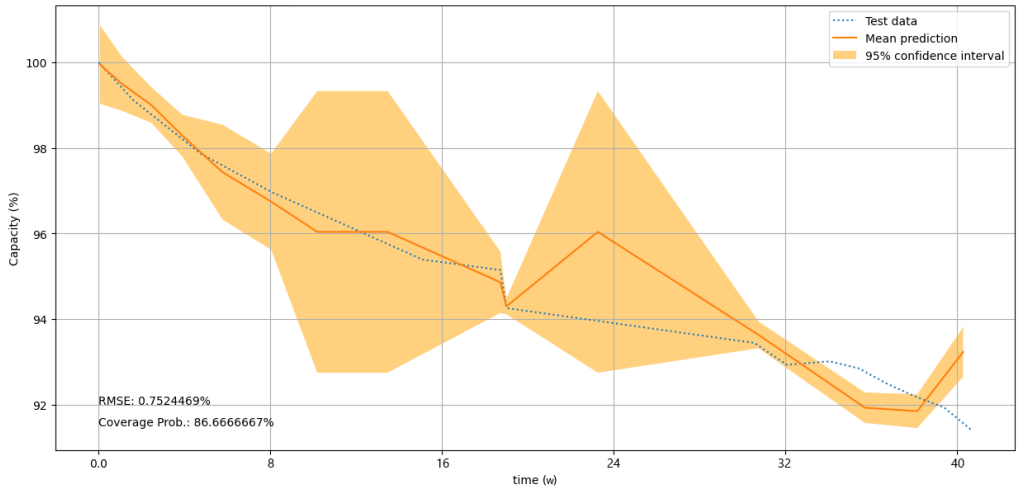

Figure 4.6: Prediction of capacity loss from initial capacity for Case 9.

4.2.3 Hybrid sampling method

The results for the cases using the hybrid sampling method are presented in Table 4.4. Considering the values of the RMSEs, Cases 12 and 15 were considered the best-performing setups for this sampling method. The predictions of these cases are shown in Figure 4.7 and Figure 4.8 respectively.

Table 4.4: Result metrics of the hybrid cases

Case	Train [kB]	Test [kB]	RMSE (%)	Coverage Probability (%)
11	21639.9	256.8	0.55915	100.0
12	121.599	1.245	0.29580	62.50
13	216.963	2.117	0.72997	93.75
14	281.397	2.775	0.73859	93.75
15	73.725	0.779	0.31855	93.75

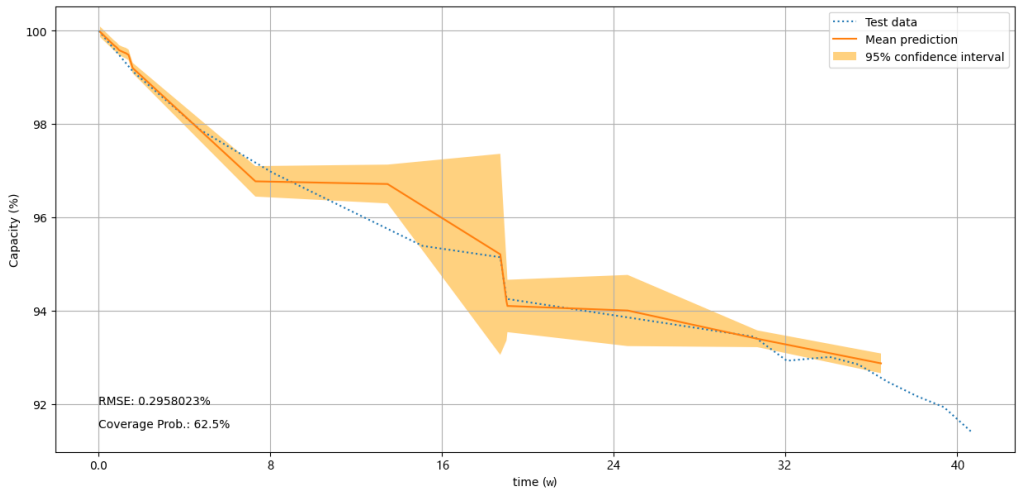

Figure 4.7: Prediction of capacity loss from initial capacity for Case 12.

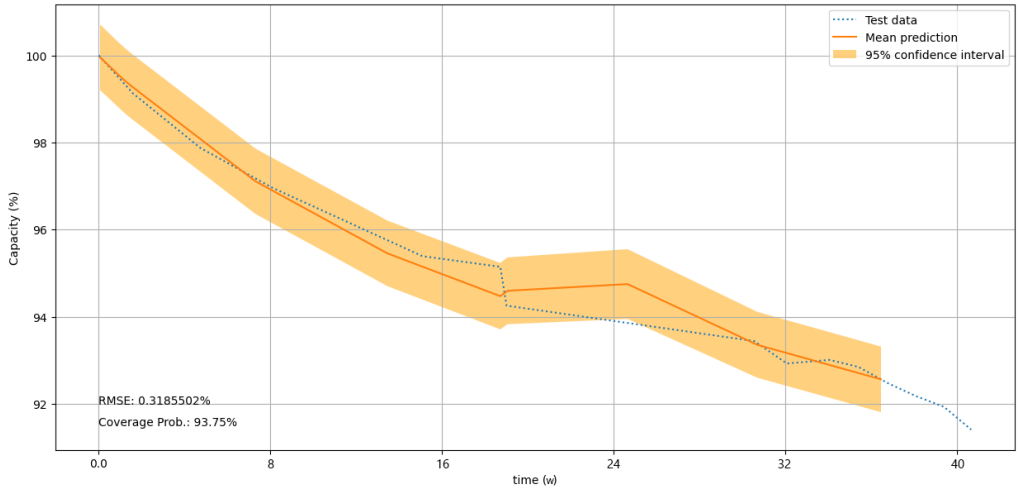

Figure 4.8: Prediction of capacity loss from initial capacity for Case 15.

This sampling method's results show promise but can be challenging to interpret since the metrics point towards different conclusions. For Case 11, only using time as input, the RMSE shows a result slightly higher than that of the baseline. This means that the hybrid sampling method would need features to improve the performance of the ML model to achieve the same accuracy as the baseline.

Similarly to Case 9 previously shown, Case 14 showed a confidence interval of the prediction that took on a shape deviating from the ones previously seen together with a high RMSE. The prediction plot of Case 14 is shown in Figure 4.9. Notably, Case 13 had a result similar to the result of Case 14. The reasoning for these deviating results would be the same as for Case 9 and 8, and a conclusion could be drawn that feature sets 2 and 3 do not give as good performance as their correlations would suggest also for the hybrid sampling method.

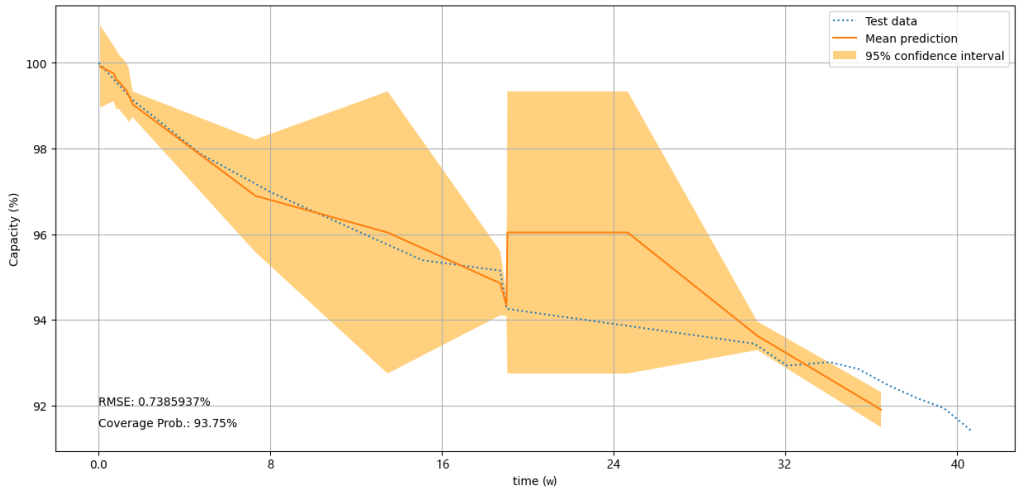

Figure 4.9: Prediction of capacity loss from initial capacity for Case 14.

Some of the other results from the hybrid sampling showed more promise. For Case 15, a high coverage probability was noted but with an RMSE showing a decrease of 41.07% compared to the baseline. Another case having significantly better RMSE is Case 12. This is the case with the lowest RMSE using a new sampling method with a decrease of 45.28% compared to the baseline. It is also the second best RMSE out of all cases investigated, after Case 5 using the periodic sampling method. However, the coverage probability for Case 12 is notable with a low value of 62.50%. The low coverage probability doesn't change the fact that this case is the one with the lowest error but indicates that the ML model is overconfident when predicting the interval of the prediction. Considering these three cases, Case 5, 12, and 15, one can argue that Case 12 would be the most sufficient option because of the previous discussion about only using time as in Feature set 4. Cases 15 and 5 are only based on F6, meaning that the prediction is only based on time. By using more features, as in Case 12, the ML model will get information on events happening during usage that may affect capacity degradation. This is probably why we see a better RMSE for Case 12 and would mean that Case 12 would be a more reliable prediction in actual usage. Therefore this case is considered the best result out of all the cases.

4.3 Discussion of related matters

Relevant factors that might affect the results and/or conclusion of the paper are discussed below. First is analysing the features and their implementation, then talking about the general data reduction and lastly how the chosen dataset has limited the model and work.

4.3.1 Features

As can be seen in Figure 4.1 showing the correlation matrix for the features used in the feature sets, the chosen features seem to have quite high correlation to the capacity. This indicates that the features would give a good result when training the model to predict capacity degradation. Surprisingly, the results in Section 4.2.1 show that the majority of the features show a worse result for the prediction than when running the model only using time as the input. This shows that the features may need some improvement to give good results but can also indicate that the data in some cases may have some faults. For example, it was noted that for some cells, there were periods where the cycles were prolonged compared to the average cycle which may affect some of the time-dependent features negatively.

To verify the features chosen, a correlation matrix of the features and capacity using the test data was extracted and shown in Figure 4.2. As can be seen in the figure, the features had a high correlation to the capacity also for the test data which would strengthen the argument of the features, making them useful also for testing of the model. For the feature selection, features having a correlation to each other stronger than 0.85 were removed to avoid training the model with redundant data. As can be seen in Figure 4.2, the removal of strongly correlated features didn't count for the

test data in this specific case, leading to strong correlations between features. This does not deteriorate the results but may cause usage of redundant data that will not increase the accuracy of the model and thereby unnecessarily increase the data usage.

When extracting features F1-F4, only the durations of step indices 9 and 10 were taken into consideration. As previously described, each cycle included two CC charging durations and two CV charging durations. For the first CC duration, the cells had varying C-rates. To make sure the features were comparable, the duration instead chosen was the second CC duration where all cells were charged under the same conditions. As a consequence of only the second CC duration being used to extract F1, only the second CV mode was used for extraction of F2 due to the relationships between them. With that said, the first CC and CV durations might have similar trends and could be investigated as features for future work despite the varying conditions between the cells.

Other than the selection of duration for the extraction of F1-F4, F3 and F4 have other factors affecting their correlation to the capacity. Since these features are represented by slopes at specific locations on the voltage curve, the resolution of the data is crucial to get a value representing the shape correctly. Since the data is down-sampled in the preprocessing for the event-based and hybrid sampling methods and even further in the periodic sampling method, the accuracy of the slope may be affected. This means a smaller downsampling factor in the preprocessing or periodic sampling could increase the capacity correlation of these features. On the other hand, a smaller down-sampling factor would increase the amount of data used and therefore require more computational resources.

From Figure 4.1 it can be deduced that F6 is the feature showing the strongest correlation to the capacity. One can also conclude from Section 4.2 that Feature set 4, only including F6, shows the best result for both the periodic and event-based sampling method. From these results, one can argue that F6 is the best feature extracted. The drawback of using F6 as the only feature, Feature set 4, is that it only contains information about time. This means that it does not give the ML model any information of the actual behaviour of the cell and deviating events are not detected or taken into consideration when predicting the capacity degradation. Considering this drawback of F6 it can be concluded that it is a good feature showing a strong correlation to the capacity but that it should preferably be used in combination with another feature able to detect unexpected events during usage.

When analyzing the capacity correlations of the window-based features F7-F26, we can see that F7 has the highest correlation. F8, F9, F19 and F20 were also included in the feature sets but showed a notably lower correlation to the capacity than F7. The remaining window-based features were either too correlated to one of the selected features or had too low correlation to the capacity to be chosen for one of the feature sets. In an attempt to increase the number of useful features, one could adjust the thresholds of the windows. This way the correlations between the features may decrease or the correlation to the capacity may increase, leading to

more useful features.

Figure 4.1 indicates that the ICA-based features F27, F28 and F29 show a low correlation to the capacity. This is an unexpected result since other works state a strong indication of the capacity degradation of these features. The reason for the weak results using these features may be because of noise in the data or a fault in its implementation. To improve these features, more or other filters could be investigated together with an even higher resolution of the data to clarify the trend of the features related to the capacity fade. Another possible reason for this unexpected result may be the high C-rates used in the dataset. Other works using this method to extract features use lower C-rates for the data which may cause clearer indications of the capacity degradation.

4.3.2 Decrease of data

In Section 4.2 it was noted that by using the proposed sampling methods and feature sets, the sizes of the train data and test data were significantly reduced. The reduction of both these data sets is beneficial since they affect the usage of our proposed system. However, when comparing the different cases, the reduction of the test data was considered the most important factor. This is because the train data only has to be used once to train the model, while the test data is needed every time inference with the model occurs. Therefore, the size of the test data is the data required to run the model. If put in a real-life scenario, this means that the training is done once when developing the system into a product and might then be used many times. We therefore prioritize the test data over the training data in this thesis.

When analyzing the data needed for the different cases it was noted that using any feature set reduces the test data size by at least 72% and on average achieves a reduction of 93.80%. The decrease is inversely proportional to the number of features in each set. It follows that Feature set 4 causes the larges reduction with only one feature and while Feature set 3 causes a large reduction, it is still the smallest one with 8 features.

4.3.3 The Dataset

Some limitations of the dataset affect the results through the training and results of our model. While the resolution of the data is great and allows for advanced features such as ICA (F27-29) and end slope of the voltage curve at CC (F3), the number of cells in it limits the scenarios on which the model can be trained. With only 10 cells, out of which one is dedicated to testing, the model's flexibility becomes constrained by only having seen a handful of possibilities. This is noticeable in the second half of the data as only five cells remain, as can be seen in Figure 3.2, and towards the end only three cells remain. This problem is compounded further as one of these three cells is used for testing pushing it down to a count of only two

remaining cells. This means that the model doesn't have sufficient data in the end, making the model overtrained and unreliable for prediction. This lack of multiple cells may for example cause the spike seen at week 40 in Figure 4.6 as the training data jumps up to a single value of the lone cell left.

The cell being used for testing, cell W8, directly affects the results and has some notable behaviours. As mentioned, it is one of the only cells that go through the entire time range which is why it was chosen. It also behaved somewhat as expected and had a C-rate beneath 1C. With that being said, it has one section at $\sim$ week 18 where the capacity unexpectedly drops. This doesn't affect the model itself but instead causes a deviation between the mean prediction and the test data, clearly seen in the baseline. This heavily impacts the baseline RMSE value as it pushes the otherwise aligned lines apart. This remains true when time is the only feature, but when using the other features that take voltage and current into consideration, such as those in Feature set 1, this drop is detected and followed in the prediction. Removing the time feature from Feature set 1 leaves, by process of elimination, at least one of the remaining two features, F6 and F7, as being able to characterize this drop. Either way, it strengthens the argument that more features than only time are necessary to describe the behaviour of LIBs accurately.

5
Conclusion

This work has investigated new sampling methods for LIBs that can be used to make accurate predictions of the SOH while lowering the computational cost for the BMS. For this work, SOH has been defined as the capacity loss from a LIB cell's initial capacity. An open-source battery aging dataset under dynamic driving profiles with 10 LIB cells over 28 months was used to train and evaluate a GP regression model that predicted the capacity loss. To make the GP regression model smaller and more efficient, 29 features were extracted and 8 of the ones most correlated with capacity were selected.

The sampling methods developed were designed with an event-based approach, improving efficiency by only acting when necessary and when new information is available. In total three methods were made, two different event-based methods, referred to as event-based and hybrid, and one periodic to act as a baseline.

The results show that both new sampling methods show promise with predictions equivalent to the baseline but with significantly less data. Out of the two proposed methods, hybrid using the three best features shows the greatest potential with an RMSE decrease of up to 45.28% and a data reduction of 99.4% compared to the baseline. With the improved sampling method and use of features, the computational resources of the BMS can be used more efficiently.

5.1 Future work

Features are a huge area with a lot of development all the time, and while significant time was spent on this, more focus can be allocated to investigating them in-depth. It will be of great importance for future work since good features can significantly increase the accuracy of the ML model while reducing its size. To enhance the value of the features, the previously discussed adjustments of the extracted features could be investigated as well as new features that haven't been included in the work.

The extracted features were based on voltage, current, time and capacity measurements but more sources can be used. As mentioned in Section 2.1.2, the internal resistance is a commonly used indicator for the SOH which would make it an interesting feature to train the model with. Using the resistance as a feature could give the model additional insights into the behaviour of the battery which could

improve the accuracy of the estimation. Other than the resistance, temperature is a commonly used feature when predicting SOH. While the data set used in this work didn't include it, it would be an interesting feature to investigate for future work with a data set that includes information about it.

As mentioned in the discussion about the dataset, the number of cells used to train the model is very limited and doesn't cover a large time span. It would be intriguing to investigate using a dataset with more cells to cover the downfalls of the current one. With a larger set of data, the already limited computational resources would be spread even thinner necessitating better usage of those resources. One way to do this, which we briefly validated, was the use of a GPU to train and inference the model. This is something that we recommend to test further in a potential follow-up work.

Bibliography

[1] Gerald B. Raines. *Electric Vehicles: Technology, Research and Development*. Nova Science Publishers, New York, 2009.

[2] Trafikanalys. Fordon 2023. Technical report, 2024.

[3] Tariq Muneer, Lal Kohle Mohan, Koki Ogura, Aisling Doyle, Irene Illescas García, Matjaž Knez, Girard Aymeric, Simon François, Eulalia Jadraque Gago, Parimita Mohanty, Yash Kotak, T.P. Chathuri Madusha, Michael Jeffrey, and Ross Milligan. *Electric Vehicles: Prospects and Challenges*. Elsevier, Amsterdam, 2017.

[4] Selamat Muslimin, Zainuddin Nawawi, Bhakti Yudho Suprapto, and Tresna Dewi. Comparison of Batteries Used in Electrical Vehicles. *Proceedings of the 5th FIRST T1 T2 2021 International Conference (FIRST-T1-T2 2021)*, 9:421–425, 2 2022.

[5] Prashant Singh, Research Associate, Arkabrata Dattaroy, Naqui Anwer, and Avik Bhattacharya. Comparative Analysis of Battery Energy Storage Systems for Mobile Substation and Grid Storage System. 2022.

[6] Bharat Balagopal, Cong Sheng Huang, and Mo-Yuen Chow. Effect of Calendar Aging on Li Ion Battery Degradation and SOH. In *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, 2017.

[7] Muhammad Nizam, Hari Maghfiroh, Rizal A. Rosadi, and Kirana D.U. Kusumaputri. Design of Battery Management System (BMS) for Lithium Iron Phosphate (LFP) Battery. *ICEVT 2019 - Proceeding: 6th International Conference on Electric Vehicular Technology 2019*, pages 170–174, 11 2019.

[8] Huang Zhang, Yang Su, Faisal Altaf, Torsten Wik, and Sebastien Gros. Interpretable Battery Cycle Life Range Prediction Using Early Degradation Data at Cell Level. *IEEE Transactions on Transportation Electrification*, 9(2):2669–2682, 4 2022.

[9] Rui Xiong, Yongzhi Zhang, Ju Wang, Hongwen He, Simin Peng, and Michael Pecht. Lithium-Ion Battery Health Prognosis Based on a Real Battery Management System Used in Electric Vehicles. *IEEE Transactions on Vehicular Technology*, 68(5):4110–4121, 5 2019.

[10] Haokai Ruan, Hongwen He, Zhongbao Wei, Zhongyi Quan, and Yunwei Li. State of Health Estimation of Lithium-Ion Battery Based on Constant-Voltage Charging Reconstruction. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, 11(4):4393–4402, 8 2023.

[11] Shaohua Li and Yue Yuan. Estimation of State-of-health for Lithium-ion Battery Based on Increment Capacity Analysis Method and Long Short-term Mem-

ory Neural Network. *IEEE IAS Industrial and Commercial Power System Asia*, 2023.

[12] Xing Shu, Shiquan Shen, Jiangwei Shen, Yuanjian Zhang, Guang Li, Zheng Chen, and Yonggang Liu. State of health prediction of lithium-ion batteries based on machine learning: Advances and perspectives. *iScience*, 24(11):103265, 11 2021.

[13] Wuzhao Yan, Bin Zhang, Xiaofeng Wang, Wanchun Dou, and Jingcheng Wang. Lebesgue-Sampling-Based Diagnosis and Prognosis for Lithium-Ion Batteries. *IEEE Transactions on Industrial Electronics*, 63(3):1804–1812, 3 2016.

[14] Guangxing Niu, Xuan Wang, Enhui Liu, and Bin Zhang. Lebesgue Sampling Based Deep Belief Network for Lithium-Ion Battery Diagnosis and Prognosis. *IEEE Transactions on Industrial Electronics*, 69(8):8481–8490, 8 2022.

[15] Gabriele Pozzato, Anirudh Allam, and Simona Onori. Lithium-ion battery aging dataset based on electric vehicle real-driving profiles. *Data in Brief*, 41:107995, 4 2022.

[16] LG Chem. Product Specification, Rechargeable Lithium Ion Battery, model: INR21700 M50T 18.20 Wh, 2018.

[17] Valentin Sulzer, Peyman Mohtat, Antti Aitio, Suhak Lee, Yen T. Yeh, Frank Steinbacher, Muhammad Umer Khan, Jang Woo Lee, Jason B. Siegel, Anna G. Stefanopoulou, and David A. Howey. The challenge and opportunity of battery lifetime prediction from field data. *Joule*, 5(8):1934–1955, 8 2021.

[18] Amnesty International and IBGDH. DRC: Powering Change or Business as Usual?, 9 2023.

[19] Mohamed Elmahallawy, Tarek Elfouly, Ali Alouani, and Ahmed M. Massoud. A Comprehensive Review of Lithium-Ion Batteries Modeling, and State of Health and Remaining Useful Lifetime Prediction. *IEEE Access*, 10:119040–119070, 2022.

[20] Helena Berg. *Batteries for Electric Vehicles: Materials and electrochemistry*. Cambridge University Press, University Printing House, Cambridge CB2 8BS, United Kingdom, 2015.

[21] Robert R. Richardson, Michael A. Osborne, and David A. Howey. Battery health prediction under generalized conditions using a Gaussian process transition model. *Journal of Energy Storage*, 23:320–328, 6 2019.

[22] Weihan Li, Neil Sengupta, Philipp Dechent, David Howey, Anuradha Annaswamy, and Dirk Uwe Sauer. One-shot battery degradation trajectory prediction with deep learning. *Journal of Power Sources*, 506:230024, 9 2021.

[23] Jochen Görtler, Rebecca Kehlbeck, and Oliver Deussen. A Visual Exploration of Gaussian Processes. *Distill*, 4 2019.

[24] infinite curiosity. Interactive Gaussian Process Visualization. URL: http://www.infinitecuriosity.org/vizgp/ [Date accessed: 10/04/2024], 4.

[25] Pingwei Gu, Zhongkai Zhou, Shaofei Qu, Chenghui Zhang, and Bin Duan. Influence Analysis and Optimization of Sampling Frequency on the Accuracy of Model and State-of-Charge Estimation for LiNCM Battery. *Energies 2019, Vol. 12, Page 1205*, 12(7):1205, 3 2019.

[26] K. J. Åström and B. M. Bernhardsson. Comparison of Riemann and Lebesgue sampling for first order stochastic systems. *Proceedings of the IEEE Conference on Decision and Control*, 2:2011–2016, 2002.

[27] Duo Yang, Xu Zhang, Rui Pan, Yujie Wang, and Zonghai Chen. A novel Gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve. *Journal of Power Sources*, 384:387–395, 4 2018.

[28] W. Waag and D. U. Sauer. Secondary Batteries - Lead- Acid Systems | State-Of-Charge/Health. *Encyclopedia of Electrochemical Power Sources*, pages 793–804, 1 2009.

[29] Robert Khizbullin, Boris Chuvykin, and Ronald Kipngeno. Research on the Effect of the Depth of Discharge on the Service Life of Rechargeable Batteries for Electric Vehicles. *Proceedings - 2022 International Conference on Industrial Engineering, Applications and Manufacturing, ICIEAM 2022*, pages 504–509, 2022.

[30] Yang Chang, Huajing Fang, Sai Li, and Yan Yan. Prognostics for lithium-ion battery operating under different depth of discharge using hybrid method. *Proceedings of the 30th Chinese Control and Decision Conference, CCDC 2018*, pages 6239–6244, 7 2018.

[31] Samuel Greenbank, David Howey, and Senior Member. Automated Feature Extraction and Selection for Data-Driven Models of Rapid Battery Capacity Fade and End of Life. *IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS*, 18(5), 2022.

[32] Xiaoyu Li, Changgui Yuan, Xiaohui Li, and Zhenpo Wang. State of health estimation for Li-Ion battery using incremental capacity analysis and Gaussian process regression. *Energy*, 190:116467, 1 2020.

[33] Abraham Savitzky and Marcel J.E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 7 1964.

Appendices

A

Voltage curves

Voltage curves during the second CV and CC steps extracted from the data set. Used to extract features F1-F4 in Section 3.3.

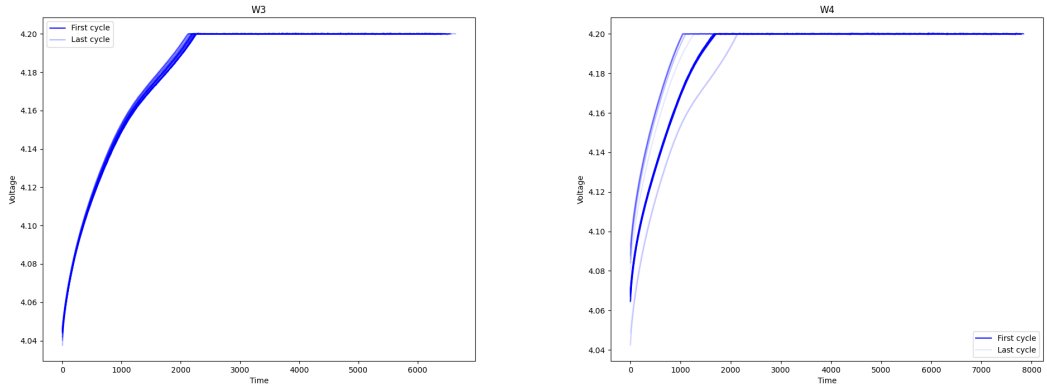

Figure A.1: Comparison of the voltage in constant current and constant voltage charging modes for every 5th cycle of cell W3 and W4

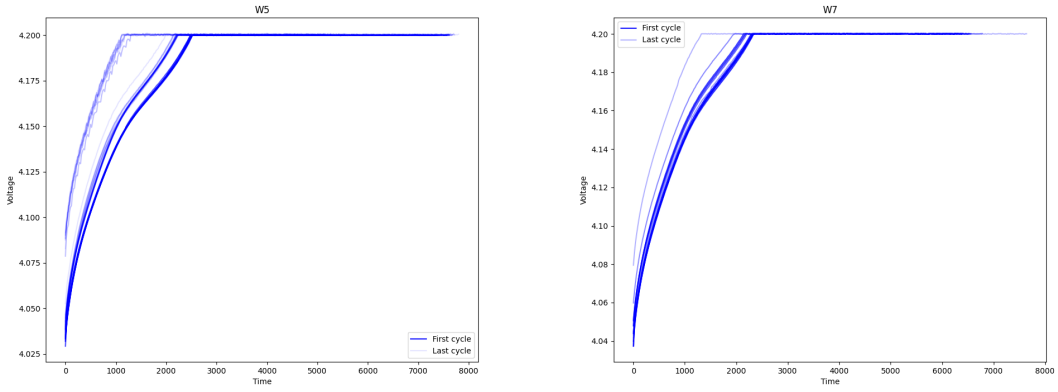

Figure A.2: Comparison of the voltage in constant current and constant voltage charging modes for every 5th cycle of cell W5 and W7

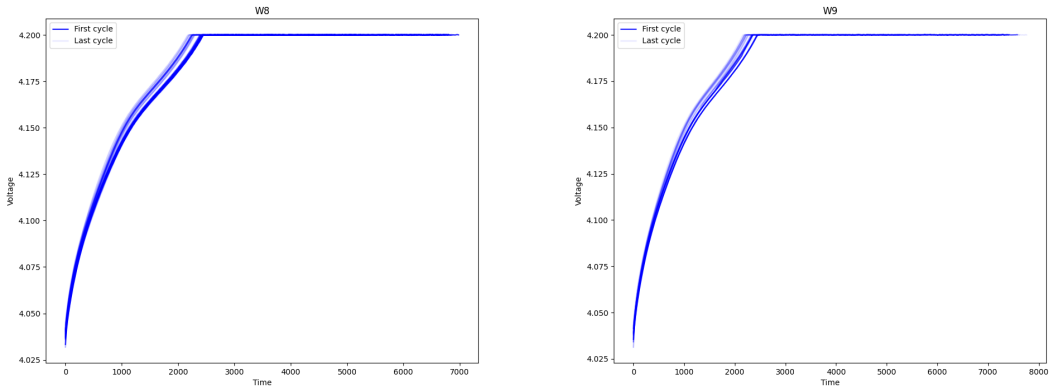

Figure A.3: Comparison of the voltage in constant current and constant voltage charging modes for every 5th cycle of cell W8 and W9

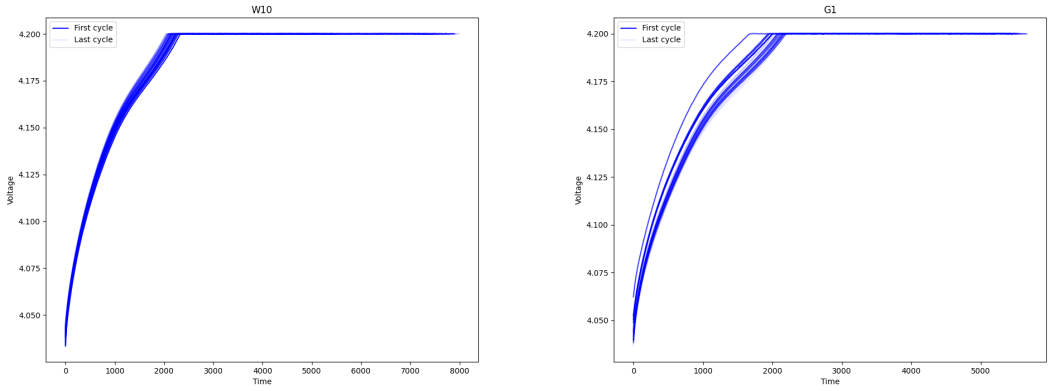

Figure A.4: Comparison of the voltage in constant current and constant voltage charging modes for every 5th cycle of cell W10 and G1

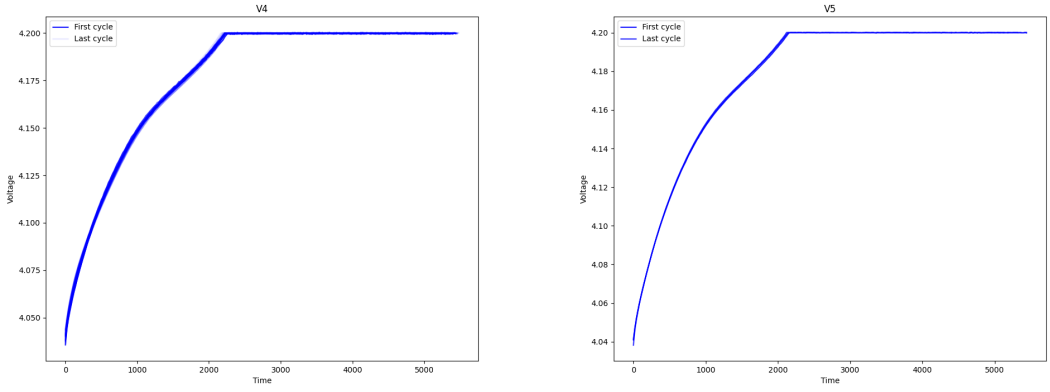

Figure A.5: Comparison of the voltage in constant current and constant voltage charging modes for every 5th cycle of cell V4 and V5

B

Correlation matrices

For the first part of the feature selection, the window-based features were compared separately. These correlation matrices are shown in Figures B.1 and B.2. The combined matrix made from these can be found in Section 4.1.

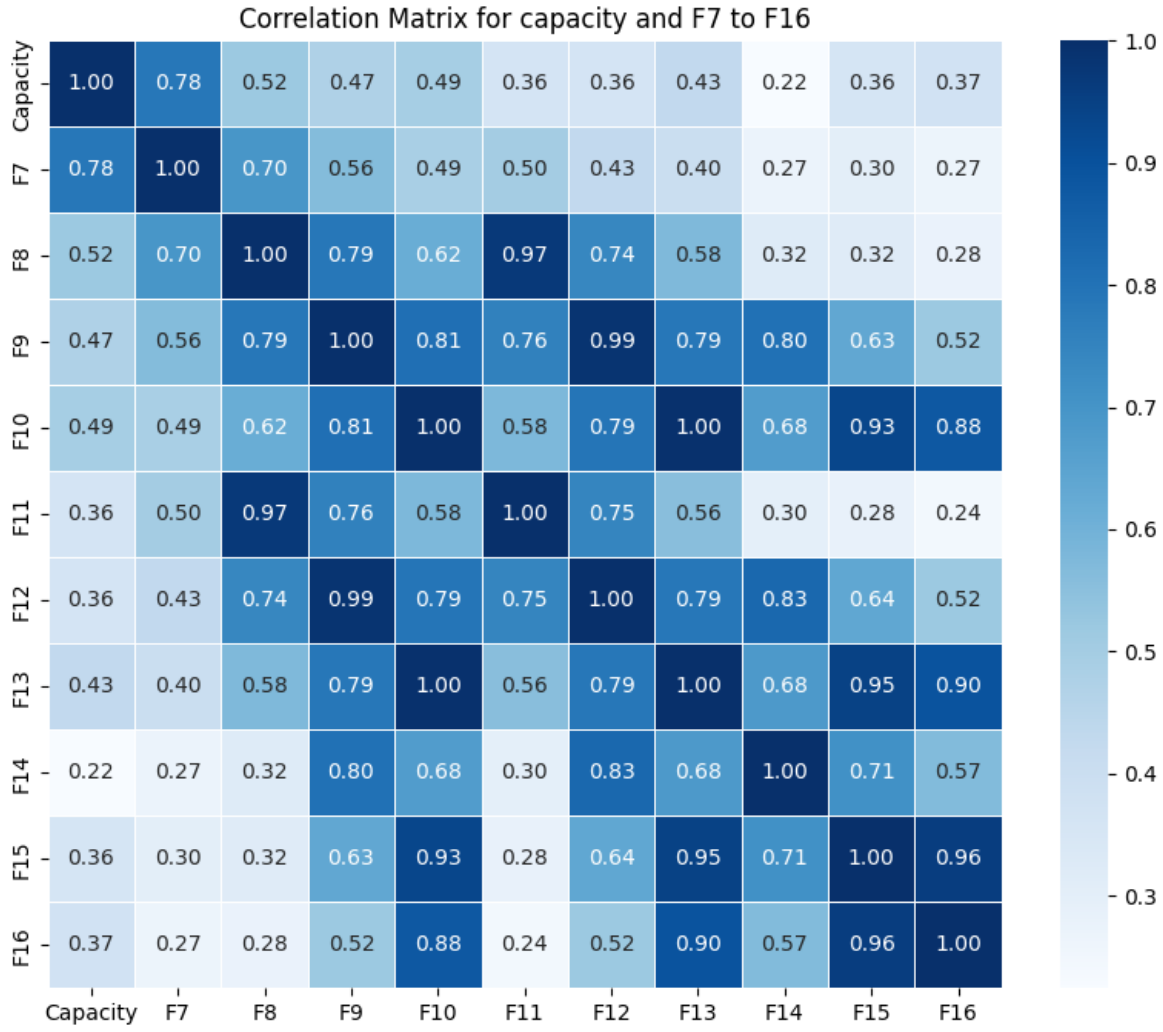

Figure B.1: Correlation matrix of capacity and histogram-based voltage features for the training data

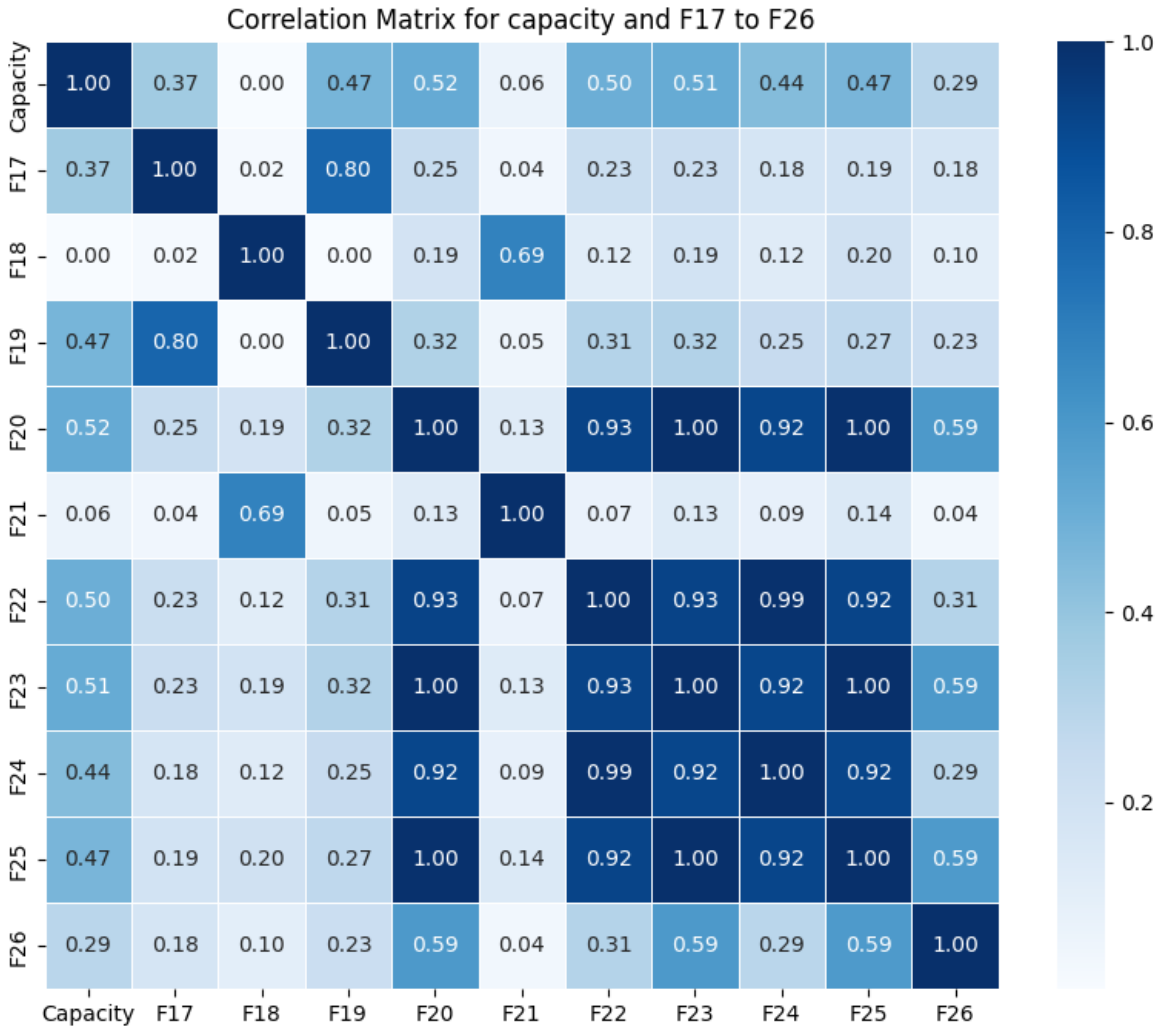

Figure B.2: Correlation matrix of capacity and histogram-based current features for the training data

After removing some of the window-based features, the remaining were integrated into a correlation matrix together with the rest of the features. This correlation matrix is shown in Figure B.3.

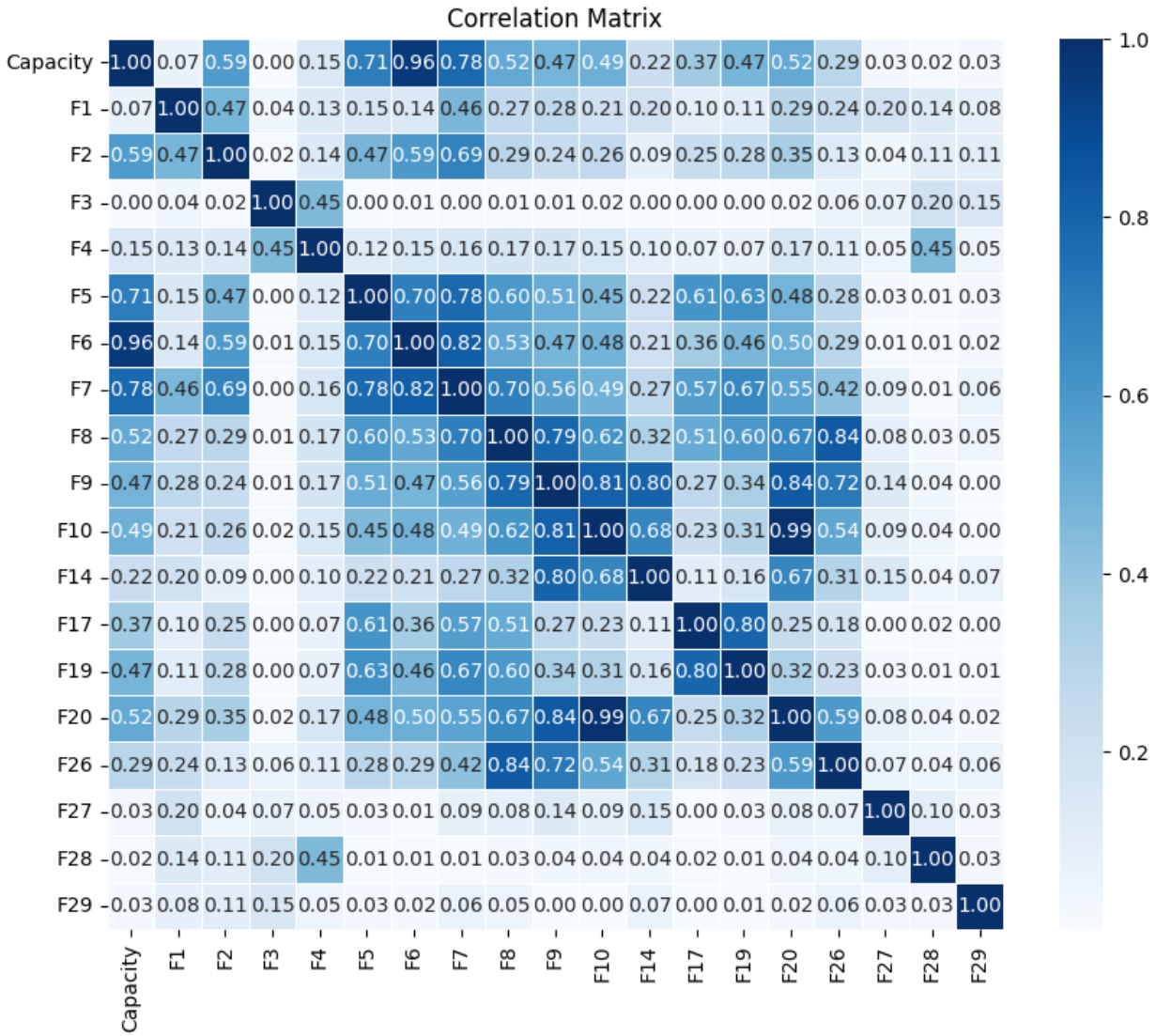

Figure B.3: Correlation matrix for features after removing strongly correlated features in the histogram-based feature set using the training data

C
Predictions

The plots from all cases tested in Section 4.2 are presented in this chapter. For setup of each case see Table 3.9.

Periodic

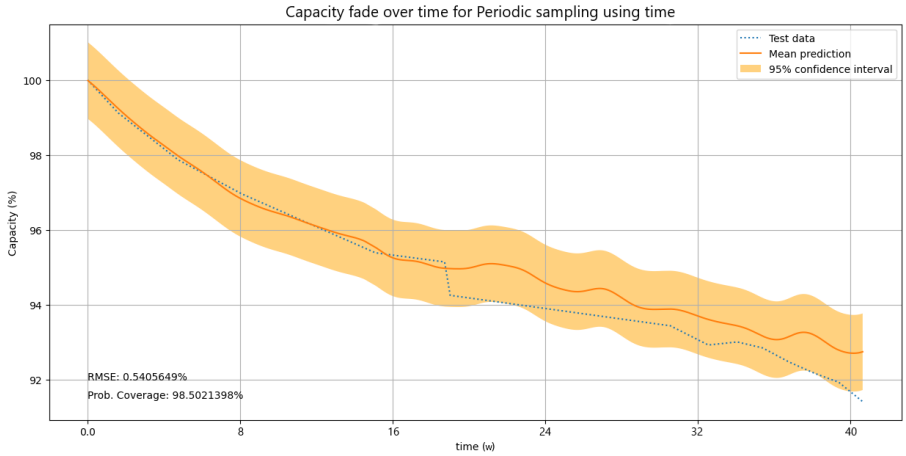

Figure C.1: Prediction of capacity loss from the initial capacity for the baseline case

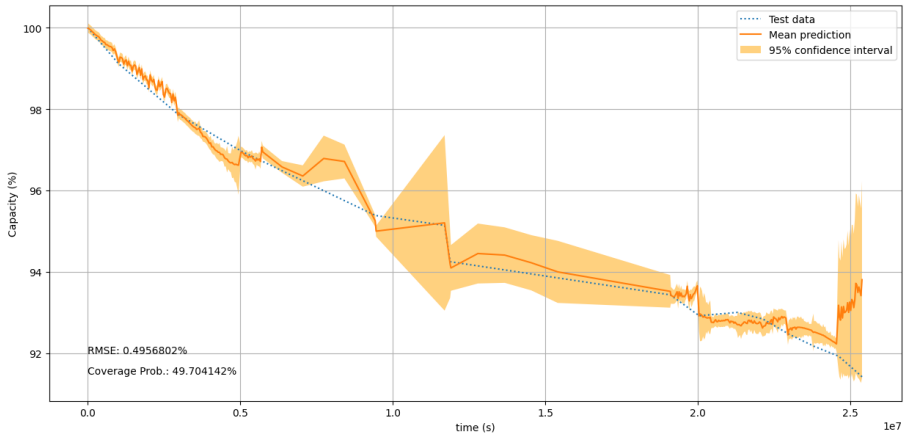

Figure C.2: Prediction of capacity loss from the initial capacity for Case 2

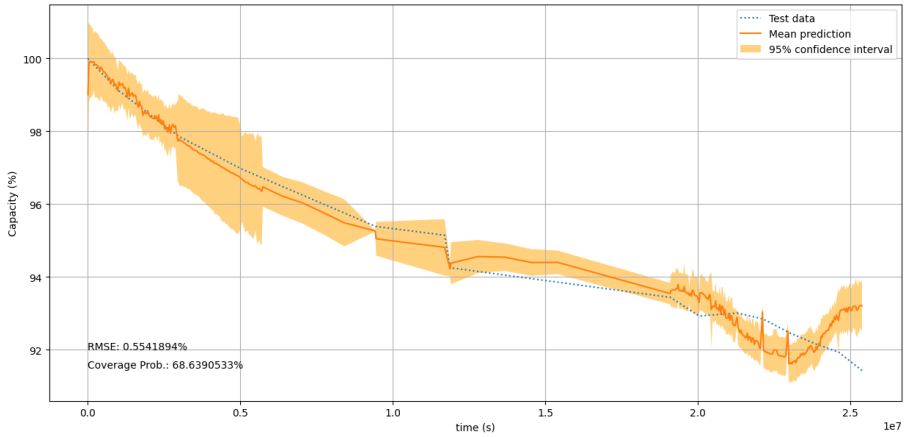

Figure C.3: Prediction of capacity loss from the initial capacity for Case 3

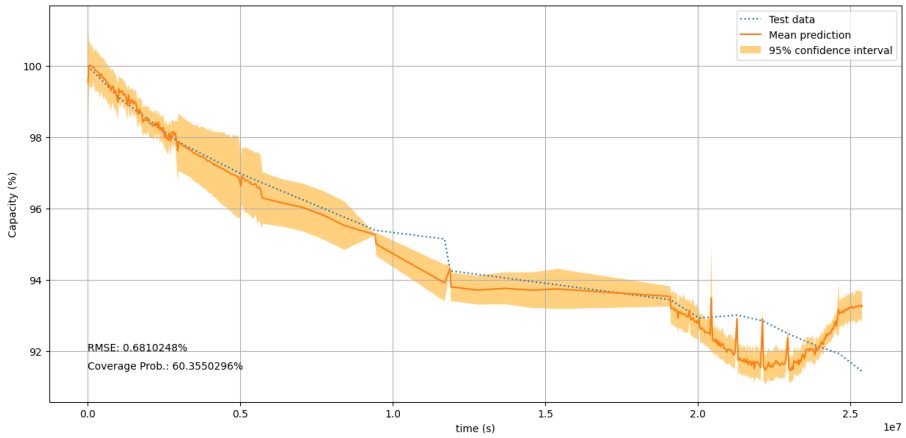

Figure C.4: Prediction of capacity loss from the initial capacity for Case 4

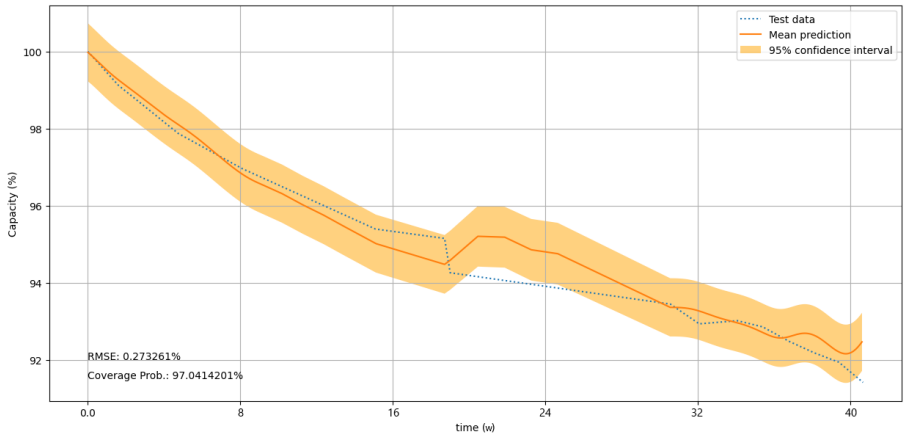

Figure C.5: Prediction of capacity loss from the initial capacity for Case 5

Event-based

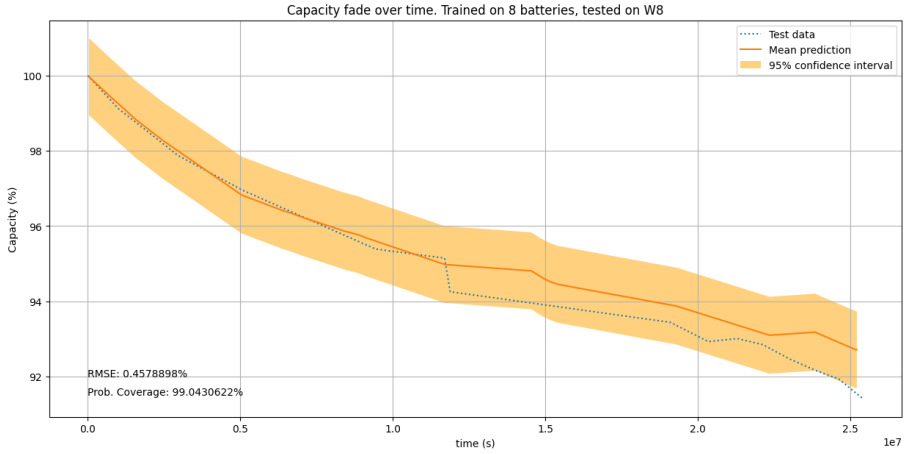

Figure C.6: Prediction of capacity loss from the initial capacity for Case 6

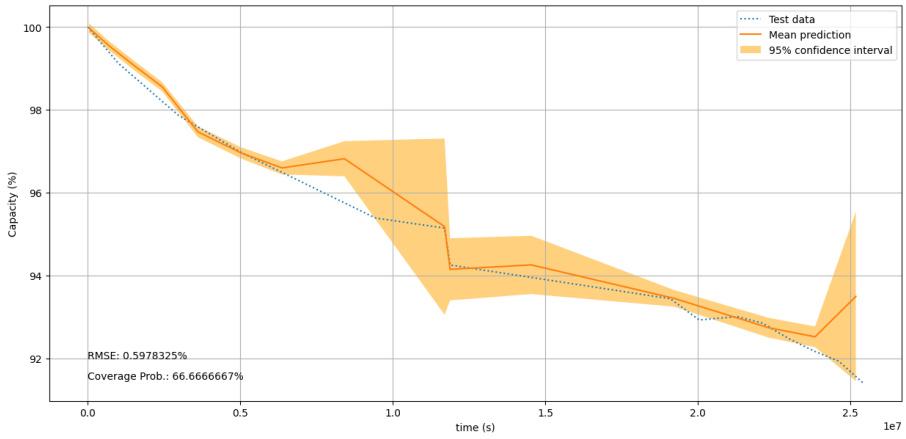

Figure C.7: Prediction of capacity loss from the initial capacity for Case 7

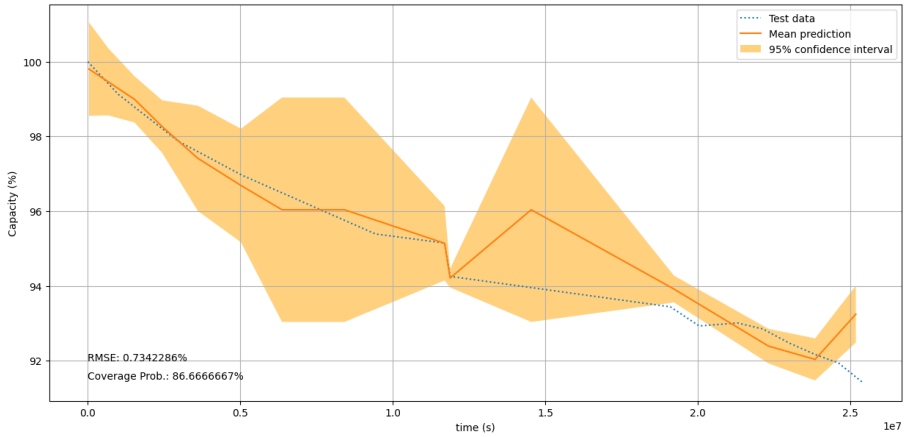

Figure C.8: Prediction of capacity loss from the initial capacity for Case 8

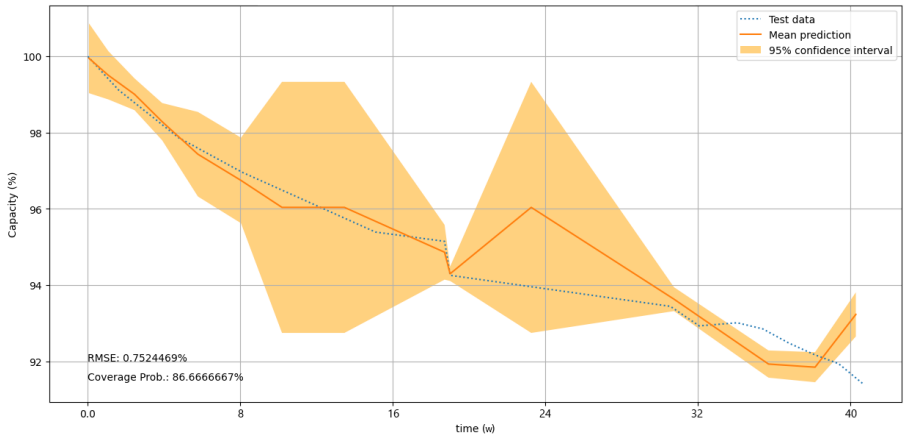

Figure C.9: Prediction of capacity loss from the initial capacity for Case 9

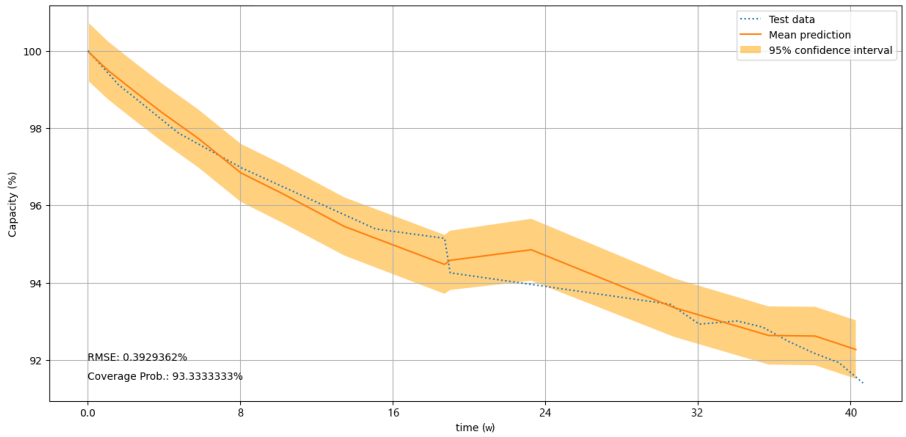

Figure C.10: Prediction of capacity loss from the initial capacity for Case 10

Hybrid

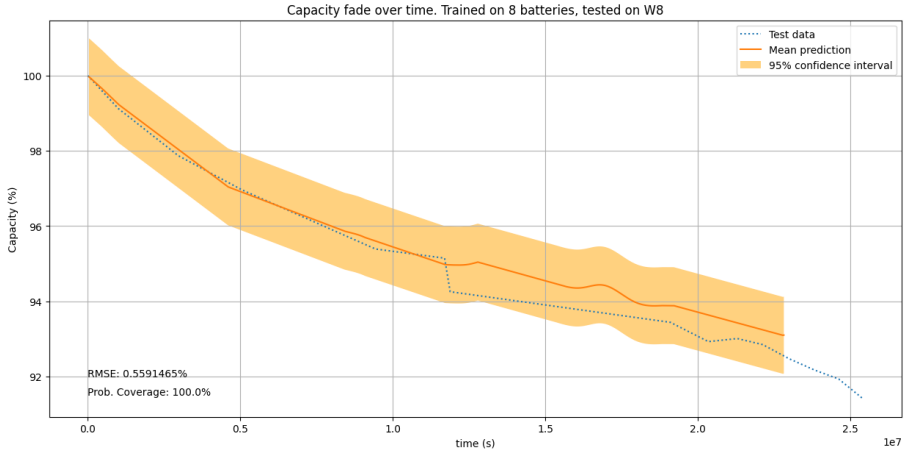

Figure C.11: Prediction of capacity loss from the initial capacity for Case 11

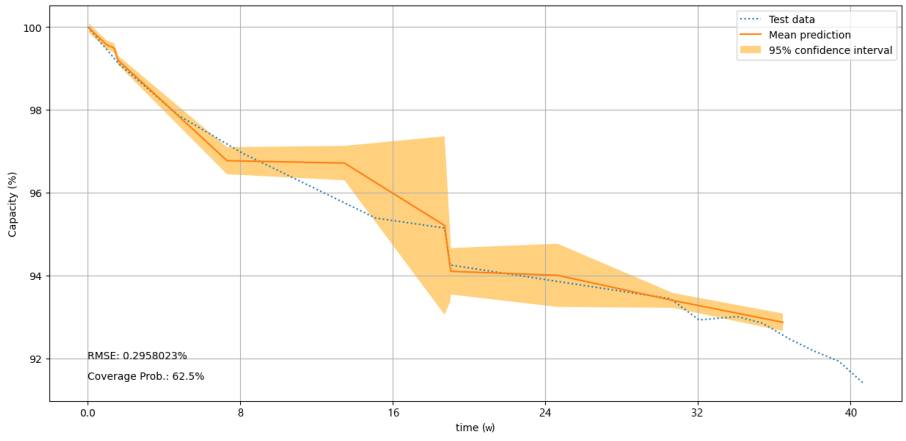

Figure C.12: Prediction of capacity loss from the initial capacity for Case 12

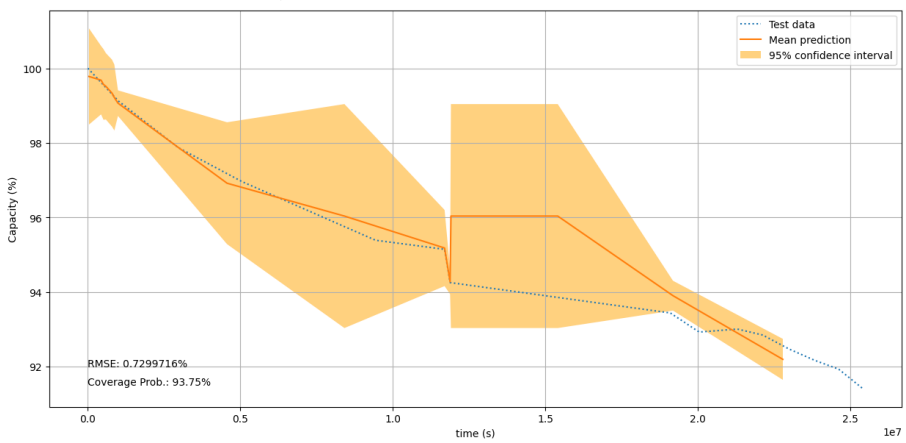

Figure C.13: Prediction of capacity loss from the initial capacity for Case 13

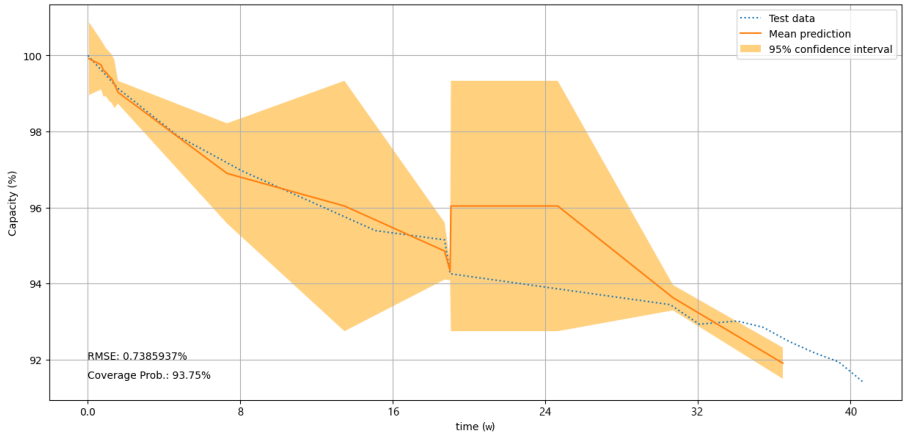

Figure C.14: Prediction of capacity loss from the initial capacity for Case 14

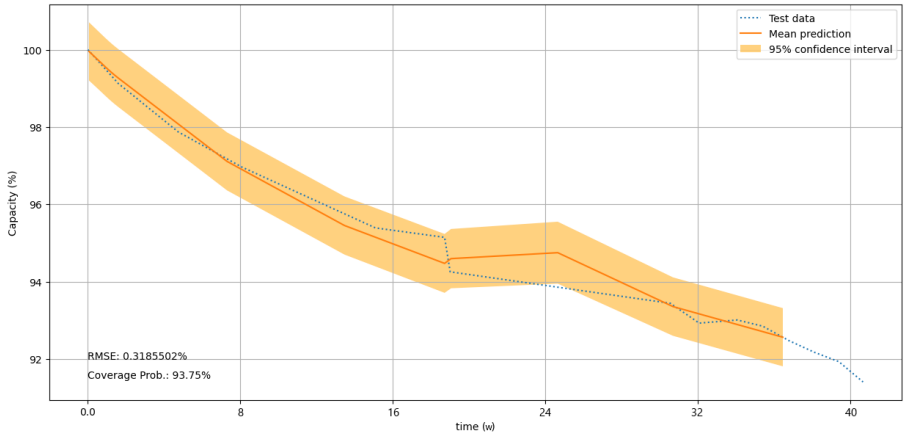

Figure C.15: Prediction of capacity loss from the initial capacity for Case 15