Development of data-driven method for capacity estimation and prognosis for lithium-ion batteries

Koul, Akhilesh

2020

Koul, A. (2020). Development of data-driven method for capacity estimation and prognosis for lithium-ion batteries. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/137313>

<https://doi.org/10.32657/10356/137313>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 22:53:08 SGT

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**DEVELOPMENT OF DATA-DRIVEN
METHOD FOR CAPACITY
ESTIMATION AND PROGNOSIS FOR
LITHIUM-ION BATTERIES**

**AKHILESH KOUL
SCHOOL OF ELECTRICAL & ELECTRONIC
ENGINEERING**

2020

DEVELOPMENT OF DATA-DRIVEN METHOD FOR CAPACITY ESTIMATION AND PROGNOSIS FOR LITHIUM-ION BATTERIES

AKHILESH KOUL

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Master of Engineering

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

21-02-20

.....

Date

.....

AKHILESH KOUL

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

21-02-20

.....

Date

Xu Yan .

.....

Dr XU YAN

Authorship Attribution Statement

This thesis does not contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

21-02-20

.....

Date

A handwritten signature in black ink, appearing to read 'Akhilesh', written over a horizontal dotted line.

.....

AKHILESH KOUL

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to everyone whose direct or indirect support enabled me to complete this thesis. First and foremost, I am grateful to my supervisor, Dr Xu Yan, who saw the potential in me and agreed to be my mentor for this research. The amount of effort he puts in supervising and consulting students is ever-inspiring. He never misses a chance to give the right advice, be it research or other career matters.

Secondly, I would like to express my thankfulness to all the colleagues and supervisors, namely, Dr Koh Leong Hai and Dr Koh Liang Mong from Energy Research Institute @NTU, Dr Gou Bin, Benjamin Chew and all the members of my research group, who were a part of my projects and have guided me throughout my research. It was a great learning experience working with everyone. I am also grateful to Energy Research Institute @NTU for providing financial support for this work.

I would like to thank the School of Electrical and Electronic Engineering to provide me with a platform to show my research potential as a masters student. Notably, the resources provided by the Clean Energy Research Lab, managed by the Lab-in-charge Mdm Chia-Nge Tak Heng, offered me a conducive environment to perform my studies and work on my research. I am thankful to Nanyang Technological University for providing a multi-disciplinary, diverse and holistic environment, not to mention all the housing and other prominent facilities, which made the entire research process easier to bear.

Last but not least, I would not miss the chance to genuinely show my gratitude to my parents who are always there for me and friends who supported me while I went through various twists and turns during my candidature as a master's student at Nanyang Technological University.

Table of Contents

Statement of Originality	iii
Supervisor Declaration Statement	v
Authorship Attribution Statement	vii
Acknowledgements	ix
Table of Contents	xi
Abstract	xiii
List of Figures	xiv
List of Tables	xvii
Acronyms	xix
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	8
1.3 Organization of the Thesis	9
2 Literature Review	11
3 Framework	19
3.1 Stochastic Gradient Boosting Regression	19
3.1.1 Function Estimation	21
3.1.2 Numerical Optimization	22
3.1.3 An Optimisation in Function Space	23
3.1.4 Gradient Boosting Algorithm	25
3.2 Auto-Regressive Integrated Moving Average	28
3.3 Feature Extraction and Selection	32
3.4 System Design	37
3.4.1 Offline / Training Phase	37
3.4.2 Online / Runtime Phase	43

4	Dataset and Results	47
4.1	NASA's Battery Dataset	47
4.1.1	Feature Selection	48
4.1.2	ARIMA Order Identification	50
4.1.3	Capacity Estimation	54
4.1.4	Capacity Prognosis	56
4.2	CALCE's CX2 Lithium-ion Cells Dataset	60
4.2.1	Feature Selection	61
4.2.2	ARIMA Order Identification	62
4.2.3	Capacity Estimation	66
4.2.4	Capacity Prognosis	70
4.3	NASA's Randomized Battery Usage Dataset	70
4.3.1	Feature Selection	73
4.3.2	ARIMA Order Identification	74
4.3.3	Capacity Estimation	77
4.3.4	Capacity Prognosis	81
5	Conclusion	83
5.1	Future Works	85
A	Fuzzy Logic Rules	87
	Bibliography	89

Abstract

With an ongoing transition from traditional energy sources to renewable energy sources, which inherently are intermittent in nature, the electrical energy storage is becoming more and more important for managing energy production and demand. Due to this, lithium-ion batteries have emerged as the key technologies in the field of energy storages. However, to ensure that adequate safety and better performance, there is a need for health monitoring of the current battery state and parameters such as capacity, state-of-health, remaining useful life. Monitoring of these parameters ensures that the batteries are being used efficiently. Also, to maximise battery life, such health parameter monitoring opens up the possibility of optimisation of battery usage. Capacity, which quantifies the available energy in a fully charged Li-ion battery, is an vital index that can be used in interfering the state-of-health or the remaining useful life. This work explores the data-driven method that can be used in estimating the current capacity and forecasting the capacity trend for the future charge cycles of the battery whose internal health parameters are difficult to gauge. The features used to develop data-driven capacity degradation model are obtained from voltage, current and time measurements observed during the charging phase of the battery, which is operated under the constant current - constant voltage charging protocol. With the developed model, capacity degradation can be estimated in respect with cyclic ageing of the battery. Stochastic gradient boosting regression (SGBR) ensemble with an autoregressive integrated moving average (ARIMA) is used for capacity estimation and prognosis. Features obtained are used to train respective SGBR models with a target value as the actual capacity or true capacity obtained using coulomb counting method from consecutive discharge cycle. For prognosis, ARIMA models are developed to forecast the features for future unobserved cycles using observed features and used as an input feature in another SGBR model to provide the predicted capacity for the unobserved cycles with the confidence interval. In actual operation, batteries are seldom fully charged/discharged, therefore during online capacity estimation,

not all the features will be available. To solve the issue of data unavailability during partial charge/discharge, the presented method does not require the full range of measurements for prediction. Instead, it uses the sets of time window, i.e. measurements belonging to different voltage and current ranges, and since the features are obtained during charging of the battery, it does not affect the normal working on the battery. Fuzzy intelligent system is implemented to deal with the missing value cases in the runtime phase. In addition to this, the proposed method also presents a method for missing value imputation in case one of the voltages or current feature set is not observed using one-step forecasting via the ARIMA model. The results are presented using three independent experimental dataset provided by the National Aeronautics and Space Administration(NASA) Prognostic Center and Center for Advanced Life Cycle Engineering (CALCE) Battery Research Group based in the University of Maryland, in which various experiments were performed on a lithium-ion battery. The results demonstrate the effectiveness and accuracy of the proposed framework for battery capacity estimation and prognosis.

List of Figures

1.1	Price development of Li-ion batteries 2005-2030	6
1.2	Gravimetric Energy Density and Specific Power of Different Available Battery Technologies	7
3.1	Typical Charging Curve for Li-ion Battery Operated under CC-CV Protocol	33
3.2	Trend in Voltage Charge Curve as Li-ion Battery Ages	34
3.3	Trend in Current Charge Curve as Li-ion Battery Ages	34
3.4	Offline or Training Phase System Design Chart	43
3.5	Online or Runtime Phase Flow Chart	45
4.1	Feature Importance observed in CC mode for NASA's Battery Dataset	49
4.2	Feature Importance observed in CV mode for NASA's Battery Dataset	49
4.3	ACF and PACF plot for $\log_Set_{V_1}$ observed in B0005	51
4.4	ACF and PACF plot for $\log_Set_{V_2}$ observed in B0005	51
4.5	ACF and PACF plot for $\log_Set_{I_1}$ observed in B0005	52
4.6	ACF and PACF plot for $\log_Set_{I_2}$ observed in B0005	52
4.7	ACF and PACF plot for $\log_Tot_en_Discha$ observed in B0005	53
4.8	ACF and PACF plot for $\log_Tot_en_Cha$ observed in B0005	53
4.9	Capacity Estimation for B0018 plotted against Cycle Number	54
4.10	Capacity Estimation for B0007 plotted against Cycle Number	55
4.11	Capacity Estimation for B0006 plotted against Cycle Number	55
4.12	Capacity Estimation for B0005 plotted against Cycle Number	56
4.13	Capacity Prognosis for B0018 forecasted at 30 th Cycle Number	57
4.14	Capacity Prognosis for B0018 forecasted at 60 th Cycle Number	58
4.15	Capacity Prognosis for B0018 forecasted at 90 th Cycle Number	58
4.16	Capacity Prognosis for B0018 forecasted at 120 th Cycle Number	59
4.17	Feature Importance observed in CC mode for CALCE's CX2 Battery Dataset	61
4.18	Feature Importance observed in CV mode for CALCE's CX2 Battery Dataset	62
4.19	ACF and PACF plot for $\log_Set_{V_1}$ observed in CX2.33	63
4.20	ACF and PACF plot for $\log_Set_{V_2}$ observed in CX2.33	63

4.21	ACF and PACF plot for $\log_Set_{I1}$ observed in CX2.33	64
4.22	ACF and PACF plot for $\log_Set_{I2}$ observed in CX2.33	64
4.23	ACF and PACF plot for $\log_Tot_en_Discha$ observed in CX2.33	65
4.24	ACF and PACF plot for $\log_Tot_en_Cha$ observed in CX2.33	65
4.25	Capacity Estimation for CX2.38 plotted against Cycle Number	66
4.26	Capacity Estimation for CX2.37 plotted against Cycle Number	67
4.27	Capacity Estimation for CX2.36 plotted against Cycle Number	67
4.28	Capacity Estimation for CX2.35 plotted against Cycle Number	68
4.29	Capacity Estimation for CX2.34 plotted against Cycle Number	68
4.30	Capacity Estimation for CX2.33 plotted against Cycle Number	69
4.31	Feature Importance observed in CC mode for NASA's Randomized Battery Usage Dataset	73
4.32	Feature Importance observed in CV mode for NASA's Randomized Battery Usage Dataset	73
4.33	ACF and PACF plot for $\log_Set_{V1}$ observed in RW3	74
4.34	ACF and PACF plot for $\log_Set_{V2}$ observed in RW3	75
4.35	ACF and PACF plot for $\log_Set_{I1}$ observed in RW3	75
4.36	ACF and PACF plot for $\log_Set_{I1}$ observed in RW3	76
4.37	ACF and PACF plot for $\log_Tot_en_Discha$ observed in RW3	76
4.38	ACF and PACF plot for $\log_Tot_en_Cha$ observed in RW3	77
4.39	Capacity Estimation for RW15 plotted against Cycle Number	78
4.40	Capacity Estimation for RW16 plotted against Cycle Number	79
4.41	Capacity Estimation for RW1 plotted against Cycle Number	79
4.42	Capacity Estimation for RW2 plotted against Cycle Number	80
4.43	Capacity Estimation for RW7 plotted against Cycle Number	80
4.44	Capacity Estimation for RW8 plotted against Cycle Number	81

List of Tables

1.1	Comparison of current battery technologies	7
3.1	Feature Space for different SGBR Models	39
3.2	Index-Based Series used for ARIMA	42
3.3	Feature Space for SGBR model used for Prognosis	42
4.1	Details of Experimental NASA’s Battery Dataset	48
4.2	Capacity Estimation Results for NASA’s Battery Dataset	56
4.3	Capacity Prognosis Results for NASA’s Battery Dataset	60
4.4	Details of Experimental CALCE’s CX2 Battery Dataset	61
4.5	Capacity Estimation Results for CALCE’s CX2 Battery Dataset	69
4.6	Capacity Prognosis Results for CALCE’s CX2 Battery Dataset	70
4.7	Load Setpoint with Probability	71
4.8	Details of Experimental NASA’s Randomized Battery Usage Dataset	72
4.9	Capacity Estimation Results for NASA’s Randomized Battery Usage Dataset	81
4.10	Capacity Prognosis Results for NASA’s Randomized Battery Usage Dataset	82

Acronyms

ACF	Partial Autocorrelation Function
Ah	Ampere-Hour
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AR	Auto-regressive
ARIMA	Autoregressive Integrated Moving Average
BMC	Bayesian Monte Carlo
CC	Constant Current
CI	Confidence Interval
CV	Constant Voltage
DOD	Depth of Discharge
EKF	Extended Kalman Filter
EOL	End of Life
ESS	Energy Storage Systems
EV	Electric Vehicles
GBM	Gradient Boosting Machines
HEV	Hybrid Electric Vehicles
HI	Health Indicators
KS	Kernel Smoothing
Li-ion	Lithium-ion
MA	Moving Average
NN	Neural Network
OCV	Open Circuit Voltage
PACF	Autocorrelation Function
PF	Particle Filter

PV	Photovoltaic
RPF	Regularised Particle Filter
RUL	Remaining Useful Life
RVM	Relevance Vector Machine
RVR	Relevance Vector Regression
SEI	Solid Electrolyte Interphase
SGBR	Stochastic Gradient Boosting Regression
SOC	State of Charge
SOH	State of Health
SVM	Support Vector Machine
SVR	Support Vector Regression
UKF	Unscented Kalman Filter
UPF	Unscented Particle Filter
USD	United States Dollar
Wh	Watt-Hour

Chapter 1

Introduction

1.1 Motivation

By observing current trends in energy sector, we can see there is much enthusiasm for renewable energy systems, such as wind turbines and solar photovoltaic cells. Many peoples see these renewables as the archetypal ‘sustainable technologies’, i.e. technologies that can continue to be used in the future without irreparably or irreversibly damaging the ecosystem. With concerns about climate change growing, the rapid development of renewable energy technologies looks increasingly important. However, these technologies face an arduous struggle in trying to become established. While the development of the technology itself is relatively straight forward, the social and institutional implementation problems are often much harder to resolve.

Renewable energy technologies rely on the use of natural energy resources such as solar radiation, winds, ocean waves and tides, which are continually replenished and will therefore not run out. Besides, the use of these renewable sources does not, in general, involve the production of pollutants or other environmentally damaging emissions. It seems likely that, as the environmental costs of the existing energy technology become more apparent, the renewable energy technology will increasingly come into the force. The adoption of more efficient energy use patterns and techniques can perhaps help us to reduce overall demand in most sectors of the economy thus energy conservation plays a significant role. However, we will still

need to find new supplies of energy.

Fortunately, the renewable resource potential is enormous, and the economics of the conversion technologies are beginning to look reasonable. For example, wind power can now be seen as competitive with conventional sources in some contexts, and its costs are continually falling. There are already around 9000 megawatts of wind-powered generating capacity around the world, equivalent to about four nuclear plants. The USA led the way, but Europe has caught up, with Germany's major wind power programme overtaking the initial lead held by Denmark [1]. The UK, which has the Europe's largest wind energy resource, is following suit, with more than 30 wind farms already in place and more planned [2].

Wind power is not the only option. For example, in addition to developing a very successful wind turbine export business, Japan has launched a very visionary 'New Sunshine' sustainable energy programme, which will involve the expenditure of about 1.55 trillion Yen (around 13 billion USD) by 2030, with a strong emphasis on 'photovoltaic' solar cells. The USA and Europe also have major photovoltaic programmes. Interest in various types of biofuels is also increasing around the world, and there is a range of other renewable options under development, including wave energy systems and devices for extracting energy from tidal currents [3].

While there are many advantages of this transition, such as a less carbon footprint and subsequently lower environmental impact with an increase in energy security, there are many challenges on the way which should be faced first. Of all the problems, one of the most critical challenges is the question of electrical energy storage. Due to the intermittent nature of renewable energy, the production of energy depends on the factors like time of day, climate and weather, which is in contrast to energy production from traditional sources. Therefore, it becomes very critical for energy storages to manage the variable demand for electricity in an economical and efficient manner. In simple words, there should be an energy storage system in place when there is an excess of energy production, and during peak demand which cannot meet production, previously stored energy should be used into to reduce the strain on the energy system. Today in the industry there exists a multitude of suggested technologies, so-called energy storage systems (ESS) which are

used for this purpose. Based on the form in which the energy is stored, these energy storage systems can be divided into the electrical, which comprises capacitors, super-capacitors; mechanical, which includes flywheels, pumped hydroelectric systems, compressed air; chemical, can be hydrogen or other chemical storage; thermal, like hot water, molten salts and electrochemical, which comprises of batteries [4].

Among the above-listed energy storage systems, batteries offer many vital advantages like better efficiency, low in pollution, fast response time and low maintenance. Because of the modular nature of battery technologies, it provides better flexibility and adaptability [5]. During peak-shaving, intensive grid loads and even as a back-up system for controlling voltage drops in the energy grid, battery power represents a desirable option for renewable energy storage. Moreover, with the increased interest in the transition from traditional energy source towards renewable energy, batteries will become much more critical due to renewable energy's associated dependance on climate and weather [6–8].

Inside the wide varieties and technologies in the battery energy storage segment, lithium-ion batteries are the most popular and are currently the market-leading technology [7, 9, 10]. This interest in lithium-ion batteries is due to their suitability for mobile applications, especially electric vehicles (EVs), hybrid electric vehicles (HEVs) and applications in Smart Grids [6]. Due to all these applications and growing markets, there has been an increased interest in research within the field of the lithium-ion battery technology [11]. In recent times, areas like battery materials, efficiency, reliability, safety and management systems are being developed very rapidly to meet customer's demands for various applications. There is an overlap in requirements for batteries for electric vehicles and stationary energy storage. In both applications, there is a requirement of long-term stability, safety, low costs and high energy density. The main differences in both the application are mainly in the type of energy density (volumetric density for stationary power, gravimetric density for mobile applications), and demands for flexibility of power deliverance [5]. Therefore, it is very advantageous to have synergistic development in both applications. There is also an economic factor involved because there is a possibility that “worn-out” EV/HEV lithium-ion batteries can be reused for

stationary applications, thus allowing the battery to have a second life and with respect to this the lifetime can be increased by as much as a factor of two [6].

In a relatively short period, lithium-ion batteries have become an essential technology in ESS. Also, by the current trend in the market, it will remain a key competitor in the battery sector for many years to come mainly due to innovations in safety, cost, and energy density [12]. The goal of this background section is to provide a general introduction and overview into the lithium-ion battery technology at large by considering historical, economic, physical and chemical aspects.

The lithium-ion battery technology that we all know today was first developed during the 1980s by Asahi Chemicals and Sony, and commercially it was available in 1991. Observing the growing market of mobile applications, mainly small portable electronics such as mobile phones, video cameras and laptop computers, the main aim was to develop a novel battery technology with increased energy densities compared to existing solutions for the market [13]. Because of its low molecular weight, leading to high energy density and fast diffusion, there was increased interest in lithium-based batteries [5, 12]. At the time, nickel-cadmium (Ni-Cd) and nickel-metal hydride (Ni-MH) were the leading competing technologies, but both were lacking in terms of gravimetric and volumetric energy density compared to the new lithium-ion technology. Unlike in other batteries technologies, the memory effect (permanent loss of capacity when not fully charging the battery) was also lower for lithium-ion batteries [14], which also resulted in its popularity. It was long believed that lithium-ion batteries suffered from no memory effect at all until recently when it has been proven there is at least some memory effect to a minimal extent in lithium-ion batteries [15].

As with the use of organic solvents instead of an aqueous solution as the electrolyte for these batteries, in construction, the cells should be pressurised. Due to this concern regarding safety, special requirements were raised in the structure of the cells. When abused or improperly assembled or designed lithium-ion batteries can infamously explode, one such case is the widely published Samsung Note scandal. Due to concerns like these, testing protocols for lithium-ion batteries should be very rigorous and stringent, which naturally is an added cost over other more

inherently safe battery technologies [16].

Also, economically, the manufacturing cost of lithium-ion batteries is relatively higher than that of contemporary solutions and technologies. This high cost is due to the comparatively expensive transition metal used as cathode [17]. For long-term sustainability, there exists some concern regarding the use of lithium metal in the manufacturing process, due to limited world inventory. Due to the limited supply of lithium, there is research going on to produce batteries based on sodium-ion intercalation compounds, which are very much similar in chemistry. As sodium being an alkali metal like lithium and the available supply of sodium overall is much better than that of lithium [5]. Besides all the listed concerns, in literature, authors have however claimed that the price of lithium is in fact not a significant driver of battery prices or will at least not be so long term [18].

Since the introduction of lithium-ion battery technology, its development and market share has changed very drastically. During the year between 1995 and 2005, energy density was doubled, and prices were halved. Compared to 1991, today the prices are one-tenth of what they were back in 1991, and similarly, sales have been increased dramatically. In the year 2013 alone, a total of five billion lithium-ion cells were sold just for powering of portable electronics. With all the advancement and increased market share, there is a growing concern within the industry that lithium-ion battery performance will soon reach its peak after continual improvements for over 25 years. Today, researchers believe that the limit of lithium-ion Batteries gravimetric density can only be increased by further 30 per cent and not more than that [19].

Presently the lithium-ion batteries are about four to eight times more expensive than equivalent lead-acid alternatives and one to four times more costly than Ni-MH [6]. As per the goal set for 2022 by the American Department of Energy, the average price of lithium-ion batteries should drop further to around USD 125/kWh, down from USD 190/kWh in 2016, so that market-wide penetration in the stationary application market is achieved [18]. With an approximately 7 percent annual decrease in cost per watt-hour between 2007 and 2014 or 60 to 70 percent in total, prices of first-generation lithium-ion batteries are expected to keep on decreasing,

despite the already dramatic lowering these past years along with purchasing costs for EV manufacturers dropping even more [18, 20]. Nykvist et al. [20] have noted that there exists a large price discrepancy amongst suppliers as depicted in Figure 1.1 [20] and have projected a cost reduction rate of around the same magnitude.

FIGURE 1.1: Price development of Li-ion batteries 2005-2030

To become cost-competitive with the traditional vehicle powered by fossil fuel, it is generally agreed that prices of lithium-ion batteries need to fall to around USD 150/kWh, only then the lithium-ion battery-powered electric vehicles can sustain in the market [20]. This means within the transport sector paradigm shift has to happen to see lithium-ion technology more economically choice, which would be followed by the stationary applications at a later stage [20]. In Figure 1.2 [5], gravimetric energy densities and specific power of different battery technologies are presented, and the current state of the most common battery technologies is summarised in Table 1.1.

As can be seen in Figure 1.2, the lithium-ion technology has an advantage in terms of both gravimetric and volumetric energy density over other common battery types. Also, it can be noted that there exists a significant difference in energy density within the category of the lithium-ion batteries itself. As it can be seen in Table 1.1 [6], the lithium-ion battery technology encompasses a diverse group of chemistries.

FIGURE 1.2: Gravimetric Energy Density and Specific Power of Different Available Battery Technologies

TABLE 1.1: Comparison of current battery technologies

	Lead-acid	Ni-Cd	Ni-MH	LiCoO ₂	LiMn ₂ O ₄	LiFePO ₄
Gravimetric energy density (Wh/kg)	30-50	45-80	60-120	150-190	100-135	90-120
Cycle life (Δ SoC = 80 % each cycle)	200-300	1000	300-500	500-1000	500-1000	1000-2000
Fast charge time (h)	8-16	1	2-4	2-4	1 or less	1 or less
Self-discharge per month at room temperature	5 %	20 %	30 %	<10 %	<10 %	<10 %
Nominal cell voltage (V)	2	1.2	1.2	3.6	3.8	3.3
Maintenance requirements	3-6 months	30-60 days	60-90 days	Not required	Not required	Not required
Toxicity	High	High	Low	Low	Low	Low
Peak C-rate (Ah/h)	5 C	20 C	5 C	>3 C	>30 C	>30 C

As observed in Table 1.1 and Figure 1.2, the main advantages of the lithium-ion battery technology over other technologies are its long cycle life, high nominal voltage, low need for maintenance and high energy density.

The critical index to measure or estimate the health condition is the battery's capacity. The capacity is defined as the available energy stored in a fully charged the lithium-ion battery cell can deliver and is measured in Ampere-Hour (Ah) [21, 21]. With the available estimate of the capacity of the battery, the remaining useful life which refers to the available service time left before the capacity fade reaches an unacceptable level [22–24] can be derived. Therefore, it is very critical to accurately estimate and forecast the capacity parameters in order to monitor the present battery State-of-Health (SOH) and Remaining Useful Life (RUL) so as to enable failure prevention through timely and effective maintenance actions. Capacity being dynamic and it depends largely on usage pattern and environmental condition, therefore it is imperative to estimate it correctly and prognosis of the same.

1.2 Objectives

The primary objective of this work is to develop a data-driven model to estimate and forecast the capacity in the lithium-ion battery. Since model-based methods rely heavily on accurate physics-based model or complex signal processing techniques, which require extensive expert involvement, the aim for this work is to build a data-driven framework which will not require much subject matter information in the construction and chemical composition of the lithium-ion battery.

Most data-driven methods are based on the variants of the neural network, which sometimes can be difficult to train or may require different architecture/parameters for different batteries. In this work, the goal was to use the same parameters for different dataset so that the results are not dependent on the architecture of the prediction network.

Lastly, batteries are seldom fully charged/discharge in real-life scenarios, therefore during online capacity estimation, full range of voltage or current measurements may not be available, so the aim for this work is also to develop a framework which can deal with the missing value cases.

1.3 Organization of the Thesis

The organisational structure is as follows. In Chapter 1, the theme for this work is introduced, along with motivation and the historical background of lithium-ion batteries. This is followed by explaining the utmost importance of determining the capacity of the lithium-ion battery. In addition to this, objectives for this work is clearly defined in the introduction.

Chapter 2 provides a comprehensive literature review for the research done in the past for developing methods to estimate capacity. Three different ways to determine capacity are studied, and the merits of each of them are thoroughly studied. By careful studying all the past work, some gaps are identified in the literature and have been tried to solve in this work.

Chapter 3 is titled as “Framework”. In this Chapter, the mathematical framework used in this work is explained, followed by the feature extraction and selection. Since in any data-driven method, features are the backbone, the feature extraction and selection are explained thoroughly. Lastly, the overall system design used in both the offline and the online stages is presented in the chapter.

Dataset used in work and results are discussed in Chapter 4. Three independent battery dataset sourced from different experiments are used to validate the method. In this chapter, first, the dataset description is presented, followed by the results and performance.

Finally, the conclusion for this work, along with some future work suggestion, are discussed in Chapter 5.

Chapter 2

Literature Review

This chapter provides a comprehensive review of the research done on the relevant topics in capacity estimation, along with regression techniques and prognostics methods. Extensive research has been conducted to understand capacity, state of health and predict the battery life. In general, approaches to estimate and forecast are studied and classified in three typical categories; model-based approaches, data-driven approaches and fusion approaches, later one being an ensemble of two former methods.

Model-based are approaches involve the knowledge of a systems failure mechanisms (e.g. crack, growth) to build a mathematical description of the systems degradation process to estimate the status and remaining useful life (RUL). The mathematical model quantitatively characterises a systems behaviour using physics or first principles. The identification of the model parameters usually requires specifically designed experiments and extensive empirical data. In prognostics, condition data are often used to identify and update the model parameter using statistical methods (e.g. regression, Bayesian update). In this thesis, prognostics models which involve first principles and statistical methods for parameter identification are still categorised as model-based approaches.

Based on the mathematical representations of the batteries degradation evolution, model-based filtering methods use variants of particle filters such as extended

Kalman filter (EKF) and unscented Kalman filter (UKF) to estimate the state-of-health (SOH) or remaining useful life of batteries [25, 26]. Hu et al. [27] proposed a particle filtering (PF) technique combined with a kernel smoothing (KS). This particle filtering - kernel smoothing simultaneously estimates the degradation state and the unknown parameters of degrading components, while significantly overcoming the problem of particle impoverishment. Based on the updated degradation model (with unknown parameters replaced by the estimated ones), the RUL prediction is then obtained by simulating future particles evolutions. Miao et al. [28] introduced an improved PF algorithm unscented particle filter (UPF) for the battery remaining useful life prediction. In this approach firstly a PF algorithm and UPF algorithm are presented distinctly, after which a degradation model is constructed based on the understanding of lithium-ion batteries.

Under physics-based performance degradation modelling, Santhanagopalan et al. [29] applied an unscented filter algorithm to model the electrochemical reactions that cause the battery's capacity fade based on a rigorous electrochemical model proposed by Doyle et al. [30]. Safari et al. [31] proposed a multimodal physics-based model that considers an isothermal, multimodal, physics-based aging model for life prediction of lithium-ion batteries, for which solvent-decomposition reaction leading to the growth of a solid electrolyte interphase (SEI) at the carbonaceous anode material is considered as the source of capacity fade. Ning et al. [32] developed a physics-based model employing the idea that lithium deposition, electrolyte decomposition, active material dissolution, phase transition inside the insertion electrode materials, and further passive film formation on the electrode and current collectors can affect the capacity fade of lithium-ion batteries up to different degrees. Their approach was to quantify these degradation processes to improve the predictive capability of battery models and help to elucidate the mechanism of capacity fade.

Park et al. [33] considered various electrical circuits and studied the battery capacity utilisation based on how the battery is discharged and validated the developed models on commercially available lithium coin cell batteries. Similarly, Mehdi et al. [34] considered a battery as a circuit with two RC subnetworks, which represented the fast and slow transient responses of the terminal voltage. In this approach,

the linear part of the model is not observable, and the nonlinear behaviour of the open-circuit voltage versus the state of charge is included in the model. The proposed observer targeted the problems in attaining a reliable estimation of the state of charge, which was used in online parameter estimation of the battery's state of health. Moreover, compared to the methods in which the non-linearities or uncertainties in the model are disregarded, or those terms are discarded using a conventional sliding-mode observer, an analytical method was proposed to estimate the additive nonlinear or uncertainty parameter in the model.

Li et al.[35] proposed a model based on the mixture of the Gaussian process models and particle filter to forecast the state of health estimation of lithium-ion batteries. In this work an integrated approach based on a mixture of the Gaussian process (MGP) model and particle filtering (PF) was presented for lithium-ion battery SOH estimation under uncertain conditions. Instead of directly assuming a specific state-space model for capacity degradation, Li et al. [35] presented that the distribution of the degradation process is learnt from the inputs based on the available capacity monitoring data. Another approach is to measure cell resistance and developed a capacity degradation model. Remmlinger et al. [36] proposed a model by monitoring the increase of the internal resistance of the battery cells observed over the whole lifetime of the battery. After which a mathematical model was deduced from an equivalent circuit which contained the parameter based on the degradation of the battery cell in an electric vehicle. Later again Remmlinger et. al. [37] developed an onboard state-of-health monitoring system of lithium-ion batteries using linear parameter varying models by measuring internal resistance over the short-term and mid-term voltages and the capacity influences deviations in open-circuit voltage (OCV) due to charge and discharge, thus quantifying the state-of-health with the change of the internal resistance.

In general, model-based approaches can well reflect the physical and electrochemical properties of batteries, but it is complicated and challenging to monitor internal state, and accurate physics-based models are challenging to obtain. This issue is faced due to the lack of understanding of the degradation evolution, and it may not be available to build a mathematical model for such complex systems.

In contrast to model-based approaches, when the explicit degradation mechanism is unknown, but enough historical data is available, data-driven methods can be used. This data-driven approach is purely based on the extracted features from observed data, such as voltage, current, capacity and impedance to predict the batteries health conditions [23, 38–40]. As a result, it is highly reliant on not only the quantity but also the quality of historical data. Hu et al. [40] proposed an ensemble approach for the data-driven prognostics with three weighting schemes, the accuracy-based weighting, diversity-based weighting and optimisation-based weighting, were proposed to determine the weights of member algorithms. The k-fold cross-validation was employed to estimate the prediction error required by the weighting schemes, which gave more precise RUL predictions compared to any single algorithm. Wang et al. [23] used a conditional three-parameter capacity degradation model to fit the representative training vectors by using relevance vector machine. Then the extrapolation of the degradation model was employed to estimate the remaining useful life of lithium-ion batteries. The relevance vector machine was used to derive the relevance vectors that can be used to find the representative training vectors containing the cycles of the relevance vectors and the predictive values at the cycles of the relevance vectors. The conditional three-parameter capacity degradation model was then developed to fit the predictive values at the cycles of the relevance vectors. Extrapolation of the conditional three-parameter capacity degradation model to a failure threshold was used to estimate the remaining useful life of lithium-ion batteries.

Nuhic et al. [39] implemented a support vector machine (SVM) to embed diagnosis and prognostics of system health with a goal to estimate the state of health and RUL of lithium-ion batteries. The features used for training included initial capacity value, capacity degradation in Ah between the initial and target capacity measurements, along with SOC and temperature measures. Patil et al. [41] presented a multi-stage SVM approach for RUL prediction of lithium-ion batteries which integrated classification and regression attributes of the support vector (SV) based machine learning technique. Cycling data of lithium-ion batteries under different operating conditions were analysed, and the critical features were extracted from the voltage and temperature profiles. The classification and regression models for RUL were built based on the critical features using SVM. The classification model provided a gross estimation, and the Support Vector Regression (SVR) was

used to predict the accurate RUL if the battery is close to the end of life (EOL).

Liu et al. [42] used the Box-Cox transformation (BCT) to approximate a relationship between the extracted features and capacities of lithium-ion batteries after which they used relevance vector machine (RVM) to predict RUL based on the extracted features. In this work, health indicators (HI) extraction and optimisation framework requiring only the operating parameters of lithium-ion batteries were proposed for battery degradation modelling and RUL estimation. The framework carried out raw HI extraction, transformation, correlation analysis, and verification and evaluation to achieve HI enhancement. In particular, the Box-Cox transformation was adopted to improve the correlation between the extracted HI and the battery's actual degradation state. To estimate the battery's RUL using the enhanced HI, an optimised relevance vector machine algorithm was utilised, which can be performed in a flexible and agile way.

Besides the kernel techniques [39, 41, 42], neural network (NN) techniques were also used to predict the RUL of lithium-ion batteries. Liu et al. [43] presented an adaptive recurrent NN (ARNN) for system dynamic state forecasting to predict the RUL of lithium-ion batteries. In this work, the developed ARNN was constructed based on the adaptive/recurrent neural network architecture, and the network weights were adaptively optimised using the recursive Levenberg-Marquardt (RLM) method. The deficiency of the purely data-driven models is their difficulty in interpreting the model parameters and predicting variables. The approaches which are purely model-based filtering and those which are purely data-driven have their limitations, with the former one relying on the physical model for state prediction and the latter one not accounting for the physical process. A single method is often insufficient in dealing with high non-linearity and varying operating conditions of batteries. Fusion or hybrid approaches have potentials to overcome the deficiencies in a single method and thus proposed to integrate the merits of the two types to overcome the limitations.

Liao et al. [44] categorised hybrid approach for prediction into five types. According to category H4 in Ref. [44] integrating model-based filtering approaches and data-driven approaches for the battery prognostic problem can be divided into

three main categories. The first type is in which a data-driven model is developed that is made to compensate for the physical state or measurement model. Due to the complex nature of the degradation system, it may be very tedious or even impossible to obtain a degradation model. The data-driven model is an alternative to replace the complex physics-based model. Liu et al. [45] adopted a data-driven non-linear degradation autoregressive (ND-AR) model as the observation model for regularised particle filter (RPF) to estimate battery RUL. In this work, firstly, the nonlinear degradation feature of the lithium-ion battery capacity degradation was analysed, and then, the nonlinear accelerated degradation factor was extracted to improve prediction ability of linear AR model followed by an introduction of an optimized nonlinear degradation autoregressive (ND-AR) time series model for remaining useful life (RUL) estimation of lithium-ion batteries. After this, the ND-AR model was used to realise the multi-step prediction of the battery capacity degradation states. Finally, to improve the uncertainty representation ability of the standard PF algorithm, the regularised particle filter was applied to design a fusion RUL estimation framework of the lithium-ion battery. Yan et al. [46] developed a Lebesgue sampling-based method for battery RUL prediction. This work introduced the concept of Lebesgue sampling (LS) in fault diagnosis and prognosis and proposed an LS-based fault diagnosis and prognosis (LS-FDP) framework. In the proposed LS-FDP, a diagnostic philosophy of “execution-only when necessary” was developed to do reduction in computation cost. For prognosis, the proposed approach defines the prognostic horizon on the fault state axis. With a reduced prognostic horizon, the LS-FDP naturally benefited the uncertainty management.

A second type is a data-driven approach in which a future trend is predicted by observing the measurement values. During the prediction period without new observation, the predicted measurements from the data-driven approach is used as new measurements for the filtering-based approach. Liu et al. [47] proposed a data-model fusion approach to improve prediction performance. A data-driven predictor was used to predict the future battery measurements, which incorporated into PF for long-term prediction. In this work, particle filtering was applied for system state estimation based on Bayesian learning in parallel with parameter identification of the prediction model (with unknown parameters). Simultaneously, a data-driven predictor was employed to learn the system degradation pattern from historical data to predict system evolution (or future measurements). An important feature

of the proposed fusion prognostic framework was that the predicted measurements (with uncertainties) from the data-driven predictor would be properly managed and utilised by the particle filtering to further update the prediction model parameters, thus enabling markedly better prognosis as well as improved forecasting transparency. Liao et al. [44] presented two different data-driven models. First one was used as a measurement model to establish the mapping between the battery's internal state and the measurement. The second one was used to predict future measurements which were fed into the PF to predict the battery RUL. Li et al. [48] developed a multistep-ahead prediction model based on the mean entropy and RVM. In this work, a wavelet denoising approach was introduced into the RVM model to reduce the uncertainty and to determine trend information. The mean entropy-based method was then used to select the optimal embedding dimension for correct time series reconstruction. Finally, RVM was employed as a novel non-linear time-series prediction model to predict the future SOH and the remaining life of the battery.

In a third type, a data-driven method is used to estimate and predict the model parameters for the physical-based methods. Saha et al. [49, 50] used relevance vector regression (RVR) to estimate the parameters for the battery model, which were fed into the particle filter (PF) or Rao-Blackwellized particle filter (RBPF) to predict the battery RUL. This work explored how the remaining useful life (RUL) could be assessed for complex systems whose internal state variables were either inaccessible to sensors or hard to measure under operational conditions. Models of electrochemical processes in the form of equivalent electric circuit parameters were combined with statistical models of state transitions, ageing processes, and measurement fidelity in a formal framework. Consequently, a Bayesian statistical approach was applied to indirect measurements, anticipated operational conditions, and historical data. Xing et al. [51] proposed an ensemble model, which fused the regression model and Particle Filter (PF) algorithm, to predict the RUL of the lithium-ion battery. In this work, the regression model was fused with an empirical exponential and a polynomial regression model to track the battery's degradation trend over its cycle life based on experimental data analysis. Model parameters were adjusted online using a particle filtering (PF) approach. Dong et al. [52] proposed a prognostic approach using Support Vector Regression-Particle Filter (SVR-PF) considering an electric circuit with resistance as parameters. In this

work, a method for battery SOH monitoring was developed to analyse the proposed capacity degradation parameters online and build a novel RUL prediction model which was able to update the RUL probability distribution parameters. Moreover, a support vector regression-particle filter (SVR-PF) algorithm was implemented in the research to improve the standard PF against the degeneracy phenomenon. He et al. [22] proposed an empirical model based on the physical degradation behaviour of lithium-ion batteries. In this work, Dempster-Shafer theory (DST) was applied to select the initial model parameters, and then the model parameters were updated using Bayesian Monte Carlo (BMC), and RUL was predicted based on the data available through battery capacity monitoring. In this work, as more data became available, the accuracy of the model in predicting RUL improved.

Existing prognostics methods have produced good results for lithium-ion battery RUL prediction and capacity estimation. However, due to complexity and diversity of lithium-ion batteries, existing methods show some limitations: (1) Most model-based methods rely heavily on accurate physics-based models or complex signal processing techniques, which require extensive expert involvement. (2) Most data-driven methods are based on the variants of neural networks which are very sensitive to the architecture of the model and the parameters will have to be changed according to the dataset. (3) In the case of partially charged/discharged of batteries, the above-listed literature does not explain how to deal with missing values. Understanding these gaps in the literature, a stochastic gradient boosting regression model is used as a base structure which can be trained using fewer data, the hyperparameter tuning is fast and need not be changed according to the dataset. The partial charge/discharge and missing value issue are taken care by carefully selecting features and using an auto-regressive integrated moving average process to deal with the missing value imputation. Overall the whole system is designed to estimate and forecast the capacity as accurately as possible taking all the real-life scenarios into consideration.

Chapter 3

Framework

This section provides the framework for the techniques used for the developing data-driven regression model followed by the statistical techniques used in this work. Section 3.1 describes stochastic gradient boosting regression (SGBR) to build a regression model which is a base structure to estimate capacity; Section 3.2 explains the autoregressive integrated moving average (ARIMA) model which is used for both forecasting and missing value imputation. This is followed by the feature selection and extraction method in Section 3.3. Lastly both the offline or training phase and online or runtime phase is described in Section 3.4.1 and 3.4.2, which presents an overall overview of the full system. To make the system more intelligent and flexible with the data availability, the fuzzy logic system is used in the online phase.

3.1 Stochastic Gradient Boosting Regression

In any machine learning applications, the goal is to build a non-parametric regression or classification model from the data. With the domain-specific knowledge, one way is to design a model based in domain knowledge and then adjust its parameters using the observed data. However, in most cases such analytical model is not available, not even initial expert-driven guesses about the potential relationship between the input and output variables. The lack of a model can be solved if one applies non-parametric machine learning, or any other algorithm like a neural network, support vector machine or any other another algorithm at one's

discretion, to build a data-driven model. These models are built in the supervised manner. This means that the model consists of a pair, an input object (typically a vector) and the desired output value (also called as the target or response variable).

The most popular and frequent approach in data-driven modelling is to build a single string predictive model. However, a different approach would be to build an ensemble of models by combining a large number of relatively weak, simple models to obtain a stronger ensemble prediction. In literature, the most prominent examples of such ensemble machine learning techniques are random forest [53] and neural network ensemble [54], which led to many successful applications in different domains.

The principle idea of boosting is to add new models to ensemble sequentially. At each iteration a base learner model is trained with respect to the errors obtained from the whole ensemble learnt so far. This base learner is a weak learner which in machine learning terminology means an algorithm (for regression/classification) that provides an accuracy slightly better than random guessing. The boosting techniques were purely algorithm-driven during the initial phase of research and development, which made a detailed analysis of their properties and performance rather difficult [55]. This observation led to several speculations as to why this algorithm either outperformed every other method or on the contrary, were inapplicable due to severe overfitting [56]. This criticism led to the formulation of the statistical framework. Thus, a gradient-descent based formulation was derived in 1997 by Freund and Schapire [57], and later improvements were contributed by Friedman [58]. This formulation of boosting methods and the corresponding models were called gradient boosting machines (GBM). This framework also provided the essential justifications for the model hyperparameters and established the methodological base for the further development in gradient boosting modelling. In comparison with the artificial neural networks, which have the memory of the learned patterns distributed within the connections of the artificial neurons, in the boosted ensemble, the base learners play the role of the memory medium. The patterns are captured sequentially and are formed gradually, increasing the level of pattern details at each iteration.

In gradient boosting machines (GBMs), the learning procedure consecutively fits new models to provide a more accurate estimate of the response variable error obtained from the whole ensemble learnt so far. The fundamental idea behind this algorithm is to construct the new base-learners, which are maximally correlated with the negative gradient of the defined loss function, associated with the whole ensemble. The loss functions applied can be arbitrary, but to give a better intuition, if the error function is the classic squared-error loss, then the learning procedure would result in consecutive error-fitting. In general, the choice of the loss function is up to the user, choosing either from a wide variety of loss functions derived so far and or developing one's own loss function. This high flexibility makes the GBMs highly customizable to any data-driven task. The GBMs have shown considerable success not only in practical applications but also in various machine-learning and data-mining challenges [59, 60]. The remaining part of this section presents SGBR algorithm formulation.

3.1.1 Function Estimation

Consider the function estimation problem in which the system consists of “target” or “response” variable y and an “explanatory input” variables $= \{x_1, \dots, x_n\}$ which can be a vector. Given the training set $\{y_i, x_i\}_1^N$ of known (y, x) values, the goal is to reconstruct or find the functional dependence $x \xrightarrow{F} y$. with an estimate $\hat{F}(x)$, such that the loss function $\Psi(y, F)$ is minimised:

$$\hat{F}(x) = y \quad (3.1)$$

$$\hat{F}(x) = \arg \min_{F(x)} \Psi(y, F(x)) \quad (3.2)$$

Expressing the estimation problem as expectations, the equivalent formulation would be to minimise the expected loss function over the target variable $E_y(\Psi[y, F(x)])$, conditioned on the observed explanatory data x .

$$\hat{F}(x) = \arg \min_{F(x)} \underbrace{E_x[E_y(\Psi[y, F(x)]) | x]}_{\text{expectation over the whole dataset}} \quad (3.3)$$

The selection of loss function differs and depends on the application and can be decided upon looking at the distribution of the response variables. Let us assume if the response variable is a binary set, i.e. $y \in \{0, 1\}$, the binomial loss function can be applied. If the response variable is a continuous, i.e. $y \in R$, a classical L_2 squared loss function can be used. For the other response distribution families like Poisson-counts, a specific loss function can be used.

A standard procedure is to make the function estimation tractable; therefore, the function search space can be restricted to the parameterised class of function $F(x, \theta)$. This would change the function optimisation problem into a parameter estimation problem

$$\hat{F}(x) = F(x, \hat{\theta}) \quad (3.4)$$

$$\hat{\theta} = \arg \min_{\theta} E_x[E_y(\Psi[y, F(x, \theta)])|x] \quad (3.5)$$

In general, the closed-form solution for the parameter estimation is not possible. Therefore, to perform the estimation, iterative numerical procedures are considered.

3.1.2 Numerical Optimization

The parameter estimation can be written in the incremental form given the M successive iteration (steps or boosts) steps as:

$$\hat{\theta} = \sum_{i=1}^M \hat{\theta}_i \quad (3.6)$$

Steepest-descent is the most frequently used and the simplest procedure used in numerical minimisation methods. Given N data points $(x, y)_{i=1}^N$, we want to decrease the empirical loss function $J(\theta)$ over this observed data:

$$J(\theta) = \sum_{i=1}^N \Psi(y_i, F(x_i, \hat{\theta})) \quad (3.7)$$

In the classical steepest descent optimisation, the continuous improvements are based along the direction of the gradient of the loss function $\nabla J(\theta)$. As mentioned earlier, the parameter estimation $\hat{\theta}$ is presented iteratively. There is a need to define the notation for a well follow up. The subscript index of the estimates $\hat{\theta}_t$ will be an indicator of the t^{th} iterative step of the estimate $\hat{\theta}$. The superscript index $\hat{\theta}^t$ will correspond to the collapsed estimate of the whole ensemble, i.e., the sum of all the estimates from the initial step till step t . The steps of steepest optimisation are as follows:

1. Initialize the parameter estimate $\hat{\theta}_0$ for each iteration t , repeat:
2. Obtain a compiled parameter estimate $\hat{\theta}^t$ from all of the previous iterations:

$$\hat{\theta}^t = \sum_i^{t-1} \hat{\theta}_i \quad (3.8)$$

3. Evaluate the gradient of the loss function $\nabla J(\theta)$., given the obtained parameter estimates of the ensemble:

$$\nabla J(\theta) = \{\nabla J(\theta_i)\} = \left[\frac{\partial J(\theta)}{\partial J(\theta_i)} \right]_{\theta=\hat{\theta}^t} \quad (3.9)$$

4. Calculate the new incremental parameter estimate $\hat{\theta}_t$:

$$\hat{\theta}_t \leftarrow -\nabla J(\theta) \quad (3.10)$$

5. Add the new estimate $\hat{\theta}_t$ to the ensemble

3.1.3 An Optimisation in Function Space

The principle difference between a traditional machine learning technique and the boosting techniques is that the optimisation is the function or prediction space in the latter, i.e. the function estimate $\hat{F}$ is parametrised in the additive functional form:

$$\hat{F}(x) = \hat{F}^M(x) = \sum_{i=0}^M \hat{F}_i(x) \quad (3.11)$$

M is the number of iteration, $\widehat{F}_0$ is the initial guess and the $\{\widehat{F}_i\}_{i=1}^M$ are the function increments or boosts. In terms of parametrisation the family of function, the base-learner function can also be written as parameterised form as $h(x, \theta)$ to distinguish them from the overall ensemble function estimates $\widehat{F}(x)$. There are different families of base-learners, such as decision trees or splines. In this implementation, the decision tree base learner is used. The idea to use a decision tree is to divide the space of the input variable into homogenous rectangle areas by a tree-based rule system. Corresponding to the if-then rule, each tree is split over some input variables. This structure of a decision tree naturally encodes and model the interaction between predictor variables. These trees are commonly parameterised with a number of splits, or equivalent, the interaction depth. A case of a decision tree with only one split is known as a tree stump or in case of regression, a regression tree stump with a single root and two children or left and right leaf, that splits on a single feature or variable - a threshold. If a test feature value is less than the threshold, the models yield the average of the training target samples in the left leaf. If the test feature value is greater than or equal to the threshold, the model yields the average if the training target samples in the right leaf. In many practical applications, small trees and tree-stump provide considerably accurate results [61]. From the literature, it is evident that even a complex model with a tree structure (interaction depth > 20) provide almost no benefit over a compact tree (interaction depth ~ 5). Hastie et al. [62] commented that typically $4 \leq J \leq 8$ iteration depth works well for boosting and results were relatively insensitive to the number of depths in this range.

The boosting approach, as mentioned before, follows a stagewise method rather than stepwise approaches. For the formulation of this “Greedy Stagewise” approach, an optimal step-size ρ is introduced and should be specified at each iteration. Therefore, the optimisation rule for the t^{th} iteration is defined as:

$$\widehat{F}_t \leftarrow \widehat{F}_{t-1} + \rho_t h(x, \theta_t) \quad (3.12)$$

$$(\rho_t, \theta_t) = \arg \min_{\rho, \theta} \sum_{i=1}^N \Psi(y_i, \widehat{F}_{t-1}) + \rho h(x_i, \theta) \quad (3.13)$$

3.1.4 Gradient Boosting Algorithm

One can specify both the loss function and the base-learner models arbitrarily. In practice, given some specific loss function $\Psi(y, F)$ and/or a custom base-learner $h(x, \theta)$, the solution to the parameter estimates can be very hard. To deal with this issue, it was proposed to choose a new function $h(x, \theta_t)$ to be the most parallel to the negative gradient $\{g_t(x_i)\}_{i=1}^N$ along with the observed data:

$$g_t(x) = E_y \left[\frac{\partial \Psi(y, F(x))}{\partial F(x)} \middle| x \right]_{F(x) = \hat{F}^{t-1}(x)} \quad (3.14)$$

In place of searching the general solution for the boost increment in the function space, one can choose the new function increment to be the most correlated with $-g_t(x)$. This permits the replacement of a potentially arduous optimisation task with the classic least-squares minimisation one:

$$(\rho_t, \theta_t) = \arg \min_{\rho, \theta} \sum_{i=1}^N [-g_t(x_i) + \rho h(x_i, \theta)]^2 \quad (3.15)$$

To summarise, we can formulate the complete form of the gradient boosting algorithm, as initially proposed by Friedman [58]. The exact form of the derived algorithm with all the corresponding formulas will heavily depend on the design

choices of $\Psi(y, F)$ and $h(x, \theta)$.

Algorithm 1: Freidman's Gradient Boosting Regression

Inputs : Input data $(x, y)_{i=1}^N$
 Number of iterations M
 Loss-function $\Psi(y, F)$
 Base-learner model $h(x, \theta)$

Output: $F(x)$

```

1 Initialize  $\hat{f}_0$  with a constant value ;
2 for  $t = 1$  to  $M$  do
3   Compute the negative gradient  $g_t(x)$ ;
4   Fit a new base-learner function  $h(x, \theta_t)$ ;
5   Find the best gradient descent step-size  $\rho_t$ ;
    $\rho_t = \arg \min_{\rho} \sum_{i=1}^N \Psi(y_i, \hat{F}_{t-1}(x_i) + \rho h(x_i, \theta_t))$  ;
6   Update the Function Estimate;
    $\hat{F}_t \leftarrow \hat{F}_{t-1} + \rho_t h(x, \theta_t)$ ;
7 end
```

Another improvement suggested by Friedman [63] was to introduce stochastic nature while building in the iteration. So at each iteration, a sub-sample of the training data is drawn at random (without replacement) from the full training dataset.

The random sampling $\{\pi(i)\}_i^N$ is calculated as a random permutation of integers $\{1, \dots, N\}$ given training set as $\{y_i, x_i\}_1^N$. Then a random subsample of size $\tilde{N} < N$ is given by $\{y_{\pi(i)}, x_{\pi(i)}\}_1^{\tilde{N}}$ which is then used instead of the full sample, to fit the base learner. Together the learning rate and sub-sample ratio are tuned as a part of hyperparameter to achieve better results.

The algorithm for this implementation is summarised in following pseudo-code Algorithm 2.

Algorithm 2: Stochastic Gradient Boosting Regression

Inputs : Input data $(x, y)_{i=1}^N$

Number of iterations M

Loss-function $\Psi(y, F)$

Base-learner model $h(x, \theta)$

Sub-sampling ratio

Output: $F(x)$

```

1 Initialize  $\hat{f}_0$  with a constant value ;
2 for  $t = 1$  to  $M$  do
3   Random Sampling of training data;
    $\{\pi(i)\}_i^N = \text{rand\_perm}\{i\}_i^N$ ;
4   Compute the negative gradient  $g_t(x)$ ;
5   Fit a new base-learner function  $h(x, \theta_t)$ ;
6   Find the best gradient descent step-size  $\rho_t$ ;
    $\rho_t = \arg \min_{\rho} \sum_{i=1}^N \Psi(y_i, \hat{F}_{t-1}(x_{\pi(i)}) + \rho h(x_{\pi(i)}, \theta_t))$  ;
7   Update the Function Estimate;
    $\hat{F}_t \leftarrow \hat{F}_{t-1} + \rho_t h(x, \theta_t)$ ;
8 end
```

Apart from being used to build the regression model, another use of gradient boosting is that after the trees are constructed, it is relatively straightforward to retrieve the importance of each feature [64, 65]. Feature importance is an indicative of how useful each feature was in the construction of the boosted regression tree within the model. The information gained can be used to select only those features which are very significant to model, and the rest of the features can be discarded without losing much accuracy. In general sense, the more an attribute is used to make key decisions in constructing a decision tree, the higher will be its relative importance. Based on how each attribute split point improves the performance measure, importance is calculated for a single decision tree. Feature extraction/selection and importance are explained in Section 3.3

3.2 Auto-Regressive Integrated Moving Average

An autoregressive integrated moving average (ARIMA) is a statistical model used for forecasting a time series data first popularized by Box and Jenkins [66]. It is a generalisation of the autoregressive moving average model and adds the notion of integration. ARIMA models are applied in the cases when the nature of the data is non-stationary, where an initial differencing step is applied one or more times to eliminate non-stationarity [67]. There are three parts to the ARIMA model, first is the autoregressive (AR) part, second is integrated (I) part, and last is the moving average (MA) part. An ARIMA model, denoted as $\text{ARIMA}(p, d, q)$ is used, where p is the AR order, d is the I order, and q is the MA order.

In order to explain the mathematical model of ARIMA, we must first introduce the back-shift operator B . The definition of B makes the mathematical model more elegant and understandable in explaining general cases. The backshift operator B is an operator when multiplied with the time series/index-based observations results in it to shift backwards by one period. Let us consider Y as a sequence series with period t and using B we can express,

$$BY_t = Y_{t-1} \quad (3.16)$$

Similarly, multiplication of a higher power of B correspondingly yields the backward shift by the raised power.

$$B^n Y_t = Y_{t-n} \quad (3.17)$$

ARIMA solves the non-stationarity issue by introducing the notion of the difference, ironically termed as Integrated. Using the definition of B , the difference can also be expressed in terms of B . First, let consider y as the first difference of series Y , then for any time period t

$$y_t = Y_t - Y_{t-1} = Y_t - BY_t = (1 - B)Y_t \quad (3.18)$$

Similarly, the second difference series of Y and a general difference can be denoted using B as follows.

$$z_t = y_t - y_{t-1} = (1 - B)y_t = (1 - B)((1 - B)Y_t) = (1 - B)^2 Y_t \quad (3.19)$$

where z_t is the second difference of Y time series.

$$d_t = (1 - B)^d Y_t \quad (3.20)$$

where d_t is the d^{th} difference of Y time series.

Using the developed definitions and understanding, let us first formulate the $ARIMA(1, 1, 1)$ mathematically, after which the generalized expression can be easily deduced. In order to explain $ARIMA(1, 1, 1)$, the following equations are defined first.

$$y_t = Y_t - Y_{t-1} \quad (3.21)$$

$$y_t = \phi_1 y_{t-1} - \theta_1 \epsilon_{t-1} + \epsilon_t \quad (3.22)$$

where Y_t is the observed time/index-based series, y_t is the first difference as explained before, ϕ and θ are model coefficients of AR and MA part respectively and ϵ_t is a random shock (noise) occurring at time t . ϕ and θ values are being obtained by using a numerical optimisation technique like curve fitting, on the observed time-series data.

Eq. (3.21) and (3.22) can also be expressed in terms of B in the following manner.

$$y_t = (1 - B)Y_t \quad (3.23)$$

$$y_t = \phi_1 B y_t + \epsilon_t - \theta_1 B \epsilon_t \quad (3.24)$$

By rearranging terms in Eq. (3.23) and substituting it in Eq. (3.24), the $ARIMA(1, 1, 1)$ can be summarised only in terms of Y_t and ϵ_t as :

$$(1 - \phi_1 B)(1 - B)Y_t = (1 - \theta_1 B)\epsilon_t \quad (3.25)$$

Similarly solving further and going up in order, the general case $ARIMA(p, d, q)$ can be expressed as follows,

$$(1 - \sum_{i=1}^p \phi_i B^i)(1 - B)^d Y_t = (1 - \sum_{j=1}^q \theta_j B^j)\epsilon_t \quad (3.26)$$

The next step in modelling is to select the order of ARIMA. As described by Box et al. [66], this step is called Identification. One of the prerequisites for the ARIMA process is that the sequence series should be stationary in nature. In the case of non-stationarity of the time series, Box et al. [66] introduced differencing and using logarithm operation on time-series values to ensure stationarity. First, the order d is chosen by evaluating the stationarity.

Secondly, the most important part of modelling identification is a selection of the AR and MA order. By plotting graphs for the autocorrelation function (ACF) and partial autocorrelation function (PACF), the AR and MA order is determined [68]. Alternatively, the order can also be determined using grid search and selecting those values which result in the most negative AIC [69]. In this work, ACF and PACF plots are being used for selection of the order, and the results are presented in Chapter 4. The ACF plot depicts the magnitude of correlation of the observed time or index-based series with lagged value, and the PACF plot summarises the correlations for observation with lag values that is not accounted for by prior lagged observations. The value of p and q are selected from ACF and PACF plots, respectively by observing the intersection of horizontal lines drawn at 95% confidence interval values on the plot. The plots that cross the confidence interval are worth noting, i.e. more significant. After identification of the order, the calculation of AR and MA coefficients is done using numerical optimisation techniques like curve fitting.

Later time periods such as $t + 1$, for which data is not yet available, the auto-regressive equation is used to make the forecast. The auto-regression can be used to forecast an arbitrary number of periods into the future. For the period t , the

first period at which data is not yet available, the values of Y one period before t is used to forecast $\hat{Y}_t$. Preceding values are used in the auto-regression equation while setting the error term ϵ_t equal to zero (because we forecast Y_t to equal its expected value, and the expected value of the unobserved error term is zero).

Therefore, the output of the autoregressive equation is the forecast for the first unobserved period. For further iteration for which the data is not observed, the autoregressive equation is used again to make the forecast, with one difference that the value of Y one period before the one being forecast now is not known, so its expected value - the predicted value arising from the previous forecasting step - is used instead. Similarly, for the later periods, the same procedure is used, each time using one more forecast value on the right-side of the predictive equation until, after setting h predictions, all h right-side values are predicted from preceding steps.

Along with the forecasted value, the forecast intervals (confidence intervals for forecasts) for ARIMA models are also calculated. The confidence intervals are based on the assumptions that the residuals are uncorrelated and normally distributed. If either of these assumptions does not satisfy, then the forecast intervals may be incorrect. For the unobserved time period h the 95% confidence interval, the distribution of the forecasted value is calculated as,

$$95\% \text{ Forecast Interval : } \hat{y}_{T+h|T} \pm 1.96\sqrt{v_{T+h|T}} \quad (3.27)$$

where $v_{T+h|T}$ is the variance of $y_{T+h}|y_1, \dots, y_T$

When $h = 1$, the variance is $v_{T+h|T} = \hat{\sigma}^2$ for all ARIMA models regardless of parameters and orders. For ARIMA(0,0,q),

$$y_t = \epsilon_t + \sum_{i=1}^q \theta_i \epsilon_{t-i} \quad (3.28)$$

$$v_{T+h|T} = \hat{\sigma}^2 \left[1 + \sum_{i=1}^{h-1} \theta_i \epsilon_{t-i} \right], \text{ for } h = 2, 3, \dots \quad (3.29)$$

In general, as the forecast horizon increases the forecast intervals also increases in ARIMA models.

3.3 Feature Extraction and Selection

In all of the data-driven approaches, feature extraction and selection is not only necessary but very important to develop any machine learning model. This section presents a detailed overview on how features are extracted and selected. First and foremost is the extraction of the target or response variable used for the training of the degradation model. As mentioned earlier in Section 1.1, the only way to know the “actual capacity” or “true capacity” of the lithium-ion battery is to use Coulomb counting [70] is the discharging phase of the battery use. The dataset used in this work have reference discharging cycles to measure the true capacity at certain intervals to determine the accuracy of the method. During these reference cycles, batteries are discharged from 100% SOC to 0% SOC to calculate the total Ampere-hour (Ah) that the battery would have been able to deliver in that cycle. Coulomb counting is defined as an integration of the active flowing current (measured in Amps) over time to derive the total sum of energy or leaving (discharging) or entering (charging) the battery pack.

$$Capc = \int_{t_A}^{t_B} I(t)dt \quad (3.30)$$

where t_A is initial time measurement, which starts when the battery is at 100% SOC, and t_B is the final time measurement, which is when the battery reaches 0% SOC. The “actual capacity” or “true capacity” is used as the target or response variable to train the data-driven model.

Since our objective is to predict cyclic ageing, it can be stated that as the battery goes through the cycle of charge and discharge, the battery ages resulting in the decrease of the capacity over time. Keeping this in mind, the critical input feature developing a degradation model is the charge cycle number “*cycle_number*”. It should be noted that one of the objectives is to estimate the capacity in the partial charge/discharge cases, during which the depth of discharge (DOD) will

not be 100%, i.e. discharging will not occur from full charge battery to lowermost threshold. Thus the cycle number is updated on pro-rata basis in the online system, i.e. dependant on the DOD%.

Typically, lithium-ion batteries follow a constant current - constant voltage charging protocol (CC-CV). This method limits the amount of current to a pre-set level until the battery reaches a pre-set voltage level. The current is then reduced as the battery becomes fully charged. The same charging protocol was followed in the all of the experimental dataset. This system allows fast charging without the risk of over-charging. The typical charging curve of the lithium-ion battery is presented in Figure 3.1

FIGURE 3.1: Typical Charging Curve for Li-ion Battery Operated under CC-CV Protocol

During exploratory data analysis, a trend was observed in the charge cycle for voltage and current measurements as battery ageing progresses. In can be seen in Figure 3.2 that there is leftwards movement in the voltage charge curve profile with the increase in the charge cycle number observed in the constant current phase. Similarly, in Figure 3.3, the trend can be observed in the current charge curve profile with the increase in the charge cycle number in the constant voltage phase.

Thus for feature extraction, the full voltage curve measurement in the CC phase and current curve measurement in the CV phase were extracted along with the

FIGURE 3.2: Trend in Voltage Charge Curve as Li-ion Battery Ages

FIGURE 3.3: Trend in Current Charge Curve as Li-ion Battery Ages

time measurements. Then the time difference between consecutive voltage and current measurement were calculated at every increment of $0.01V$ and $0.01Amp$ respectively. These features are denoted as ΔT_V and ΔT_I and in generic form presented as equation 3.31 and 3.32:

$$\Delta T_{V_{ab}} = t_{V_a} - t_{V_b} \quad (3.31)$$

$$\Delta T_{Icd} = t_{I_c} - t_{I_d} \quad (3.32)$$

where t_{V_a} is the time measurement when the terminal voltage is equal to V_a and t_{V_b} is the time measurement when the terminal voltage is equal to V_b with the condition that $V_b - V_a = 0.01V$. Similarly t_{I_c} is the time measurement when the terminal current is equal to I_c and t_{I_d} is the time measurement when the terminal current is equal to I_d with the condition that $I_c - I_d = 0.01Amp$. The range of voltage and current sweep will depend on the battery specification, but the method for extraction remains the same.

As listed in the objectives, the goal is to estimate the capacity irrespective of full or partial charge/discharge case. So, developing a model using all the features extracted will not make sense as the full observation will not be available in an online scenario. To solve this problem, the importance of features is taken into consideration using gradient boosting, as explained in Section 3.1.4.

The feature selected will be different for batteries with different specifications. In this work, three different dataset are used, so the feature extracted will be different for each of them, but the methodology will remain the same. As mentioned in Section 3.1, gradient boosting in addition to build prediction models can also provide the importance of each feature. The idea of feature importance to select only that portion of the charge curve, which is more important in the construction of regression trees. More an attribute is used to make critical decisions in constructing a decision tree; the higher will be its relative importance. The index for the feature here is called F-Score, which is the number of times a feature is used to split the data across all trees.

In this implementation, four sets of time windows consisting of seven consecutive time difference features are selected, two belonging to CC stage and the other two CV stage, which are relatively important than other features. For the sake of simplicity, these four sets of time windows containing seven data points, are each termed as Set_{V1} , Set_{V2} , Set_{I1} , Set_{I2} . Selection of the four sets of windows is made keeping in mind that they belong to different SOC ranges. These different windows are selected to fulfil the objective as and if one or more than one window

set is not observed in partial charge scenario.

For prognosis and missing value imputations, the features are slightly modified, the logarithm of combined the time difference of the full window, i.e. time difference between the initial point of the window and the final point of the window. This is represented as $\log\Delta T_V$ and $\log\Delta T_I$

$$\log\Delta T_{V_{a'b'}} = \log(t_{V_{a'}} - t_{V_{b'}}) \quad (3.33)$$

$$\log\Delta T_{I_{c'd'}} = \log(t_{I_{c'}} - t_{I_{d'}}) \quad (3.34)$$

where $t_{V_{a'}}$ is the time measurement when the terminal voltage is equal to $V_{a'}$ and $t_{V_{b'}}$ is the time measurement when the terminal voltage is equal to $V_{b'}$ with the condition that $V_{b'} - V_{a'} = 0.07V$. Similarly $t_{I_{c'}}$ is the time measurement when the terminal current is equal to $I_{c'}$ and $t_{I_{d'}}$ is the time measurement when terminal current is equal to $I_{d'}$ with the condition that $I_{c'} - I_{d'} = 0.07Amp$. All these time difference features are used to build respecting regression model and will be stated towards the end of this Section. Again for the sake of simplicity these data-points are termed as $\log_Set_{V1}$, $\log_Set_{V2}$, $\log_Set_{I1}$ and $\log_Set_{I2}$

Apart from the time difference feature, two more features are extracted: total energy charged and total energy discharged using coulomb counting. During charging, the amount of energy charged is calculated, and similarly, during the discharging, the amount of energy charged is calculated using Eq. (3.30). For the subsequent charge cycles, the same process is repeated, and the calculated value is subsequently added to previous energy charged or discharged respectively. Therefore, at the end of each cycle, we will have the total energy charged, and the total energy discharged values till that cycle. These features are denoted as “*Tot_en_Cha*” and “*Tot_en_Discha*”.

For prognosis, the idea is to forecast the features for the un-observed cycles rather than the target value (capacity) itself. Then the forecasted values for the features along with the own distribution will be fed into the regression model to forecast the capacity for the future cycles along with the distribution. ARIMA is

explained in Section 3.2 is used to forecast the features by observing the previous features. Also, as mentioned in Section 3.2, to achieve the stationarity of time or index-based series, the logarithm operator is used. Therefore the modified feature are used for forecasting using ARIMA are $\log_{Tot_en_Cha}$, $\log_{Tot_en_Discha}$, $\log_{Set_{V1}}$, $\log_{Set_{V2}}$, $\log_{Set_{I1}}$ and $\log_{Set_{I2}}$ i.e. a total of 6 index-based series.

In the online system, there is a probability that one or more windows selected is not observed. So, if the regression model needed all the four windows, it would not be able to give a better estimate. To solve this issue, different regression models are built combined with different windows and fuzzy logic rules [71] consisting of IF-Then condition as to select respective model given the observation. These rules are presented in in Appendix A. Further, the input and output feature space are summarized in Table 3.1 for the respective prediction and forecasting models along with description is presented in next Section 3.4

3.4 System Design

This section gives a detailed view of the complete system design implemented. The system design consists of two phases. Offline, i.e. training phase and followed by online, i.e. testing or runtime phase.

3.4.1 Offline / Training Phase

In the offline phase, features as described in Section 3.3 are obtained from training. The features include the time difference features, total energy charged and discharged, cycle number as input feature space and true capacity as the response variable. Next, the time difference features are selected according to the feature importance, as explained in Section 3.3. This is followed by the training of the regression model using the stochastic gradient boosting regression algorithm. As mentioned in the previous section, various regression models are built taken into consideration all the permutation and combination of the observed time window. All these feature space and regression models are summarised in Table 3.1 along with the description. For naming purposes, the regression model is termed as

$SGBR_i$, where i is just the iteration for the different Stochastic Gradient Boosting Regression model.

TABLE 3.1: Feature Space for different SGBR Models

Model	Input Feature	Target Value	Description
$SGBR_1$	$Set_{V1}, Set_{V2}, Set_{I1}, Set_{I2},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	All the four time windows, total energy charged/discharged along with cycle number are used
$SGBR_2$	$Set_{V1}, Set_{V2}, Set_{I1},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	Two CC and one CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_3$	$Set_{V1}, Set_{V2}, Set_{I2},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	Two CC and one CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_4$	$Set_{V1}, Set_{V2},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	Two CC stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_5$	$Set_{V1}, Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CC stage time window, total energy charged/discharged along with cycle number are used
$SGBR_6$	$Set_{V2}, Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CC stage time window, total energy charged/discharged along with cycle number are used

$SGBR_7$	$Set_{V1}, Set_{I1}, Set_{I2},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CC and two CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_8$	$Set_{V2}, Set_{I1}, Set_{I2},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CC and two CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_9$	$Set_{I1}, Set_{I2},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	Two CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_{10}$	$Set_{I1}, Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_{11}$	$Set_{I2}, Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_{12}$	$Set_{V1}, Set_{I1},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CC stage and one CV stage time windows, total energy charged/discharged along with cycle number are used

$SGBR_{13}$	$Set_{V1}, Set_{I2},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CC stage and one CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_{14}$	$Set_{V2}, Set_{I1},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CC stage and one CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_{15}$	$Set_{V2}, Set_{I2},$ $Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	One CC stage and one CV stage time windows, total energy charged/discharged along with cycle number are used
$SGBR_{16}$	$Tot_en_Cha,$ $Tot_en_Discha,$ $cycle_number$	$Capc$	Total energy charged/discharged along with cycle number are used

For forecasting of features, all the features mentioned in Section 3.3 are fed into ARIMA identification and estimation process as explained in Section 3.2. The parameters learned are then stored for the respective ARIMA models. All the features that are used for forecasting are summarised in Table 3.2. For naming purposes, the ARIMA model is termed as $ARIMA_i$, where i is just the iteration for a different model.

TABLE 3.2: Index-Based Series used for ARIMA

Model	Index-Based Series
$ARIMA_1$	$\log_Set_{V1}$
$ARIMA_2$	$\log_Set_{V2}$
$ARIMA_3$	$\log_Set_{I1}$
$ARIMA_4$	$\log_Set_{I2}$
$ARIMA_5$	$\log_Tot_en_Cha$
$ARIMA_6$	$\log_Tot_en_Discha$

Since we are forecasting the features in unobserved time period instead of capacity itself, correlation is required to develop the mapping from input feature space and Capacity. For this again the stochastic gradient boosting regression model and its details are presented in Table 3.3.

TABLE 3.3: Feature Space for SGBR model used for Prognosis

Model	Input Feature	Target Value	Description
$SGBR_{17}$	$\log_Set_{V1}, \log_Set_{V2},$ $\log_Set_{I1}, \log_Set_{I2},$ $\log_Tot_en_Cha,$ $\log_Tot_en_Discha,$ $cycle_number$	$\log_Capc$	All the four time windows, total energy charged/discharged along with cycle number are used

Therefore, a total of seventeen different regression models and parameters for six different $ARIMA$ models are learned in the off-line phase. These will be used in runtime operation. The off-line phase is also summarised in Figure 3.4

FIGURE 3.4: Offline or Training Phase System Design Chart

3.4.2 Online / Runtime Phase

In an Online or runtime phase, all the features as mentioned in Section 3.3 are extracted during the charging and discharging cycle. After which the validity of data is checked. Here validity means completeness of data of features required for capacity estimation. The respective SGBR model is used as per the availability

of features, and all the condition are listed in Appendix A. If either of the time difference features are missing, then the one-step ahead *ARIMA* forecasting is done to complete the missing value for the current cycle. After missing value imputation, the database is updated comprising of both past and current data. The future prediction data values are fed into the respective *ARIMA* models as defined in Section 3.4.2 along with the user-defined $\widehat{cycle_number}$ ahead charge cycle for which the predictions are required. The *ARIMA* model gives the foretasted along with confidence interval (C.I.), and for this implementation, 95% C.I. is chosen. The forecasted feature data with C.I. distribution is then fed into *SGBR*₁₆ which intrinsically produces the forecasted estimated capacity along with distribution for $\widehat{cycle_number}$ ahead of the unobserved charge cycle. The online phase is also summarised in Figure 3.5.

FIGURE 3.5: Online or Runtime Phase Flow Chart

Chapter 4

Dataset and Results

In this work, three independent dataset sourced from different experiments are used, and the details of each one are presented in this chapter. Each of the dataset will be presented one by one, followed by the results for each case separately. Feature importance followed by ARIMA parameter extraction will be described. Finally results for the capacity estimation and prognosis are presented. The computer programming to implement the method is written in 'Python' language [72] along with using 'XGBoost' library [65] for SGBR modelling and 'statsmodel' library [73] for ARIMA modelling.

4.1 NASA's Battery Dataset

In this dataset, a set of four Li-ion batteries (B0005, B0006, B0007 and B0018) [74] were run through two different operational profiles, i.e. charge and discharge at room temperature. Charging was carried out in a constant current (CC) mode at 1.5A until the battery voltage reached 4.2V and then continued in a constant voltage (CV) mode until the charge current dropped to 20mA. The discharge was carried out at a CC level of 2A until the battery voltage fell to 2.7V, 2.5V, 2.2V and 2.5V for batteries B0005, B0006, B0007 and B0018 respectively. Repeated charge and discharge cycles result in accelerated ageing of the batteries. The experiments were stopped when the batteries reached the end-of-life (EOL) criterion, which was a 30% fade in rated capacity (from 2Ah to 1.4Ah). Details of experimental dataset is given in Table 4.1.

TABLE 4.1: Details of Experimental NASA's Battery Dataset

Cell Num.	Cutoff voltage(V)	Discharge Current	Temperature($^{\circ}$ C)
B0005	2.7	2A @CC	24
B0006	2.5	2A @CC	24
B0007	2.2	2A @CC	24
B0018	2.5	2A @CC	24

For training and testing, the battery dataset is used in round-robin order since there are limited number of batteries i.e. three battery dataset are used for training, and the remaining one is used for testing. This is followed for every possible combination.

As per the system design which was explained in Section 3.4, the data is processed and features are extracted in the training stage. Then as per the feature importance, features are extracted and the respective models are trained. As mentioned in Chapter 1, one of the objectives is to use the standard or same hyper-parameter for the regression model, so that the model with the same hyper-parameter could be used irrespective of the selection of the dataset. The traditional way of performing hyperparameter optimization is grid search, or a parameter sweep, which is simply an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm. After grid-searching following hyper-parameter were found for stochastic gradient boosting regression models: *tree - depth* = 4, *sub - samplingratio* = 0.8; *loss - function* = Mean Absolute Error. For training, number of iterations sufficiently high number = 100000 is selected, with the early stopping criteria that when the error function is not showing any improvement, stop the further iteration and save the respective model.

4.1.1 Feature Selection

As mentioned in section 3.3, in addition to building prediction models, SGBR can also provide the importance of each feature. The idea of feature importance to select only that portion of the charge curve, which is more important in the construction of tress. The index to measure feature importance is called F-Score, which is the number of times a feature is used to split the data across all trees.

More an attribute is used to make key decisions in constructing a decision tree; the higher will be its F-Score. The feature importance for NASA's Battery Dataset is presented in Figure 4.1 and 4.2 belonging to CC and CV mode respectively.

FIGURE 4.1: Feature Importance observed in CC mode for NASA's Battery Dataset

FIGURE 4.2: Feature Importance observed in CV mode for NASA's Battery Dataset

Features are selected, which are relatively more important than other features and also belong to two different ranges of SOC. As clearly observed from Figure 4.1

and 4.2, Set_{V1} comprises the voltage range from $3.85V$ to $3.92V$, Set_{V2} comprises the voltage range from $4.06V$ to $4.13V$, Set_{I1} comprises of the current range from $1.13A$ to $1.20A$ and finally Set_{I2} comprises of the current range from $0.91A$ to $0.98A$. The voltage and current measurements that doesn't fall in these ranges are not used offline as well as online phase for this dataset.

4.1.2 ARIMA Order Identification

Forecasting features listed in Section 3.3 is done using ARIMA process. To identify the order of the ARIMA process, two functions namely autocorrelation function (ACF) and partial autocorrelation function (PACF). ACF is a measure of the correlation between the series and with a lagged version of itself. PACF is the correlation between series with lagged versions of itself. A partial autocorrelation is the amount of correlation between a variable and a lag of itself that is not explained by correlations at all lower-order-lags.

Along with correlation measure of correlation ACF and PACF plots, the two dashed lines on either sides of 0 are plotted. These lines represent the confidence interval of 95%. From the ACF plot, AR order i.e. p is the lag value where the PACF chart crosses the upper confidence interval for the first time. Similarly from the PACF plot, MA order i.e. q is the lag value where the ACF chart crosses the upper confidence interval for the first time. Plot for the modified features $\log_Set_{V1}$, $\log_Set_{V2}$, $\log_Set_{I1}$, $\log_Set_{I2}$, $\log_Tot_en_Cha$, $\log_Tot_en_Discha$ for B0005 are presented in Figure 4.3, 4.4, 4.5, 4.6, 4.7 and 4.8 respectively.

FIGURE 4.3: ACF and PACF plot for $\log_Set_{V1}$ observed in B0005 FIGURE 4.4: ACF and PACF plot for $\log_Set_{V2}$ observed in B0005

FIGURE 4.5: ACF and PACF plot for $\log_Set_{I1}$ observed in B0005

FIGURE 4.6: ACF and PACF plot for $\log_Set_{I2}$ observed in B0005

FIGURE 4.7: ACF and PACF plot for $\log_Tot_en_Discha$ observed in B0005 FIGURE 4.8: ACF and PACF plot for $\log_Tot_en_Cha$ observed in B0005

By observing the plots for $\log_Set_{V1}$, $\log_Set_{V2}$, $\log_Set_{I1}$, $\log_Set_{I2}$, $p = 1$ and $q = 1$. Similarly, for $\log_Tot_en_Cha$ and $\log_Tot_en_Discha$, $p = 10$ and $q = 1$,

with d chosen as 1 for all the features.

Therefore, for the modified features $\log_Set_{V1}$, $\log_Set_{V2}$, $\log_Set_{I1}$, $\log_Set_{I2}$, the chosen ARIMA model order (p, d, q) is $(1, 1, 1)$ and for $\log_Tot_en_Cha$, $\log_Tot_en_Discha$ the chosen ARIMA model order (p, d, q) is $(10, 1, 1)$.

4.1.3 Capacity Estimation

Results for this dataset is presented in this section. Plots of the actual capacity and the estimated capacity values in Ah versus the cycle number are presented. Error metrics for each of the battery is presented in terms of mean square error (MSE) and mean absolute percentage error (MAPE). As mentioned in the start of this section, training and testing are done in a cyclic order, i.e. three battery's data are used for training and remaining one for testing. Therefore, if the results for B0007 is presented that means B0005, B0006 and B0018 battery's data were used for training. Results are presented in Figure 4.9, 4.10, 4.11 and 4.12 and the detailed results with error metrics are presented in Table 4.2.

FIGURE 4.9: Capacity Estimation for B0018 plotted against Cycle Number

FIGURE 4.10: Capacity Estimation for B0007 plotted against Cycle Number

FIGURE 4.11: Capacity Estimation for B0006 plotted against Cycle Number

FIGURE 4.12: Capacity Estimation for B0005 plotted against Cycle Number

TABLE 4.2: Capacity Estimation Results for NASA's Battery Dataset

Battery	MSE	MAPE(%)
B0005	0.0016546111	3.3695449839
B0006	0.0040435104	5.4896759184
B0007	0.0040741104	5.3948339532
B0018	0.0048289553	6.1593019205

As it can be noticed from Figure 4.10 and 4.12, there is a dip in the performance for capacity estimation at the last charging cycle. This is due to the data corruption that can be observed during the the last charge cycle [74]. Please note that this datapoint is not used in the calculation of error metrics.

4.1.4 Capacity Prognosis

The prognosis results for the battery B0018 are presented in Figure 4.13, 4.14, 4.15, 4.16. The plot of the actual capacity and prognosis of the capacity for the unobserved cycle at different cycle number along with the distribution of the result

are presented. It can be observed, as the battery ages, more historical data is available, so the distribution of the result is getting more narrower and thus the error is also minimizing. For the rest of the batteries, the results with error metrics are presented in Table 4.3.

FIGURE 4.13: Capacity Prognosis for B0018 forecasted at 30th Cycle Number

FIGURE 4.14: Capacity Prognosis for B0018 forecasted at 60th Cycle Number

FIGURE 4.15: Capacity Prognosis for B0018 forecasted at 90th Cycle Number

FIGURE 4.16: Capacity Prognosis for B0018 forecasted at 120th Cycle Number

TABLE 4.3: Capacity Prognosis Results for NASA's Battery Dataset

Battery	Prognosis @	MAPE(%)
B0005	30 th Cycle	8.75045
	60 th Cycle	6.72712
	90 th Cycle	4.16714
	120 th Cycle	3.36785
	150 th Cycle	2.62516
B0006	30 th Cycle	7.71056
	60 th Cycle	6.57439
	90 th Cycle	4.57153
	120 th Cycle	3.78132
	150 th Cycle	2.94711
B0007	30 th Cycle	9.44026
	60 th Cycle	7.35419
	90 th Cycle	5.81199
	120 th Cycle	3.58955
	150 th Cycle	1.81592
B00018	30 th Cycle	3.83896
	60 th Cycle	3.58760
	90 th Cycle	2.32623
	120 th Cycle	1.43307

4.2 CALCE's CX2 Lithium-ion Cells Dataset

This dataset comprise of 6 Li-ion batteries (CX2_33, CX2_34, CX2_35, CX2_36, CX2_37, CX2_38) [22]. All CX2 cells underwent the same charging profile which was a standard constant current/constant voltage protocol with a constant current rate of 0.5C until the voltage reached 4.2V and then 4.2V was sustained until the charging current dropped to below 0.05A. The discharge cut off voltage for these batteries was 2.7V. Repeated charge and discharge cycles result in accelerated ageing of the batteries. The capacity rating for each cell is 1350 mAh. Details of experimental dataset is given in Table 4.4.

TABLE 4.4: Details of Experimental CALCE's CX2 Battery Dataset

Cell Num.	Cutoff voltage(V)	Charge current Rate	Temperature(deg C)
CX2_33	2.7	0.5C	24
CX2_34	2.7	0.5C	24
CX2_35	2.7	0.5C	24
CX2_36	2.7	0.5C	24
CX2_37	2.7	0.5C	24
CX2_38	2.7	0.5C	24

For training and testing, the battery's data is used in round-robin order since there are limited number of batteries, i.e. six battery datasets are used for training, and the remaining one is used for testing. This is followed for every possible combination. The hyper-parameter for stochastic gradient boosting regression models are kept the same as mentioned in Section 4.1

4.2.1 Feature Selection

The feature importance for CALCE's Battery Dataset is presented in Figure 4.17 and 4.18 belonging to CC and CV mode respectively.

FIGURE 4.17: Feature Importance observed in CC mode for CALCE's CX2 Battery Dataset

FIGURE 4.18: Feature Importance observed in CV mode for CALCE's CX2 Battery Dataset

Features which are relatively more important than other features are selected. They belong to two different ranges of SoC. As clearly observed from Figure 4.17 and 4.18, Set_{V1} comprises of the voltage range from 3.85V to 3.92V, Set_{V2} comprises the voltage range from 4.11V to 4.18V, Set_{I1} comprises the current range from 0.67A to 0.68A and finally Set_{I2} comprises the current range from 0.48A to 0.55A. The voltage and current measurements that doesn't fall in these ranges are not used offline as well as online phase for this dataset.

4.2.2 ARIMA Order Identification

Plots for the modified features $\log_Set_{V1}$, $\log_Set_{V2}$, $\log_Set_{I1}$, $\log_Set_{I2}$, $\log_Tot_en_Cha$, $\log_Tot_en_Discha$ for CX2_33 are presented in Figure 4.19, 4.20, 4.21, 4.22, 4.23 and 4.24 respectively.

FIGURE 4.19: ACF and PACF plot for $\log_Set_{V1}$ observed in CX2.33 FIGURE 4.20: ACF and PACF plot for $\log_Set_{V2}$ observed in CX2.33

FIGURE 4.21: ACF and PACF plot for $\log_Set_{I1}$ observed in CX2.33 FIGURE 4.22: ACF and PACF plot for $\log_Set_{I2}$ observed in CX2.33

FIGURE 4.23: ACF and PACF plot for $\log_{Tot_en_Discha}$ observed in CX2.33 FIGURE 4.24: ACF and PACF plot for $\log_{Tot_en_Cha}$ observed in CX2.33

By observing the plots for $\log_Set_{V1}$, $\log_Set_{V2}$, $\log_Set_{I1}$, $\log_Set_{I2}$, $p = 1$ and $q = 1$. Similarly, for $\log_{Tot_en_Cha}$ and $\log_{Tot_en_Discha}$, $p = 40$ and $q = 1$,

with d chosen 1 for all the features.

Therefore for both the modified features $\log_Set_{V_1}$, $\log_Set_{V_2}$, $\log_Set_{I_1}$, $\log_Set_{I_2}$, the chosen ARIMA model order (p, d, q) is $(1, 1, 1)$ and for $\log_Tot_en_Cha$, $\log_Tot_en_Discha$ the chosen ARIMA model order (p, d, q) is $(40, 1, 1)$.

4.2.3 Capacity Estimation

Results for this dataset is presented in this section. Plots of the actual capacity and the estimated capacity values in Ah versus the cycle number are presented. Error metrics are presented in terms of mean square error(MSE) and mean absolute percentage error (MAPE) in Table 4.5. As mentioned in the start of this section, training and testing is done in a cyclic order, i.e. five dataset are used for training and one for testing. This is done for each dataset e.g. if the results for CX2_33 are presented, that means CX2_34, CX2_35, CX2_36, CX2_37, CX2_38 dataset were used for training. Results are presented in Figure 4.25, 4.26, 4.27, 4.28, 4.29, 4.30 and the detailed results with error metrics are presented in Table 4.5.

FIGURE 4.25: Capacity Estimation for CX2_38 plotted against Cycle Number

FIGURE 4.26: Capacity Estimation for CX2.37 plotted against Cycle Number

FIGURE 4.27: Capacity Estimation for CX2.36 plotted against Cycle Number

FIGURE 4.28: Capacity Estimation for CX2_35 plotted against Cycle Number

FIGURE 4.29: Capacity Estimation for CX2_34 plotted against Cycle Number

FIGURE 4.30: Capacity Estimation for CX2.33 plotted against Cycle Number

TABLE 4.5: Capacity Estimation Results for CALCE's CX2 Battery Dataset

Battery	MSE	MAPE(%)
CX2_33	0.0008699713	2.4121788134
CX2_34	0.0003542805	1.5150116474
CX2_35	0.0012265566	3.0373859678
CX2_36	0.0011042615	2.2659031795
CX2_37	0.0000741513	0.6711962829
CX2_38	0.0040586337	3.5101270719

As it can be observed from Figure 4.25, there is a saturation or flat line in capacity estimation towards the end of the battery life. This is due to limitation in the training set. Only in the case of battery CX2.38, it was operated till the capacity reached 0.3 Ah, rest of batteries were not operated to that extent. Therefore while training on the rest of the dataset, the behavior of the battery parameters or features were not learned if operated to that low extent. Due to this the performance of the model is not correct. Please note that this incorrect capacity estimation is used in the calculation of error metrics.

4.2.4 Capacity Prognosis

The prognosis results for this battery dataset are presented in Table 4.6. The prognosis is done at various cycle during runtime. The result along with error metrics, are presented here. It can be observed that, as the battery ages, more historical data is available, the error is also getting minimised.

TABLE 4.6: Capacity Prognosis Results for CALCE's CX2 Battery Dataset

Battery	Prognosis @	MAPE(%)
CX2_33	400 th Cycle	13.35575
	800 th Cycle	7.68191
	1200 th Cycle	4.89173
CX2_34	400 th Cycle	12.67139
	800 th Cycle	7.13572
	1200 th Cycle	4.57912
CX2_35	400 th Cycle	15.71022
	800 th Cycle	9.41504
	1200 th Cycle	5.91531
CX2_36	400 th Cycle	12.15013
	800 th Cycle	6.75013
	1200 th Cycle	4.24842
CX2_37	400 th Cycle	9.34147
	800 th Cycle	3.95911
	1200 th Cycle	2.26431
CX2_38	400 th Cycle	12.63132
	800 th Cycle	8.61591
	1200 th Cycle	5.91643

4.3 NASA's Randomized Battery Usage Dataset

This dataset comprises of a set of 12 Li-ion batteries (RW3, RW4, RW5, RW6, RW13, RW14, RW15, RW16, RW1, RW2, RW7, RW8) [75]. Four Li-ion batteries (RW3, RW4, RW5 and RW6) were continuously operated by repeatedly charging them to 4.2V and then discharging them to 3.2V using a randomised sequence

of discharging currents between 0.5A and 4A. This discharging profile is referred to here as a random walk (RW) discharging. Batteries are first charged at 2A (constant current) until they reach 4.2V, at which time the charging switches to a constant voltage mode and continues charging the batteries until the charging current falls below 0.01A. After every fifty RW cycles, a series of reference charging and discharging cycles were performed in order to provide reference benchmarks for the battery state of health.

Four Li-ion batteries (RW13, RW14, RW15 and RW16) were continuously operated by repeatedly charging them to 4.2V and then discharging them to 3.2V using a randomised sequence of discharging currents between 0.5A and 5A. This discharging profile is referred to here as a random walk (RW) discharging. A customised probability distribution is used in this experiment to select a new load setpoint every 1 minute during RW discharging operation. The custom probability distribution was designed to be skewed towards selecting lower currents.

The probabilities of selecting each potential load setpoint are shown in Table 4.7:

TABLE 4.7: Load Setpoint with Probability

Load Setpoint	Probability
0.5A	7.2%
1.0A	14.8%
1.5A	19.3%
2.0A	21.6%
2.5A	14.6%
3.0A	10.0%
3.5A	6.5%
4.0A	4.0%
4.5A	1.5%
5.0A	0.5%

Batteries are charged at 2A (constant current) until they reach 4.2V, at which time the charging switches to a constant voltage mode and continues charging the batteries until the charging current falls below 0.01A. After every fifty RW cycles, a series of reference charging and discharging cycles were performed in order to provide reference benchmarks for battery state health.

Four Li-ion batteries (RW1, RW2, RW7 and RW8) were continuously operated by repeatedly discharging them to 3.2V using a randomised sequence of discharging currents between 0.5A and 4A. This discharging profile is referred to here as a random walk (RW) discharging. After each discharging cycle, the batteries were charged for a randomly selected duration between 0.5 hours and 3 hours. Batteries are first charged at 2A (constant current) until they reach 4.2V, at which time the charging switches to a constant voltage mode and continues charging the batteries until the charging current falls below 0.01A. After every fifty RW cycles, a series of reference charging and discharging cycles were performed in order to provide reference benchmarks for the battery state of health. Details of the experimental dataset are given in Table 4.8.

TABLE 4.8: Details of Experimental NASA's Randomized Battery Usage Dataset

Cell Num.	Cutoff Voltage(V)	Charging Current	Discharging Current	Temperature (° C)
RW3	3.2	2A	0.5A-4A	24
RW4	3.2	2A	0.5A-4A	24
RW5	3.2	2A	0.5A-4A	24
RW6	3.2	2A	0.5A-4A	24
RW13	3.2	2A	0.5A-5A	24
RW14	3.2	2A	0.5A-5A	24
RW15	3.2	2A	0.5A-5A	24
RW16	3.2	2A	0.5A-5A	24
RW1	3.2	2A	0.5A-4A	24
RW2	3.2	2A	0.5A-4A	24
RW7	3.2	2A	0.5A-4A	24
RW8	3.2	2A	0.5A-4A	24

For training and testing, Six batteries namely RW3, RW4, RW5, RW6, RW13, RW14 are used for training, and the other six namely RW15, RW16, RW1, RW2, RW7, RW8 are used for testing. Note that the batteries RW1, RW2, RW7, RW8 were charged for a randomly selected duration between 0.5 hours and 3 hours, thus are considered as the partial charging category. The hyper-parameter for stochastic gradient boosting regression models are kept the same as mentioned in Section 4.1.

4.3.1 Feature Selection

The feature importance for NASA's randomized battery usage dataset is presented in Figure 4.31 and 4.32 which belong to CC and CV mode respectively.

FIGURE 4.31: Feature Importance observed in CC mode for NASA's Randomized Battery Usage Dataset

FIGURE 4.32: Feature Importance observed in CV mode for NASA's Randomized Battery Usage Dataset

Features are selected, which are relatively important than other features and also belonging to two different ranges of SoC. As clearly observed from Figure 4.31 and 4.32, Set_{V_1} comprises the voltage range from 4.01V to 4.08V, Set_{V_2} comprises of the voltage range from 4.13V to 4.20V, Set_{I_1} comprises the current range from 1.88A to 1.95A and finally Set_{I_2} comprises of the current range from 0.32A to 0.39A. The voltage and current measurements that doesn't fall in these ranges are not used offline as well as online phase for this dataset.

4.3.2 ARIMA Order Identification

Plots for the modified features $\log_Set_{V_1}$, $\log_Set_{V_2}$, $\log_Set_{I_1}$, $\log_Set_{I_2}$, $\log_Tot_en_Cha$, $\log_Tot_en_Discha$ for RW3 are presented in Figure 4.33, 4.34, 4.35, 4.36, 4.37 and 4.38 respectively.

FIGURE 4.33: ACF and PACF plot for $\log_Set_{V_1}$ observed in RW3

FIGURE 4.34: ACF and PACF plot for $\log_Set_{V2}$ observed in RW3 FIGURE 4.35: ACF and PACF plot for $\log_Set_{I1}$ observed in RW3

FIGURE 4.36: ACF and PACF plot for $\log_Set_{I1}$ observed in RW3 FIGURE 4.37: ACF and PACF plot for $\log_Tot.en.Discha$ observed in RW3

FIGURE 4.38: ACF and PACF plot for $\log_Tot_en_Cha$ observed in RW3

By observing the plots for $\log_Set_{V1}$, $\log_Set_{V2}$, $\log_Set_{I1}$, $\log_Set_{I2}$, $p = 1$ and $q = 1$. Similarly, for $\log_Tot_en_Cha$, $p = 40$ and $q = 1$ and for $\log_Tot_en_Discha$ $p = 20$ and $q = 1$, with d chosen 1 for all the features.

Therefore for both the modified features $\log_Set_{V1}$, $\log_Set_{V2}$, $\log_Set_{I1}$, $\log_Set_{I2}$, the chosen ARIMA model order (p, d, q) is $(1, 1, 1)$ and for $\log_Tot_en_Cha$ the chosen ARIMA model order (p, d, q) is $(40, 1, 1)$ and for $\log_Tot_en_Discha$ the chosen ARIMA model order (p, d, q) is $(20, 1, 1)$.

4.3.3 Capacity Estimation

This dataset is different from the previous two datasets. In this dataset, the actual capacity is only measured accurately in the reference cycles, as mentioned already. Therefore charging cycle, which is not a reference cycle, does not have a target or response variable for training. To solve this problem, simple interpolation capacity is used in between the reference cycle. Please note that this interpolation is used in the training phase and not in the testing Phase. Plots of the actual capacity and the estimated capacity values in Ah versus the cycle Number are presented.

Actual capacity values are plotted only for reference cycle.

Error metric is presented in terms of Mean Square Error(MSE) and Mean Absolute Percentage Error (MAPE) in Table 4.9. Note that the Error Metrics are being calculated only for Reference Cycle because this is the only instance where there a ground truth present. For rest of the estimates apart from the reference cycle are not taken into account for calculating the error.

As mentioned at the start of this section, six batteries namely RW3, RW4, RW5, RW6, RW13, RW14 are used for training and the other six namely RW15, RW16, RW1, RW2, RW7, RW8 are used for testing. Results are presented in Figure 4.39, 4.40, 4.41, 4.42, 4.43, 4.44 and the detailed results with error metrics are presented in Table 4.9.

FIGURE 4.39: Capacity Estimation for RW15 plotted against Cycle Number

FIGURE 4.40: Capacity Estimation for RW16 plotted against Cycle Number

FIGURE 4.41: Capacity Estimation for RW1 plotted against Cycle Number

FIGURE 4.42: Capacity Estimation for RW2 plotted against Cycle Number

FIGURE 4.43: Capacity Estimation for RW7 plotted against Cycle Number

FIGURE 4.44: Capacity Estimation for RW8 plotted against Cycle Number

TABLE 4.9: Capacity Estimation Results for NASA's Randomized Battery Usage Dataset

Battery	MSE	MAPE(%)
RW15	0.0030758179	4.4418840636
RW16	0.0010634005	2.5611858579
RW1	0.0027139716	4.3829433638
RW2	0.0052972080	5.6202275315
RW7	0.0023145062	3.9202339083
RW8	0.0037709649	5.2657861710

4.3.4 Capacity Prognosis

The prognosis results for this battery dataset are presented in Table 4.10. The prognosis is done at various cycles during runtime. The results along with error metrics, are presented here. It can be observed that, as the battery ages, more historical data is available and thus the error is also getting minimised.

TABLE 4.10: Capacity Prognosis Results for NASA's Randomized Battery Usage Dataset

Battery	Prognosis @	MAPE(%)
RW15	300 th Cycle	16.17133
	600 th Cycle	10.17289
	900 th Cycle	6.01463
RW16	300 th Cycle	17.58248
	600 th Cycle	11.14735
	900 th Cycle	7.71844
RW1	300 th Cycle	17.17359
	600 th Cycle	12.62791
	900 th Cycle	8.63743
RW2	300 th Cycle	19.56191
	600 th Cycle	13.65027
	900 th Cycle	8.36191
	1200 th Cycle	5.56381
RW7	300 th Cycle	17.26413
	400 th Cycle	10.61513
	900 th Cycle	7.93132
RW8	300 th Cycle	15.41813
	400 th Cycle	9.78263
	900 th Cycle	5.95816

Chapter 5

Conclusion

This thesis started with an insight into the global trend of moving away from the traditional energy sources and towards the renewable energy source. But the intrinsic intermittent nature of renewables poses a risk in the required continuous supply of energy and stability. The issue with the irregular supply or dependence of renewables on the environment can be solved by using energy storage systems. In this regard, the use of a lithium-ion battery as an energy storage system is most desirable because of its high energy density, small size, and relatively low maintenance. However, to use the battery safely and efficiently, there is a need to estimate the health condition of the battery. In most of the literature, index based on the battery's capacity is used to monitor the health of the same. The capacity is defined as the available energy stored in a fully charged the lithium-ion battery can deliver. Thus, the main objective of this work was to estimate the current capacity of the battery and forecast the same by observing the trend or usage of the battery. In Chapter 1, a general introduction, along with motivation, was presented. In addition to this, historical background and market trends about lithium-ion battery were mentioned so that we can better relate to the current scenario.

The literature was reviewed in Chapter 2. This provided a basic understanding of the methods that were used in the past to estimate the battery's capacity or state-of-health. Estimating the current capacity and forecasting the capacity trend for the future charge cycles of the battery whose internal health parameters

are hard to gauge is difficult to attain. Three different ways to determine capacity were studied, and the merits of each of them were thoroughly studied. It was observed that existing prognostics methods had produced promising results in the estimation of the battery's state-of-health or capacity. However, there were limitations and gaps that were noticed. Most important of them was that in the case of partially charging/discharging of batteries, the listed literature did not explain how to deal with missing values. Apart from these, model-based approaches relied heavily on accurate physics-based models. Data-driven methods were mostly using variants of neural networks, which are sensitive to the architecture of the model and have to be changed according to the dataset. Understanding these gaps in the literature, a novel way to combine a data-driven method (other than the neural network) ensembled with statistical approach was developed.

In Chapter 3, the mathematical formulation used in this work was explained, followed by the feature extraction and selection. A stochastic gradient boosting regression model was used as a base structure which could be trained using fewer data, the hyperparameter tuning is fast and need not be changed according to the dataset. The partial charge/discharge and missing value issue were taken care by carefully selecting features and using an auto-regressive integrated moving average process. Keeping in mind the real-life scenarios in which batteries are seldom fully charged/discharged, during which the full range of voltage or current measurements may not be available. This was solved by using only those part of the feature, which are essential in the building of the regression model and using ARIMA one-step forecasting for missing value imputations. For prognosis, the results were presented at different instances charge cycle along with the confidence interval, which can be used for failure prevention through timely and effective maintenance actions. Overall the whole system was designed to estimate and forecast the capacity as accurately as possible considering all the real-life scenarios.

One of the significant contributions of this work is to present an efficient and fast way to select features which will be used in building data-driven models. Lastly, the overall system design used in both the offline and the online stages were presented in Chapter 3. In this work, the data-driven capacity degradation model was developed using voltage, current and time measurements.

Effectiveness and accuracy for this framework were presented by using three independent dataset, and the results were presented in Chapter 4. Stochastic gradient boosting regression (SGBR) ensembled with autoregressive integrated moving average (ARIMA) provided the basic structure for capacity estimation and prognosis. The hyperparameters for SGBR were kept the same for all three different dataset, thus achieving the adaptability irrespective of the battery specification. Results for capacity estimation were well within the mean absolute error percentage of 6%, which is a respectable accuracy compared to that of the work in the literature. For prognosis of the capacity trends, it is understandable that at the initial use of the battery it is very difficult to forecast capacity, but as more data was available, trends were getting more strong, and thus errors were also dramatically decreasing.

5.1 Future Works

There quite a few areas and scenarios where more work is needed to better understand the health parameter of the batteries.

In this implementation, the batteries were kept at room temperature. So, for further improvements, getting data in which batteries are tested in variable temperature and other operating conditions. With the varied dataset, this framework can be used to cover all possible operation scenarios. Due to the slow nature of battery ageing, this data acquisition is very time consuming, and not many open source databases are available.

It was also observed in this work that when data-driven model encounters new condition or feature on which it was not trained before, the results were not good. This can be a very promising future work direction in which the results can be extrapolated for the conditions which were not observed in the training phase.

Lastly, if an accurate mathematical modelling of the battery can be accomplished or hardware-in-the-loop simulation can be implemented, it can generate a large battery dataset to train and test all scenarios. With the availability of huge dataset, a method like reinforcement learning can be implemented to select the usage of the battery so that battery life can be maximised and better return

on the investment can be obtained. But the catch here is that if we have an accurate mathematical model, then why we should bother to develop data-driven degradation model. Nevertheless, this can be a very interesting area of work and research to work with the very basic mathematical model to build dataset and use data-driven method to estimate as well as optimise energy storage systems.

In conclusion, this framework developed can be well utilised for different lithium-ion batteries for capacity estimation, along with state-of-health and remaining useful life prediction.

Appendix A

Fuzzy Logic Rules

Fuzzy rules are used within fuzzy logic systems to infer an output based on input variables. These IF-Then rules are used for selecting the respective *SGBR* models based on the availability of the features. Please note that the models are selected on the availability of features Set_{V1} , Set_{V2} , Set_{I1} , Set_{I2} . The rules are presented here in the order of the priority. In the online phases for each charging/discharging, rules will be checked from top to bottom. If a rule matches the availability of the feature, then that respective model will be selected and the rest of the later condition would not be checked.

The Fuzzy Logic Rules are as follows:

IF Set_{V1} , Set_{V2} , Set_{I1} , Set_{I2} is Observed
THEN Select $SGBR_1$ and Exit the Fuzzy Rules Check

IF Set_{V1} , Set_{V2} , Set_{I1} is Observed
THEN Select $SGBR_2$ and Exit the Fuzzy Rules Check

IF Set_{V1} , Set_{V2} , Set_{I2} is Observed
THEN Select $SGBR_3$ and Exit the Fuzzy Rules Check

IF Set_{V1} , Set_{I1} , Set_{I2} is Observed
THEN Select $SGBR_7$ and Exit the Fuzzy Rules Check

-
- IF** $Set_{V2}, Set_{I1}, Set_{I2}$ is Observed
THEN Select $SGBR_8$ and Exit the Fuzzy Rules Check
- IF** Set_{V1}, Set_{V2} is Observed
THEN Select $SGBR_4$ and Exit the Fuzzy Rules Check
- IF** Set_{V1}, Set_{I1} is Observed
THEN Select $SGBR_{12}$ and Exit the Fuzzy Rules Check
- IF** Set_{V1}, Set_{I2} is Observed
THEN Select $SGBR_{13}$ and Exit the Fuzzy Rules Check
- IF** Set_{V2}, Set_{I1} is Observed
THEN Select $SGBR_{14}$ and Exit the Fuzzy Rules Check
- IF** Set_{V2}, Set_{I2} is Observed
THEN Select $SGBR_{15}$ and Exit the Fuzzy Rules Check
- IF** Set_{I1}, Set_{I2} is Observed
THEN Select $SGBR_9$ and Exit the Fuzzy Rules Check
- IF** Set_{V1} is Observed
THEN Select $SGBR_5$ and Exit the Fuzzy Rules Check
- IF** Set_{V2} is Observed
THEN Select $SGBR_6$ and Exit the Fuzzy Rules Check
- IF** Set_{I1} is Observed
THEN Select $SGBR_{10}$ and Exit the Fuzzy Rules Check
- IF** Set_{I2} is Observed
THEN Select $SGBR_{11}$ and Exit the Fuzzy Rules Check
- IF** No Window Set is Observed
THEN Select $SGBR_{16}$ and Exit the Fuzzy Rules Check

Bibliography

- [1] Paul Gipe. Wind power comes of age, 1995. [2](#)
- [2] Dave Elliott. Renewable energy and sustainable futures. *Futures*, 32(3-4): 261–274, 2000. [2](#)
- [3] Godfrey Boyle. Renewable energy. *Renewable Energy*, page 456, 2004. [2](#)
- [4] Anthony Price. Electrical energy storagea review of technology options. In *Proceedings of the Institution of Civil Engineers-Civil Engineering*, volume 6, pages 52–58. Thomas Telford Ltd, 2005. [3](#)
- [5] Bruce Dunn, Haresh Kamath, and Jean-Marie Tarascon. Electrical energy storage for the grid: a battery of choices. *Science*, 334(6058):928–935, 2011. [3](#), [4](#), [5](#), [6](#)
- [6] Boucar Diouf and Ramchandra Pode. Potential of lithium-ion batteries in renewable energy. *Renewable Energy*, 76:375–380, 2015. [3](#), [4](#), [5](#), [6](#)
- [7] Ana-Irina Stan, Maciej Świerczyński, Daniel-Ioan Stroe, Remus Teodorescu, and Søren Juhl Andreasen. Lithium ion battery chemistries from renewable energy storage to automotive and back-up power applicationsan overview. In *2014 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM)*, pages 713–720. IEEE, 2014. [3](#)
- [8] Matthew T Lawder, Bharatkumar Suthar, Paul WC Northrop, Sumitava De, C Michael Hoff, Olivia Leitermann, Mariesa L Crow, Shriram Santhanagopalan, and Venkat R Subramanian. Battery energy storage system (bess) and battery management system (bms) for grid-scale applications. *Proceedings of the IEEE*, 102(6):1014–1030, 2014. [3](#)

- [9] Benoit Joffres, Dorothée Laurenti, Nadège Charon, Antoine Daudin, Alain Quignard, and Christophe Geantet. Thermochemical conversion of lignin for fuels and chemicals: A review. *Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles*, 68(4):753–763, 2013. [3](#)
- [10] M Berecibar, I Gandiaga, I Villarreal, N Omar, J Van Mierlo, and P Van den Bossche. Critical review of state of health estimation methods of li-ion batteries for real applications. *Renewable and Sustainable Energy Reviews*, 56: 572–587, 2016. [3](#)
- [11] M Rosa Palacín and Anne de Guibert. Why do batteries fail? *Science*, 351 (6273):1253292, 2016. [3](#)
- [12] George E Blomgren. The development and future of lithium ion batteries. *Journal of The Electrochemical Society*, 164(1):A5019–A5025, 2017. [4](#)
- [13] Akira Yoshino. The birth of the lithium-ion battery. *Angewandte Chemie International Edition*, 51(24):5798–5800, 2012. [4](#)
- [14] Ralph J Brodd, Akiya Kozawa, and Masaki Yoshio. *Lithium-Ion Batteries: Science and Technologies*. Springer, 2009. [4](#)
- [15] Tsuyoshi Sasaki, Yoshio Ukyo, and Petr Novák. Memory effect in a lithium-ion battery. *Nature materials*, 12(6):569, 2013. [4](#)
- [16] Qingsong Wang, Ping Ping, Xuejuan Zhao, Guanquan Chu, Jinhua Sun, and Chunhua Chen. Thermal runaway caused fire and explosion of lithium ion battery. *Journal of Power Sources*, 208:210–224, 2012. [5](#)
- [17] Ralph J Brodd. Synopsis of the lithium-ion battery markets. In *Lithium-Ion Batteries*, pages 1–7. Springer, 2009. [5](#)
- [18] Rebecca E Ciez and JF Whitacre. The cost of lithium is unlikely to upend the price of li-ion storage systems. *Journal of Power Sources*, 320:310–313, 2016. [5](#), [6](#)
- [19] Richard Van Noorden. The rechargeable revolution: A better battery. *Nature News*, 507(7490):26, 2014. [5](#)
- [20] Björn Nykvist and Måns Nilsson. Rapidly falling costs of battery packs for electric vehicles. *Nature Climate Change*, 5(4):329, 2015. [6](#)

- [21] Gregory L Plett. Sigma-point kalman filtering for battery management systems of lipb-based hev battery packs: Part 1: Introduction and state estimation. *Journal of Power Sources*, 161(2):1356–1368, 2006. [8](#)
- [22] Wei He, Nicholas Williard, Michael Osterman, and Michael Pecht. Prognostics of lithium-ion batteries based on dempster-shafer theory and the bayesian monte carlo method. *Journal of Power Sources*, 196(23):10314–10321, 2011. [8](#), [18](#), [60](#)
- [23] Dong Wang, Qiang Miao, and Michael Pecht. Prognostics of lithium-ion batteries based on relevance vectors and a conditional three-parameter capacity degradation model. *Journal of Power Sources*, 239:253–264, 2013. [14](#)
- [24] Selina SY Ng, Yinjiao Xing, and Kwok L Tsui. A naive bayes model for robust remaining useful life prediction of lithium-ion battery. *Applied Energy*, 118: 114–123, 2014. [8](#)
- [25] Guang Jin, David E Matthews, and Zhongbao Zhou. A bayesian framework for on-line degradation assessment and residual life prediction of secondary batteries inspacecraft. *Reliability Engineering & System Safety*, 113:7–20, 2013. [12](#)
- [26] De Z Li, Wilson Wang, and Fathy Ismail. A mutated particle filter technique for system state estimation and battery life prediction. *IEEE Transactions on Instrumentation and Measurement*, 63(8):2034–2043, 2014. [12](#)
- [27] Yang Hu, Piero Baraldi, Francesco Di Maio, and Enrico Zio. A particle filtering and kernel smoothing-based approach for new design component prognostics. *Reliability Engineering & System Safety*, 134:19–31, 2015. [12](#)
- [28] Qiang Miao, Lei Xie, Hengjuan Cui, Wei Liang, and Michael Pecht. Remaining useful life prediction of lithium-ion battery with unscented particle filter technique. *Microelectronics Reliability*, 53(6):805–810, 2013. [12](#)
- [29] Shriram Santhanagopalan and Ralph E White. State of charge estimation using an unscented filter for high power lithium ion cells. *International Journal of Energy Research*, 34(2):152–163, 2010. [12](#)
- [30] Marc Doyle, Thomas F Fuller, and John Newman. Modeling of galvanostatic charge and discharge of the lithium/polymer/insertion cell. *Journal of the Electrochemical Society*, 140(6):1526–1533, 1993. [12](#)

- [31] M Safari, M Morcrette, A Teyssot, and C Delacourt. Multimodal physics-based aging model for life prediction of li-ion batteries. *Journal of The Electrochemical Society*, 156(3):A145–A153, 2009. [12](#)
- [32] Gang Ning, Bala Haran, and Branko N Popov. Capacity fade study of lithium-ion batteries cycled at high discharge rates. *Journal of Power Sources*, 117(1-2):160–169, 2003. [12](#)
- [33] Sung Park, Andreas Savvides, and Mani B Srivastava. Battery capacity measurement and analysis using lithium coin cell battery. In *Proceedings of the 2001 International Symposium on Low Power Electronics and Design*, pages 382–387. IEEE, 2001. [12](#)
- [34] Mehdi Gholizadeh and Farzad R Salmasi. Estimation of state of charge, unknown nonlinearities, and state of health of a lithium-ion battery based on a comprehensive unobservable model. *IEEE Transactions on Industrial Electronics*, 61(3):1335–1344, 2013. [12](#)
- [35] Fan Li and Jiuping Xu. A new prognostics method for state of health estimation of lithium-ion batteries based on a mixture of gaussian process models and particle filter. *Microelectronics Reliability*, 55(7):1035–1045, 2015. [13](#)
- [36] Jürgen Remmlinger, Michael Buchholz, Markus Meiler, Peter Bernreuter, and Klaus Dietmayer. State-of-health monitoring of lithium-ion batteries in electric vehicles by on-board internal resistance estimation. *Journal of Power Sources*, 196(12):5357–5363, 2011. [13](#)
- [37] Juergen Remmlinger, Michael Buchholz, Thomas Soczka-Guth, and Klaus Dietmayer. On-board state-of-health monitoring of lithium-ion batteries using linear parameter-varying models. *Journal of Power Sources*, 239:689–695, 2013. [13](#)
- [38] Bing Long, Weiming Xian, Lin Jiang, and Zhen Liu. An improved autoregressive model by particle swarm optimization for prognostics of lithium-ion batteries. *Microelectronics Reliability*, 53(6):821–831, 2013. [14](#)
- [39] Adnan Nuhic, Tarik Terzimehic, Thomas Soczka-Guth, Michael Buchholz, and Klaus Dietmayer. Health diagnosis and remaining useful life prognostics of lithium-ion batteries using data-driven methods. *Journal of Power Sources*, 239:680–688, 2013. [14](#), [15](#)

- [40] Chao Hu, Byeng D Youn, Pingfeng Wang, and Joung Taek Yoon. Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. *Reliability Engineering & System Safety*, 103:120–135, 2012. [14](#)
- [41] Meru A Patil, Piyush Tagade, Krishnan S Hariharan, Subramanya M Kolake, Taewon Song, Taejung Yeo, and Seokgwang Doo. A novel multistage support vector machine based approach for li ion battery remaining useful life estimation. *Applied Energy*, 159:285–297, 2015. [14](#), [15](#)
- [42] Datong Liu, Jianbao Zhou, Haitao Liao, Yu Peng, and Xiyuan Peng. A health indicator extraction and optimization framework for lithium-ion battery degradation modeling and prognostics. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(6):915–928, 2015. [15](#)
- [43] Jie Liu, Abhinav Saxena, Kai Goebel, Bhaskar Saha, and Wilson Wang. An adaptive recurrent neural network for remaining useful life prediction of lithium-ion batteries. Technical report, National Aeronautics and Space Administration Moffett Field CA AMES Research, 2010. [15](#)
- [44] Linxia Liao and Felix Köttig. Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. *IEEE Transactions on Reliability*, 63(1):191–207, 2014. [15](#), [17](#)
- [45] Datong Liu, Yue Luo, Jie Liu, Yu Peng, Limeng Guo, and Michael Pecht. Lithium-ion battery remaining useful life estimation based on fusion nonlinear degradation ar model and rpf algorithm. *Neural Computing and Applications*, 25(3-4):557–572, 2014. [16](#)
- [46] 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, 2015. [16](#)
- [47] J Liu, W Wang, F Ma, YB Yang, and CS Yang. A data-model-fusion prognostic framework for dynamic system state forecasting. *Engineering Applications of Artificial Intelligence*, 25(4):814–823, 2012. [16](#)
- [48] Hong Li, Donghui Pan, and CL Philip Chen. Intelligent prognostics for battery health monitoring using the mean entropy and relevance vector machine.

- IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(7):851–862, 2014. [17](#)
- [49] Bhaskar Saha, Kai Goebel, Scott Poll, and Jon Christophersen. Prognostics methods for battery health monitoring using a bayesian framework. *IEEE Transactions on Instrumentation and Measurement*, 58(2):291–296, 2008. [17](#)
- [50] Bhaskar Saha, Scott Poll, Kai Goebel, and Jon Christophersen. An integrated approach to battery health monitoring using bayesian regression and state estimation. In *2007 IEEE Autotestcon*, pages 646–653. Ieee, 2007. [17](#)
- [51] Yinjiao Xing, Eden WM Ma, Kwok-Leung Tsui, and Michael Pecht. An ensemble model for predicting the remaining useful performance of lithium-ion batteries. *Microelectronics Reliability*, 53(6):811–820, 2013. [17](#)
- [52] Hancheng Dong, Xiaoning Jin, Yangbing Lou, and Changhong Wang. Lithium-ion battery state of health monitoring and remaining useful life prediction based on support vector regression-particle filter. *Journal of Power Sources*, 271:114–123, 2014. [17](#)
- [53] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. [20](#)
- [54] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 10:993–1001, 1990. [20](#)
- [55] Robert E Schapire. The boosting approach to machine learning: An overview. In *Nonlinear Estimation and Classification*, pages 149–171. Springer, 2003. [20](#)
- [56] Martin Sewell. Ensemble learning. *RN*, 11(02), 2008. [20](#)
- [57] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. [20](#)
- [58] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001. [20](#), [25](#)
- [59] Simon J Pittman and Kerry A Brown. Multi-scale approach for predicting fish species distributions across coral reef seascapes. *PloS One*, 6(5):e20583, 2011. [21](#)

- [60] Rie Johnson and Tong Zhang. Learning nonlinear functions using regularized greedy forest. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):942–954, 2013. [21](#)
- [61] Wenxin Jiang et al. On weak base hypotheses and their implications for boosting regression and classification. *The Annals of Statistics*, 30(1):51–73, 2002. [24](#)
- [62] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005. [24](#)
- [63] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002. [26](#)
- [64] Ping Li, Qiang Wu, and Christopher J Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in Neural Information Processing Systems*, pages 897–904, 2008. [27](#)
- [65] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. [27](#), [47](#)
- [66] Spyros Makridakis and Michele Hibon. Arma models and the box–jenkins methodology. *Journal of Forecasting*, 16(3):147–163, 1997. [28](#), [30](#)
- [67] Samarjit Das. *Time series analysis*. Princeton University Press, Princeton, NJ, 1994. [28](#)
- [68] Javier Contreras, Rosario Espinola, Francisco J Nogales, and Antonio J Conejo. Arima models to predict next-day electricity prices. *IEEE Transactions on Power Systems*, 18(3):1014–1020, 2003. [30](#)
- [69] Peiyuan Chen, Troels Pedersen, Birgitte Bak-Jensen, and Zhe Chen. Arima-based time series model of stochastic wind power generation. *IEEE Transactions on Power Systems*, 25(2):667–676, 2010. [30](#)
- [70] Kong Soon Ng, Chin-Sien Moo, Yi-Ping Chen, and Yao-Ching Hsieh. Enhanced coulomb counting method for estimating state-of-charge and state-of-health of lithium-ion batteries. *Applied Energy*, 86(9):1506–1511, 2009. [32](#)

- [71] Herbert Enderton and Herbert B Enderton. *A mathematical introduction to logic*. Elsevier, 2001. [37](#)
- [72] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*, volume 620. Centrum voor Wiskunde en Informatica Amsterdam, 1995. [47](#)
- [73] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. [47](#)
- [74] B Saha and K Goebel. Battery data set. *NASA AMES Prognostics Data Repository*, 2007. [47](#), [56](#)
- [75] B Bole, C Kulkarni, and M Daigle. Randomized battery usage data set. *NASA AMES prognostics data repository*, 2014. [70](#)