Physics-Informed Feature Engineering and Deep Embedded Clustering Framework for Lithium-Ion Battery State-of-Health Analysis

A Thesis submitted

in partial fulfillment of the requirements for the Degree of

MASTER OF SCIENCE

in

MATHEMATICS

by

Marvi Tyagi

Roll No.: 24/MSCMAT/39

Under the Supervision of

Dr. Anshul Arora

Assistant Professor

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi – 110042, India

May 2026

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi – 110042, India

CANDIDATE'S DECLARATION

I, **Marvi Tyagi (24/MSCMAT/39)**, hereby declare that the work presented in this thesis entitled **"Physics-Informed Feature Engineering and Deep Embedded Clustering Framework for Lithium-Ion Battery State-of-Health Analysis"** is an authentic record of my own research and study carried out during the period from **August 2025 to May 2026** under the supervision of **Dr. Anshul Arora**, Department of Applied Mathematics, Delhi Technological University.

The material presented in this thesis has not been submitted by us for the award of any other degree of this or any other institute.

Candidate's Signature: _____

This is to certify that the student has incorporated all corrections suggested during evaluation, and the statement made by the candidate is to the best of our knowledge.

Signature of Supervisor

_____

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi – 110042, India

CERTIFICATE

This is to certify that **Marvi Tyagi (24/MSCMAT/39)** has successfully completed the thesis entitled **"Physics-Informed Feature Engineering and Deep Embedded Clustering Framework for Lithium-Ion Battery State-of-Health Analysis"** as part of the academic requirements in the Department of Applied Mathematics, Delhi Technological University, Delhi.

The work presented in this thesis has been carried out under my supervision. To the best of my knowledge, this thesis is a bona fide record of the students' effort and has not been submitted to any other institution for any academic purpose.

Signature (Dr. Anshul Arora)

Department of Applied Mathematics

Acknowledgements

The completion of this dissertation would not have been possible without the guidance, support, and encouragement of several remarkable individuals to whom I owe my deepest gratitude.

First and foremost, I express my most sincere gratitude to my supervisor, **Dr. Anshul Arora**, Assistant Professor, Department of Applied Mathematics, Delhi Technological University. His unwavering academic guidance, intellectual generosity, and consistent encouragement throughout every stage of this research have been the foundation upon which this work was built. His patience during challenges, rigorous approach to research, and genuine investment in the growth of his students have shaped not only the quality of this dissertation but also my understanding of what it means to pursue knowledge with purpose and precision. I am deeply fortunate to have worked under his supervision.

I sincerely thank the **Department of Applied Mathematics, Delhi Technological University**, for providing the academic environment, computational infrastructure, and institutional support that made this research possible.

I am profoundly grateful to my **Mother, Mrs. Reeta Tyagi**, and my **Father, Dr. Kapil Kumar**, whose love, sacrifice, and steadfast belief in my abilities, academic inspiration, have been my greatest source of strength. Their encouragement through every difficult moment and their quiet confidence that I would succeed mean more to me than words can express.

I warmly thank my **close friends and peers** whose conversations, motivation, and support during both rewarding and challenging phases of this research provided perspective and reminded me that the pursuit of knowledge is always richer when shared with others.

Marvi Tyagi
Department of Applied Mathematics
Delhi Technological University
May 2026

Abstract

Lithium-ion batteries serve as the cornerstone of modern portable electronics, electric vehicles, and large-scale energy storage systems. However, their long-term reliability is constrained by progressive degradation mechanisms that affect capacity, internal resistance, and overall performance. Accurate prediction of battery State-of-Health (SoH) has therefore become essential for enhancing battery safety, optimizing maintenance cycles, and improving energy-management strategies. While significant progress has been made in machine learning–based SoH estimation, current methods typically rely on pre-selected or domain-specific features, leaving a substantial gap in the exploration of generalizable feature engineering frameworks. Motivated by this limitation, the present work aims to systematically investigate new mathematical and data-driven feature constructs derived from real-world battery cycling datasets.

This report proposes a structured methodology for extracting, evaluating, and selecting high-value features from raw operational measurements such as voltage, current, temperature, charge/discharge capacity, and cycle count. The focus is on building a standardized feature-engineering pipeline that incorporates mathematical transformations, degradation-sensitive indicators, and time-series based statistical descriptors. Additionally, the study explores the correlation between temperature dynamics, voltage relaxation behaviour, differential capacity curves, and cycle-induced degradation trends. The derived features are assessed for their predictive relevance using machine learning models, with emphasis on model interpretability and robustness.

The proposed experimental design aims to bridge the gap between electrochemical understanding and data-driven modelling by formulating features rooted in mathematical relationships such as rate of change, gradient estimators, integral measures, and non-linear transformations. The ultimate goal is to identify a compact yet informative subset of features that can generalize across different battery types and cycling conditions.

This work lays the foundation for future development of a comprehensive feature-engineering framework for SoH prediction. The outcomes are expected to support enhanced diagnostic models, reduce dependence on handcrafted domain features, and contribute toward safer, more efficient battery-management systems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Rechargeable lithium-ion batteries (LIBs) are widely used in portable electronics [1], electric vehicles (EV) [2], and energy storage systems [3]. As the demand for clean and renewable energy grows, the diverse applications of LIBs in electrical energy storage will make a significant contribution to reducing carbon emissions and ultimately mitigating global warming [4]. Modern commercial LIBs have been highly optimized, from chemical composition to manufacturing technology, enabling a service lifespan ranging from months to decades. Energy is efficiently stored and utilized through reversible electrochemical reactions within the battery, while a certain level of irreversible degradation reactions also occurs with the cycling process, including active material loss [5] and increased impedance [6] over time. This degradation leads to performance issues like capacity fading, mechanical failure, and thermal instability [7]. To prevent unexpected failures and safety concerns, batteries in EVs should be replaced once their real-time capacity drops to 80 percent of the initial capacity [8]. Machine learning involves the study of algorithms that learn patterns and construct predictive models from data. Particularly, based on the various types of data recorded by a BMS, a machine learning algorithm can construct a predictive model that is capable of estimating the SOH of LiBs. Several studies use current, voltage and charging/discharging cycle data as input for building machine learning models and predicting the SOC or SOH of LiBs. For example, some have used classic machine learning approaches such as Gaussian process regression (GPR) [[9], [10], [11]], support vector regression (SVR) [12,13], Random Forest (RF) [14,15], and single layered feed-forward neural networks (FNN) called extreme learning machine (ELM) [16]. Whereas others have used deep learning approaches that include deep neural networks (DNN) [[17], [18], [19]], convolutional neural network (CNN) [20], recurrent neural networks (RNN) [21,22], transformer networks [5], and various model ensembles [6,23]. By contrast, existing studies used data obtained from electrochemical spectroscopy impedance (EIS), which involves applying small voltage perturbations to the batteries and measuring the difference between the original and resultant currents. EIS provides a significant amount of information on electrochemical processes occurring inside batteries when measured at different frequencies, which can be quickly and inexpensively conducted under operating conditions [4,24]. Furthermore, newly developed instrumentation methods for collecting EIS data in real time enable noninvasive extraction of EIS data from batteries, thus enabling their use in a BMS [25,26]. Some data-driven studies have used EIS data for estimating the battery health and charge status of LiBs using machine learning models, such as GPR [27,28], FNN [29,30], DNN [31], CNN [32], a combination of CNNs, long short-term memory (LSTM) networks [33], and even clustering techniques [26]. In addition, some studies have enriched the current and voltage data using the EV

driving and parking data [34,35]. Accurate estimation of the State of Health (SOH) is crucial for ensuring the performance, safety, and longevity of lithium-ion batteries in electric vehicles. Traditional methods, such as Coulomb Counting and the Extended Kalman Filter, often lack the accuracy and computational efficiency required for modern applications. This study proposes an advanced framework that leverages machine learning models to model the nonlinear degradation patterns of lithium-ion batteries by focusing on key features such as voltage, current, internal resistance, and temperature. The proposed framework incorporates optimized pre-processing techniques, including normalization, to improve data quality and ensure consistency across varying battery conditions.

As the global adoption of residential battery storage systems paired with local photovoltaic (PV) generation increases, prosumers are increasingly motivated to reduce both their electricity costs and dependence on the grid. This shift highlights the importance of accurately evaluating and predicting the battery's State of Health (SOH) and Remaining Useful Life (RUL). These factors are crucial for determining the operational costs and longevity of battery systems. Traditionally, SOH predictions have relied heavily on detailed measurement data and time-intensive simulations. In response, we introduce a new AI-based approach that simplifies SOH estimation.

This innovative AI-driven technique offers substantial benefits for evaluating the economic viability and warranty parameters of battery installations in different regions. It provides a valuable resource for both industry stakeholders and energy system planners aiming to assess and anticipate battery health outcomes efficiently. However, research gaps still remain as presented in detail in the following subsection.

Table 1.1: SOH forecasting and SOH estimation differentiation, inspired by vonBuelow2023

Key differentiator	**Estimation**	**Forecasting**
Modeling	State determination	State change
Cycling data ($SOC, T, I, ...$)		
Length	Short (In past)	Long (In future)
Information on	Electrochemical characteristics	Applied load
Provide details regarding ... of battery aging	Effect	Causes
Critical factors for a SOH forecasting model:		
1	Model includes data on forthcoming load	
2	Aggregation of forthcoming load data	
3	Reflects the increased variability of actual battery usage	
4	Transferability to new battery units	
5	Suitability for second-life battery applications	

1.1 Background

The background section provides the fundamental context required to understand both the operational principles of lithium-ion batteries and the machine learning techniques utilized in this project. It introduces the key structural elements of lithium-ion cells, explains the underlying energy conversion mechanisms, and presents a simplified overview

of the electrochemical processes involved. The section then transitions to a discussion of machine learning methodologies, with particular focus on commonly used neural network architectures relevant to this study.

1.1.1 Background and Overview for SoH prediction

This study is grounded in a substantial body of existing research on lithium-ion battery diagnostics and degradation analysis. Previous work in this domain has established effective baselines through data-driven feature extraction, electrochemical behavior characterization, and early-stage degradation monitoring. These studies collectively demonstrate that combining statistical descriptors, mathematical signal processing techniques, and machine-learning algorithms enables the extraction of informative degradation patterns from battery cycling data. Building upon these foundations, the present work advances the field by proposing a structured feature-engineering framework aimed at enhancing both the reliability and interpretability of State-of-Health (SOH) prediction models. Accurate SOH estimation for lithium-ion batteries has attracted considerable research interest due to its critical role in ensuring operational safety, performance efficiency, and long-term reliability across a wide range of applications. Existing SOH estimation techniques are commonly categorized into three broad classes: model-based, data-driven, and hybrid approaches. Each category offers distinct advantages while also presenting notable limitations. **Model-based approaches, often referred to as white-box models**, rely on electrochemical and physical principles to mathematically represent battery behavior and degradation mechanisms. These methods typically require detailed knowledge of internal chemical processes and parameter identification, which can be computationally demanding and difficult to obtain in practice. Although such models provide strong physical interpretability, their applicability is often constrained by modeling complexity and data availability. Reviews such as that by Spotnitz and Franklin provide comprehensive insights into the fundamental degradation mechanisms governing lithium-ion batteries. In contrast, **data-driven or black-box approaches** estimate SOH directly from historical operational data using machine-learning techniques, without explicitly modeling internal electrochemical reactions. These methods offer greater flexibility and scalability across different battery types and operating conditions. Recent studies have demonstrated the effectiveness of deep learning architectures, including Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs), in achieving high prediction accuracy and robustness. However, the limited interpretability of such models and their dependence on large, high-quality datasets remain significant challenges. **Hybrid or gray-box approaches** seek to combine the interpretability of model-based methods with the adaptability of data-driven techniques. By integrating physical insights with learning-based models, hybrid frameworks aim to overcome the shortcomings of purely white-box or black-box strategies. Emerging fusion-based techniques have shown promise in improving SOH estimation accuracy while maintaining model reliability and interpretability. Despite these

advancements, several challenges persist across all categories of SOH prediction methods. These include the need for extensive and representative datasets, the high computational cost of physics-based models, and the complexity associated with integrating hybrid frameworks. Furthermore, achieving robust generalization across varying battery chemistries, operating conditions, and aging mechanisms remains an open research problem. Among the many influencing factors, temperature plays a particularly significant role in accelerating degradation and affecting SOH prediction accuracy. Recent progress in machine-learning-based battery health management has led to the development of both cell-level and system-level SOH estimation strategies. Cell-level studies emphasize the importance of individualized modeling to accurately track degradation at the single-cell scale. For instance, works by Tian et al. and Hemdani et al. demonstrate the effectiveness of deep learning models in achieving reliable SOH prediction using cell-specific datasets. Accurate cell-level SOH estimation is essential for meeting stringent longevity requirements in automotive and stationary energy storage applications. Conversely, system-level approaches provide a broader assessment of battery health under diverse operational conditions and are particularly useful for evaluating compliance in large-scale battery systems.

1.1.2 The Lithium-Ion Battery

This section introduces the fundamental operating concepts of a battery by describing the primary elements that constitute a battery cell. A typical cell is composed of five core components, each of which performs a distinct role essential to the cell's overall operation and performance.

1.1.3 Components

- **Anode-The negative electrode:** The anode serves as the negative electrode during battery discharge and plays a crucial role in storing and releasing lithium ions. In lithium-ion batteries, the anode is typically composed of graphite due to its stable layered structure, which allows lithium ions to intercalate (insert) between its layers without causing significant structural damage. During the charging process, lithium ions migrate from the cathode through the electrolyte and become embedded within the anode material. During discharge, these ions are released back into the electrolyte, generating an electron flow through the external circuit. The performance of the anode directly influences key battery characteristics such as capacity, charging speed, and cycle life. However, over repeated cycles, issues such as solid electrolyte interphase (SEI) layer growth and lithium plating can occur, contributing to capacity degradation and reduced efficiency. Andrea2020.

- **Cathode-The positive electrode:** The cathode acts as the positive electrode during discharge and is the primary source of lithium ions within the battery. It is typically

made from lithium metal oxides such as lithium cobalt oxide (LiCoO), lithium iron phosphate (LiFePO), or nickel-manganese-cobalt (NMC) compounds. The cathode material determines the battery's voltage, energy density, and thermal stability. During discharge, lithium ions move from the anode to the cathode through the electrolyte, while electrons flow through the external circuit to provide electrical energy. During charging, this process is reversed. The structural integrity of the cathode is critical for long-term battery performance, as repeated cycling can lead to phase transitions, structural degradation, and loss of active material, all of which contribute to capacity fade and reduced state of health (SOH). Andrea2020.

- **Electrolyte:** The electrolyte is a chemically conductive medium that facilitates the movement of lithium ions between the anode and cathode while preventing the direct flow of electrons. It is typically composed of a lithium salt (such as LiPF) dissolved in an organic solvent mixture. The electrolyte plays a vital role in determining the ionic conductivity, operating temperature range, and safety characteristics of the battery. During battery operation, lithium ions travel through the electrolyte during both charging and discharging processes, enabling continuous electrochemical reactions. However, the electrolyte is also prone to decomposition at high voltages or temperatures, which can lead to the formation of unwanted byproducts, gas generation, and reduced battery efficiency. Its stability is therefore essential for ensuring safe and reliable battery operation. DellRand2001.

- **Separator:** The separator is a thin, porous membrane placed between the anode and cathode to prevent physical contact while allowing ionic transport. It is typically made from polymeric materials such as polyethylene or polypropylene. The separator's primary function is to act as an electrical insulator, thereby avoiding short circuits, while its porous structure enables lithium ions to pass through freely. The performance and safety of the battery are highly dependent on the separator's mechanical strength, thermal stability, and porosity. In cases of overheating, some separators are designed to shut down by melting and closing their pores, thereby stopping ion flow and preventing thermal runaway. Any damage or failure in the separator can lead to internal short circuits, making it a critical component for battery safety. DellRand2001.

- **Current Collectors:** Current collectors are conductive materials that connect the electrodes to the external circuit, allowing the flow of electrons during battery operation. Typically, copper is used as the current collector for the anode, while aluminum is used for the cathode due to their excellent electrical conductivity and compatibility with electrode materials. These collectors do not participate directly in electrochemical reactions but play an essential role in minimizing resistive losses and ensuring efficient energy transfer. The design and thickness of current collectors impact the overall weight, con-

ductivity, and efficiency of the battery. Poor contact or degradation of current collectors can lead to increased internal resistance and reduced performance.DellRand2001.

Together, these components collectively form a complete and functional electrochemical cell, which serves as the fundamental building block of any battery system. In practical applications, multiple such cells are interconnected in either series or parallel configurations, depending on the desired electrical output. A series configuration increases the overall voltage of the battery pack, while a parallel configuration enhances the total capacity and current delivery capability. These arrangements allow battery systems to be customized for a wide range of applications, from portable electronics to large-scale energy storage and electric vehicles.

In rechargeable batteries, particularly lithium-ion batteries, the internal architecture is carefully designed to support repeated and reversible electrochemical reactions during charge–discharge cycles. The electrode plates, comprising the anode and cathode materials, are typically arranged in a layered or stacked configuration. These layers are aligned in parallel, with electrodes of opposite polarity facing each other to maximize the effective surface area for electrochemical reactions. This structural arrangement ensures efficient electron flow through the external circuit and ion transport within the cell, both of which are critical for maintaining performance over extended cycling.

Separating the electrodes is a thin, porous separator, which plays a crucial dual role in battery operation. Firstly, it acts as an electrical insulator, preventing direct physical contact between the anode and cathode, which could otherwise lead to short circuits and potential safety hazards. Secondly, its porous structure allows the free passage of ions through the electrolyte, thereby enabling ionic conduction while maintaining electrical isolation. The electrolyte, which can be in liquid, gel, or solid form, serves as the medium for ion transport. During operation, lithium ions migrate through the electrolyte from the anode to the cathode during discharge, and in the reverse direction during charging.

Additionally, current collectors are attached to the electrodes to facilitate the efficient flow of electrons into and out of the external circuit. These collectors, typically made of conductive metals such as copper (for the anode) and aluminum (for the cathode), ensure minimal energy loss and contribute to the overall efficiency of the cell. The combined interaction of these components enables the battery to store, transfer, and deliver energy effectively. Over repeated cycles, however, electrochemical and mechanical changes occur within these components, leading to gradual performance degradation, which ultimately impacts the battery's capacity, power output, and lifespan. DellRand2001.

Figure 1.1 illustrates the basic structure of a battery cell and depicts the discharge process, during which ions migrate through the electrolyte while electrons flow through the external circuit. During charging, this process is reversed.

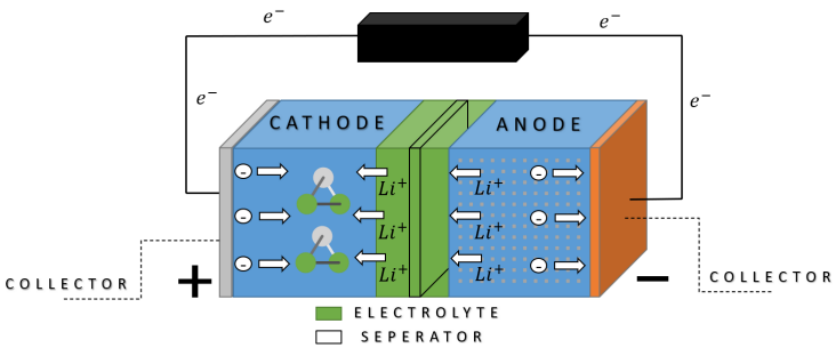

Figure 1: Illustration of a battery during discharge

Figure 1.1: Basic structure of a lithium-ion battery cell showing the discharge process.

1.1.4 Energy Principle

The operating principle of a lithium-ion battery is fundamentally governed by the difference in electrochemical potential between its two electrodes, namely the anode and the cathode. This potential difference arises due to variations in the chemical energy states of lithium ions within the host materials of each electrode. As discussed in [36], lithium ions at the anode exist in a relatively higher electrochemical energy state compared to those at the cathode, thereby creating a natural driving force for ion migration.

During the discharge process, an oxidation reaction occurs at the anode, where lithium atoms lose electrons and are converted into lithium ions. These ions then migrate through the electrolyte toward the cathode. Simultaneously, the released electrons cannot pass through the electrolyte and instead flow through the external circuit, generating an electric current that can be utilized to power external devices. At the cathode, a reduction reaction takes place, where lithium ions combine with incoming electrons and are incorporated into the cathode material. This coordinated movement of ions and electrons sustains the electrochemical process and enables continuous energy delivery.

Conversely, during the charging process, an external power source applies energy to reverse these reactions. Lithium ions are extracted from the cathode and driven back toward the anode through the electrolyte, while electrons flow in the opposite direction through the external circuit. This restores the original energy imbalance between the electrodes, effectively storing energy within the battery for future use.

The efficiency and reversibility of these redox reactions are critical for the performance of lithium-ion batteries. However, as highlighted by [36], repeated cycling leads to gradual changes in electrode structure, electrolyte stability, and interfacial properties. These changes contribute to performance degradation, reduced capacity, and ultimately a decline in the battery's state of health (SOH). Therefore, understanding the electrochemical operating principle is essential for analyzing battery behavior, modeling degradation mechanisms, and developing advanced prediction frameworks for battery health monitoring

1.1.5 Electrochemical Process

The electrochemical mechanisms governing lithium-ion battery operation are inherently complex and play a critical role in determining battery performance, safety, and longevity. A thorough understanding of these reactions is essential for selecting suitable electrode materials and electrolytes, as well as for minimizing degradation and preventing irreversible damage. Electrical energy generation within the cell arises from redox reactions occurring at the electrodes, where each electrode participates in a distinct half-cell reaction. During discharge, oxidation reactions at the negative electrode release electrons, which are then transported through the external circuit toward the positive electrode. This electron flow is driven by the electrochemical potential difference between the two electrodes and forms the basis of electrical energy delivery. The process is reversed during charging, when an external voltage source forces electrons to move back toward the negative electrode, restoring the original energy state of the system. The fundamental discharge reactions at the electrodes can be represented as follows[37]: **Anode (oxidation):**

$$M \longrightarrow M^{n+} + ne^{-} \tag{1.1}$$

Cathode (reduction):

$$nX + ne^{-} \longrightarrow nX^{-} \tag{1.2}$$

In these expressions, M and X denote the metal species and oxidizing agents, respectively, while e^{-} represents an electron involved in the electrochemical process [37].

1.2 Literature review and existing approaches

Recent advancements in battery health monitoring have increasingly leveraged machine learning and deep learning techniques to model degradation patterns and estimate State-of-Health (SOH). A common theme across existing research is the transformation of raw electrochemical data into meaningful representations through feature engineering, embedding, and clustering techniques.

Several studies, such as [38], utilize deep neural networks for implicit feature embedding, where multi-modal inputs including electrochemical, thermal, and mechanical signals are fused to learn representations. In such approaches, clustering is typically performed after feature extraction, enabling simultaneous SOH estimation and grouping of batteries based on learned characteristics.

Other works, including [39], adopt a two-stage pipeline where clustering is explicitly performed prior to prediction. In these methods, batteries are first grouped using centroid-based clustering techniques, and then cluster-specific models (e.g., LSTM) are trained.

This approach effectively handles heterogeneity in battery behavior, improving predictive performance by tailoring models to specific degradation patterns.

Feature engineering also plays a critical role in many studies. For instance, [40] employs Incremental Capacity Analysis (ICA) to extract degradation signatures from voltage curves. Instead of direct clustering, these engineered features implicitly capture latent groupings, allowing downstream models such as BiLSTM and ensemble learners to perform more robust predictions.

More advanced approaches focus on representation learning from raw signals. In [41], deep hybrid models extract latent embeddings from voltage relaxation curves, enabling implicit clustering of battery states. Similarly, [42] utilizes time-series embeddings with Deep Gaussian Processes, modeling degradation trajectories probabilistically without explicit clustering.

A growing trend in recent literature is the integration of autoencoders for dimensionality reduction and feature learning. Studies like [43] demonstrate how autoencoders can compress high-dimensional battery data into compact latent representations, improving clustering separability and noise robustness.

Building upon this idea, Deep Embedded Clustering (DEC), as discussed in [44], represents a significant advancement. DEC combines autoencoder-based feature learning with clustering optimization in a unified framework. By iteratively refining both the latent space and cluster assignments, DEC achieves superior clustering performance compared to traditional methods such as KMeans.

Table 1.2: Summary of Clustering-Based Approaches in Battery SoH Estimation

Ref	Clustering / Embedding Approach	Deep Learning Method	How Clustering is Applied	Key Contribution
[38]	Implicit feature embedding via deep neural networks	Multi-feature Deep Neural Network	Clustering performed after feature extraction using electrochemical, thermal, and mechanical signals	Joint SOH estimation and battery grouping using learned representations
[39]	Explicit centroid-based clustering	LSTM (Sequence Model)	Data is first clustered, followed by cluster-specific LSTM models	Improves prediction accuracy by handling heterogeneous battery behaviors
[40]	Feature-level grouping using degradation indicators (ICA-based)	BiLSTM + AdaBoost + PSO	No explicit clustering; ICA curves act as latent grouping mechanism	Captures degradation signatures instead of raw signals
[41]	Latent feature embedding from voltage relaxation curves	Deep hybrid learning model	Latent representations used for implicit clustering of battery states	Learns aging patterns directly from voltage dynamics
[42]	Time series embedding	Deep Gaussian Process + LSTM	No explicit clustering; relies on latent space modeling	Uses probabilistic embeddings for capacity prediction
[43]	Autoencoder-based representation learning	Deep Autoencoder	Clustering applied in learned latent space	Reduces dimensionality and improves cluster separability
[44]	Deep Embedded Clustering (DEC)	Autoencoder + Clustering Optimization	Joint optimization of embedding and clustering	Simultaneous feature learning and clustering improves cluster quality

1.2.1 Key observations from literature

A comprehensive review of existing literature reveals that battery health estimation and degradation modeling have increasingly transitioned toward data-driven and deep learning-based approaches. Most studies adopt a structured framework in which raw electrochemical measurements are transformed into informative representations through feature engineering and representation learning techniques. These representations are subsequently utilized for clustering or predictive modeling tasks.

A key observation is that clustering methodologies are applied at different stages of the ana-

lytical pipeline. In some approaches, clustering is performed prior to model development to partition the dataset into homogeneous groups, followed by the training of cluster-specific predictive models. This strategy effectively addresses the heterogeneity in battery degradation behavior. Alternatively, several studies perform clustering after feature extraction, leveraging engineered or learned features to group batteries based on degradation characteristics. More recent works incorporate implicit clustering through latent space representations, where deep learning models such as autoencoders learn compressed embeddings that naturally capture similarities among battery states without requiring explicit clustering mechanisms.

Another important trend is the increasing reliance on deep learning architectures, including recurrent neural networks and hybrid models, to capture nonlinear and temporal dependencies in battery degradation. These models demonstrate superior capability in learning complex patterns from high-dimensional data compared to traditional statistical methods. Additionally, feature engineering techniques such as incremental capacity analysis and voltage-based indicators have been widely used to encode electrochemical degradation phenomena into measurable quantities, thereby improving model interpretability and performance.

Despite these advancements, conventional clustering techniques such as K-Means remain limited due to their dependence on predefined feature spaces and inability to adapt to complex data distributions. This has led to the emergence of integrated frameworks, such as deep embedded clustering, which jointly optimize feature representation and clustering objectives. Such approaches provide a more robust and adaptive mechanism for identifying latent degradation patterns and battery health states.

1.2.2 Generalized Data-Driven Battery Health Modeling Pipeline

The overall framework adopted in existing studies, and further refined in this work, can be explicitly represented as:

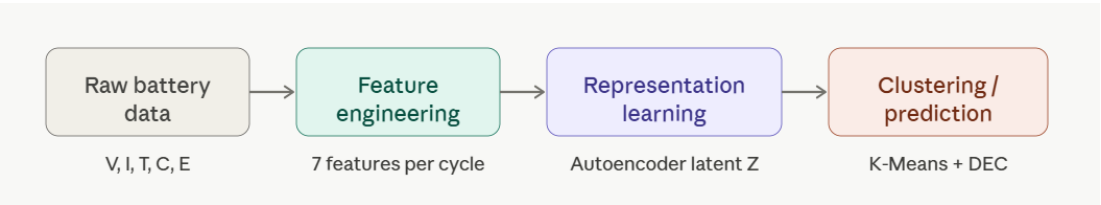

Figure 1.2: Battery SoH Estimation Pipeline: from raw data to clustering and prediction.

1. Raw Battery Data Acquisition

The process begins with the collection of raw time-series battery data, including parameters such as current, voltage, capacity, temperature, and cycle index. These measurements reflect the underlying electrochemical processes occurring during battery operation.

2. Feature Engineering

In this stage, domain knowledge is applied to extract meaningful indicators from raw data. Features such as capacity fade, voltage slope, efficiency, and state of health (SOH) are derived to capture the effects of electrochemical degradation. This step transforms raw signals into structured inputs suitable for machine learning models.

3. Representation Learning

To address high dimensionality and nonlinear relationships, representation learning techniques such as autoencoders are employed. These models compress input features into a low-dimensional latent space, preserving essential information while filtering noise. The latent representations encode hidden degradation patterns that are not directly observable in the original feature space.

4. Clustering / Health State Identification

Clustering algorithms are applied to the learned representations to group batteries into distinct health states. Traditional methods perform clustering in the original feature space, whereas advanced approaches such as Deep Embedded Clustering (DEC) jointly optimize feature learning and clustering, resulting in improved cluster separability and interpretability.

5. SOH Analysis and Interpretation

The final stage involves interpreting the identified clusters in terms of battery health conditions. Each cluster is associated with a specific degradation level (e.g., healthy, moderate, critical), enabling meaningful insights into battery aging behavior and supporting predictive maintenance strategies

This structured pipeline highlights the transition from physics-driven measurements to data-driven intelligence, enabling accurate and scalable modeling of battery degradation and health states.

1.3 Motivation

The accurate estimation and characterization of lithium-ion battery degradation remain critical challenges in modern energy storage systems, particularly in applications such as electric vehicles, grid storage, and portable electronics. Despite significant advancements in both physics-based modeling and data-driven techniques, existing approaches exhibit notable limitations in capturing the complex, nonlinear, and multi-scale nature of battery aging.

Physics-based electrochemical models, although grounded in fundamental principles, often require extensive parameterization and domain expertise. These models struggle to general-

ize across different battery chemistries, operating conditions, and usage patterns due to their reliance on idealized assumptions. Additionally, the computational complexity associated with solving coupled differential equations makes real-time deployment challenging.

On the other hand, conventional machine learning approaches primarily rely on handcrafted feature engineering followed by standard clustering or regression techniques. While these methods improve scalability, they are heavily dependent on the quality and completeness of engineered features. In many cases, important latent degradation patterns remain unobserved due to the inability of manual feature extraction to fully capture intricate temporal and electrochemical relationships. Furthermore, traditional clustering algorithms such as K-Means operate in the original feature space and assume linear separability, which limits their effectiveness when dealing with highly nonlinear degradation trajectories.

Recent deep learning-based approaches attempt to address these challenges by leveraging representation learning techniques such as autoencoders. These methods enable the transformation of high-dimensional battery data into compact latent representations. However, a significant limitation persists: clustering is often performed as a separate post-processing step, resulting in suboptimal alignment between learned features and clustering objectives. Consequently, the latent space may not be explicitly structured to enhance cluster separability, reducing interpretability and clustering performance.

Deep Embedded Clustering (DEC) emerges as a promising framework to overcome these limitations by jointly optimizing feature representation and clustering in a unified manner. By iteratively refining cluster assignments using soft labels and target distributions, DEC enhances the discriminative structure of the latent space. This leads to improved grouping of batteries based on degradation behavior without requiring explicit supervision. However, the application of DEC in battery health analysis remains relatively underexplored, particularly in conjunction with domain-specific feature engineering.

Motivated by these gaps, the proposed work aims to develop a hybrid framework that integrates physics-informed feature engineering with deep representation learning and clustering. By incorporating electrochemical insights into the feature extraction stage and leveraging autoencoder-based embeddings, the model captures both domain knowledge and hidden nonlinear patterns. The integration of DEC further ensures that the latent space is explicitly optimized for clustering, enabling more meaningful separation of battery health states.

This approach not only improves clustering accuracy but also enhances interpretability by linking latent clusters to physically meaningful indicators such as State of Health (SOH), capacity fade, and efficiency. Moreover, the proposed framework provides a scalable and data-driven alternative to traditional methods, making it suitable for real-world battery management systems where adaptability and robustness are essential.

In summary, the motivation for this work lies in bridging the gap between physics-based understanding and data-driven intelligence, while addressing the limitations of existing clustering and representation learning techniques. The proposed methodology aims to deliver a more accurate, interpretable, and scalable solution for battery degradation analysis and health state identification.

Chapter 2

Problem Formulation

2.1 Problem Statement

A major limitation in current data-driven approaches lies in their reliance on direct prediction frameworks, where engineered features are mapped to SOH values using regression or classification models. Such approaches fail to capture the underlying latent structure of degradation behavior, which may contain meaningful patterns representing different health states of the battery. Additionally, traditional clustering techniques, when used, are typically applied as a post-processing step and operate in the original feature space, limiting their ability to effectively separate complex, nonlinear degradation trajectories.

Recent advances in representation learning, particularly autoencoders, have enabled the extraction of compact latent features from high-dimensional battery data. However, in most existing works, clustering is still performed independently of representation learning, resulting in suboptimal alignment between learned features and clustering objectives. This separation restricts the model's ability to form well-defined and interpretable clusters corresponding to distinct battery health conditions.

To address these limitations, there is a need for a unified framework that can simultaneously learn meaningful feature representations and identify intrinsic degradation patterns without relying solely on labeled data. Specifically, an approach that integrates domain-informed feature engineering with deep representation learning and clustering optimization can provide improved interpretability and robustness in SOH analysis.

In this context, the problem addressed in this work is the development of a hybrid data-driven framework for lithium-ion battery degradation analysis that leverages physics-informed feature engineering, autoencoder-based latent representation learning, and Deep Embedded Clustering (DEC). The objective is to identify distinct battery health states by learning a structured latent space in which clustering is explicitly optimized using a Kullback–Leibler divergence-based objective function.

The novelty of the proposed approach lies in the integration of feature engineering and Deep Embedded Clustering within a unified pipeline, enabling simultaneous learning of representations and clustering of battery degradation patterns. Unlike conventional methods that treat clustering as a separate step, the proposed framework refines cluster assignments iteratively, leading to improved separation of health states and enhanced interpretability. Furthermore, by linking the identified clusters to SOH values, the model provides a meaningful mapping between latent degradation patterns and real-world battery health conditions.

Thus, this work aims to bridge the gap between traditional feature-based modeling and advanced deep clustering techniques, providing a scalable, interpretable, and data-driven solution for battery SOH estimation and degradation state identification.

Therefore, there is a need for effective data aggregation and modeling strategies that can manage sequential battery data while maintaining prediction accuracy. This work addresses this problem by evaluating machine-learning approaches, including regression and neural-network-based models, to improve reliable SOH forecasting.

2.2 Research objective

The primary objective of this study is to develop an advanced data-driven framework for analyzing lithium-ion battery degradation and estimating State-of-Health (SOH) through the integration of feature engineering, representation learning, and deep clustering techniques. To achieve this, the following specific objectives are defined:

1. To perform physics-informed feature engineering on battery cycling data by extracting meaningful degradation indicators such as capacity fade, SOH, efficiency, and voltage-based features.

2. To design and implement an autoencoder-based representation learning model capable of transforming high-dimensional battery data into a compact and informative latent space.

3. To apply Deep Embedded Clustering (DEC) for jointly optimizing feature representation and clustering, enabling the identification of intrinsic battery health states.

4. To evaluate the effectiveness of clustering using appropriate metrics such as Silhouette Score and Davies–Bouldin Index, ensuring meaningful separation of degradation patterns.

5. To establish a relationship between latent clusters and SOH values, enabling interpretable classification of battery health into categories such as healthy, moderate degradation, and critical condition.

6. To compare the proposed DEC-based clustering framework with traditional clustering methods (e.g., K-Means) to demonstrate performance improvements.

2.3 Battery Cell Degradation and Ageing Mechanisms Layout

The prediction of battery performance evolution and operational lifetime can be approached through multiple analytical perspectives. A fundamental and unavoidable contributor to battery performance deterioration is cell-level degradation. In lithium-ion batteries, degradation manifests across different regions of the cell and progressively alters several key performance characteristics. These include usable energy capacity, power delivery capability, and the effective service life of the battery.

As degradation advances, batteries may exhibit reduced power output, increased internal resistance, and operational instability, eventually leading to an end-of-life (EOL) condition Birkl2017. Battery degradation is governed by a combination of interacting mechanisms and external stressors, among which cycling history plays a dominant role. With repeated charge–discharge cycles, battery capacity steadily declines, and EOL is commonly defined when the available capacity falls to approximately 80% of its nominal value Borah2020.

This capacity loss arises from both cycle aging and calendar aging effects. Cycle aging is primarily driven by electrochemical and mechanical processes occurring during battery operation. Two major degradation pathways are widely recognized: loss of lithium inventory (LLI) and loss of active material (LAM). Although degradation reactions can take place throughout the cell, the anode and cathode are particularly susceptible to aging due to sustained electrochemical stress and material fatigue Xiong2020.

Chapter 3

Dataset and Pre-processing

3.1 Overview of the Dataset

The dataset used in this study consists of real-world operational records collected from a Lithium-ion cylindrical cell (a large-scale lithium-ion battery cycling dataset) undergoing repeated charge–discharge cycles. The data represents high-resolution measurements captured throughout the entire cycling process and includes electrical, thermal, temporal, and experimental cycle information.

Such datasets are essential for analyzing battery degradation, understanding state-of-health (SOH) evolution, and developing feature-engineering pipelines for SOH prediction. The richness of temporal measurements in this dataset makes it suitable for both data-driven machine learning approaches and physics-informed analysis.

The dataset captures the following key aspects:

- Dynamic behavior of the battery during each cycle
- Electrochemical responses such as voltage, capacity, and current
- Aging progression across multiple cycles
- Environmental or operational variables like temperature
- Time-series characteristics that allow extraction of degradation patterns

Overall, the dataset provides a comprehensive view of how the battery's performance evolves with usage, which is fundamental for SOH prediction and feature development.

3.2 Dataset Structure and Key Columns

The dataset contains the following variables, each carrying distinct physical significance for battery degradation analysis.

3.2.1 Test_Time (s)

The time elapsed in seconds from the start of the test. This variable serves as the primary temporal index for all measurements.

Relation to battery behavior:

- Helps identify charge and discharge phases

- Used to compute rate-of-change features such as $\mathrm{d}V/\mathrm{d}t$ and $\mathrm{d}I/\mathrm{d}t$
- Useful for extracting curve-based features including voltage profile and relaxation behavior

Relation to SOH: Time-based features help detect internal resistance changes, slower charging response, and voltage drop under load — all of which are indirect indicators of degradation.

3.2.2 Current (A)

The instantaneous charge or discharge current, where a positive value indicates charging and a negative value indicates discharging.

Effect on battery behavior:

- Lithium plating, which is dangerous at low temperatures and high charge currents
- Capacity fade due to repeated high-current cycling
- Internal resistance rise over prolonged operation

Impact on SOH: Higher current accelerates SEI (Solid Electrolyte Interphase) growth, induces mechanical stress in electrodes, and contributes to lithium inventory loss. Machine learning models frequently exploit current-based patterns to predict remaining useful life [2].

3.2.3 Capacity (Ah)

The instantaneous measured capacity at each point in the cycle. This is the most important feature for SOH estimation, as SOH is formally defined as:

$$\mathrm{SOH} = \frac{Q_{\mathrm{current}}}{Q_{\mathrm{rated}}} \times 100\% \tag{3.1}$$

where Q_{current} is the present maximum deliverable capacity and Q_{rated} is the original nominal capacity.

Relation to SOH: As the battery degrades, capacity decreases and SOH drops proportionally. It is therefore central to SOH prediction, cycle life estimation, and end-of-life forecasting.

3.2.4 Voltage (V)

The instantaneous terminal voltage of the battery. Voltage curves change subtly but consistently as the battery ages:

- Higher internal resistance causes larger voltage drops under load
- Plateau regions in the voltage curve shift with degradation
- $\mathrm{d}V/\mathrm{d}t$ patterns become distorted over cycles

Impact on SOH: The voltage profile is one of the strongest predictors of internal resistance rise, degradation mechanisms, and both cycle aging and calendar aging. Many SOH estimation models are built entirely on voltage curve analysis [2].

3.2.5 Energy (Wh)

The real-time energy delivered or absorbed by the battery, computed as:

$$E = \int_{t_0}^{t_f} V(t)\, I(t)\, \mathrm{d}t \tag{3.2}$$

Energy output decreases over cycles due to:

- Declining Coulombic and energy efficiency
- Reduction in usable capacity
- Progressive increase in internal resistance

Impact on SOH: A declining energy output directly reflects loss of usable capacity, reduced charge acceptance, and cumulative cycle aging.

3.2.6 Temperature (°C)

The cell surface or internal temperature during operation. Temperature is one of the most significant external drivers of battery degradation:

- Elevated temperature accelerates SEI layer formation
- Low temperature induces lithium plating on the anode
- Thermal fluctuations introduce mechanical stress in electrode materials

Effects on SOH: Elevated temperatures lead to faster capacity fade, increased internal resistance, and permanent structural damage to electrode materials. ML models consistently treat temperature as a core feature for accurate SOH prediction [6].

3.2.7 Cycle Index

An integer index indicating which battery cycle a given data sample belongs to. This is the most critical variable for aging analysis.

Observed trends across cycles:

- SOH decreases monotonically with cycle number
- Voltage curves progressively flatten
- Capacity declines and internal resistance rises

Impact on SOH: Cycle index enables degradation curve fitting, Remaining Useful Life (RUL) estimation, and cycle-based feature engineering — making it indispensable for any data-driven battery health model.

3.3 Significance of this Dataset for Battery Research

This dataset is well-suited for a wide range of analytical and modelling tasks, including:

- **Feature engineering** — for example, voltage relaxation features, energy efficiency ratios, and cycle-level summary statistics
- **SOH trajectory estimation** across the full cycle life of the cell
- **Predictive modelling** using supervised machine learning regression models
- **Degradation mechanism analysis** through voltage curves, capacity fade trends, and temperature effects
- **Hybrid physics–data-driven model development** combining electrochemical priors with learned features

Its high temporal resolution and cycle-based labelling make it particularly ideal for constructing new features that are mathematically grounded and physically interpretable.

3.4 Overview of the Dataset

The dataset used in this study consists of real-world operational records collected from a Lithium-ion cylindrical cell (a large-scale lithium-ion battery cycling dataset) across two experimental files, subsequently concatenated into a unified dataframe for analysis. The combined dataset comprises **329,164 measurement records** spanning **101 complete charge–discharge cycles**, collected between 14 December 2020 and 18 December 2020. File one contributes 202,516 records and file two contributes 126,648 records, with no missing values across any column.

Each record captures nine variables sampled at approximately one-second intervals, yielding an average of 3,259 data points per cycle (minimum: 2,127; maximum: 5,181). This high

temporal resolution makes the dataset well-suited for time-series feature extraction and degradation-sensitive modelling.

3.4.1 Dataset Variables and Statistical Summary

Table 3.1 presents the descriptive statistics for the key measurement variables across the full combined dataset.

Table 3.1: Descriptive Statistics of the Combined a large-scale lithium-ion battery cycling dataset Dataset (329,164 records, 101 cycles)

Variable	Min	Max	Mean	Std Dev
Current (A)	−7.207	7.207	−0.028	4.488
Capacity (Ah)	0.000	2.271	0.617	0.417
Voltage (V)	2.999	4.201	3.719	0.302
Energy (Wh)	0.000	8.111	2.312	1.576
Cycle Index	1	101	43.45	29.11

3.4.2 Correlation Analysis

To understand the inter-variable relationships within the dataset, a Pearson correlation matrix was computed across the five key numerical variables: Current (A), Capacity (Ah), Voltage (V), Energy (Wh), and Cycle_Index. Figure 3.1 presents the resulting heatmap.

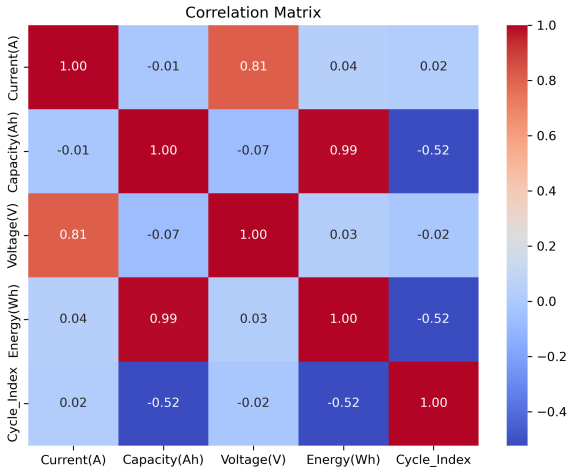

Figure 3.1: Pearson correlation matrix of key dataset variables for the a large-scale lithium-ion battery cycling dataset lithium-ion cell (combined dataset, 329,164 records).

3.4.2.1 Key Observations from the Correlation Matrix

The following significant relationships are identified:

- **Current and Voltage (r = 0.81):** A strong positive correlation exists between current

and voltage. This is physically consistent — during charging, both current and terminal voltage are elevated simultaneously, while during discharge, both decline together.

- **Capacity and Energy (r = 0.99):** A near-perfect positive correlation is observed between capacity and energy. This is expected, since energy is the product of voltage and capacity integrated over time — as capacity increases during charging, energy accumulates proportionally.
- **Capacity and Cycle_Index (r = −0.52):** A moderate negative correlation exists between capacity and cycle index. This directly reflects battery degradation — as the cycle number increases, the maximum deliverable capacity progressively declines, confirming the aging trend observed across the 101 cycles.
- **Energy and Cycle_Index (r = −0.52):** Similarly, energy output shows a moderate negative correlation with cycle index, consistent with the capacity fade and rising internal resistance over the battery's operational life.
- **Current and Capacity (r = −0.01):** Effectively zero correlation, indicating that instantaneous current magnitude does not directly predict instantaneous capacity at the raw measurement level.
- **Voltage and Cycle_Index (r = −0.02):** Near-zero raw correlation between voltage and cycle index at the instantaneous measurement level. This does not imply voltage is unimportant — rather, degradation-related voltage changes manifest in the *shape* of voltage curves rather than in instantaneous values, which is why curve-based features such as $\mathrm{d}V/\mathrm{d}t$ and voltage plateau analysis are essential for SOH estimation.

3.4.2.2 Implications for Feature Engineering

The correlation analysis provides two important insights for the feature engineering pipeline:

1. **Capacity and Energy are highly redundant** (r = 0.99). Including both raw features in a machine learning model would introduce multicollinearity. It is therefore preferable to use derived features such as Coulombic efficiency $\eta = Q_{\text{discharge}}/Q_{\text{charge}}$ and energy efficiency $\eta_E = E_{\text{discharge}}/E_{\text{charge}}$ rather than raw capacity and energy values directly.
2. **Cycle_Index has moderate linear correlation with degradation indicators** (r = −0.52 with both capacity and energy), but the relationship is non-linear in nature. This motivates the use of non-linear machine learning models such as Support Vector Regression (SVR) and Gradient Boosting rather than simple linear regression for SOH prediction.

3.4.3 Univariate and Bivariate Data Analysis

To further explore the distributional properties and pairwise relationships among the dataset variables, a pairplot was constructed comprising both univariate analysis (diagonal elements) and bivariate analysis (off-diagonal scatter plots). Figure **??** presents the complete pairplot for Current (A), Capacity (Ah), Voltage (V), Energy (Wh), and Cycle_Index.

3.4.3.1 Univariate Analysis (Diagonal)

The diagonal elements display the marginal distribution of each variable individually:

- **Current (A):** The distribution shows discrete clusters at approximately −7.2 A, −4.8 A, 0 A, +4.8 A, and +7.2 A, reflecting the fixed charge and discharge current levels used in the cycling protocol. The symmetric distribution around zero confirms that charge and discharge cycles are approximately balanced in duration.
- **Capacity (Ah):** The distribution is concentrated near zero and rises gradually up to 2.271 Ah, reflecting the accumulation of capacity during each cycle. The skewed shape indicates that most measurements are captured during the early-to-mid portion of each charge or discharge phase.
- **Voltage (V):** The distribution shows a sharp concentration near 3.0 V with a long tail extending up to 4.2 V. This reflects the cutoff voltage limits of the cycling protocol, with the majority of measurements occurring in the mid-voltage discharge region between 3.4 V and 4.0 V.
- **Energy (Wh):** Similar in shape to capacity, the energy distribution is right-skewed, peaking near zero and tapering toward 8.111 Wh. This mirrors the charge accumulation pattern observed in the capacity distribution, consistent with the near-perfect correlation (r = 0.99) between the two variables.
- **Cycle_Index:** The distribution appears approximately uniform across cycles 1 to 101, with slight variations reflecting differences in the number of data points recorded per cycle (ranging from 2,127 to 5,181 points per cycle).

3.4.3.2 Bivariate Analysis (Off-Diagonal)

The off-diagonal scatter plots reveal the pairwise relationships between variables:

- **Capacity vs. Voltage:** The scatter plot shows a characteristic butterfly-shaped pattern, reflecting the distinct voltage profiles during charging (rising voltage) and discharging (falling voltage). The spread of this pattern across cycles reveals subtle curve shifts consistent with progressive degradation.

- **Capacity vs. Energy (and vice versa):** A near-linear relationship is visible, confirming the very high correlation (r = 0.99) between these two variables. The slight spread around the diagonal reflects cycle-to-cycle efficiency variations.
- **Capacity vs. Cycle_Index:** The plot shows a clear downward curving pattern — higher cycle indices are associated with lower maximum capacity values, visually confirming the 80.14% capacity fade observed across 101 cycles.
- **Energy vs. Cycle_Index:** Similar to the capacity trend, energy output decreases with increasing cycle index, displaying a curved degradation trajectory consistent with accelerating aging in later cycles.
- **Voltage vs. Cycle_Index:** The scatter shows multiple horizontal bands corresponding to fixed voltage levels in the cycling protocol. No strong linear trend is visible at the raw measurement level, confirming that voltage degradation manifests in curve shape rather than in instantaneous values — further motivating the use of $\mathrm{d}V/\mathrm{d}t$ and differential capacity features for SOH estimation.
- **Current vs. Voltage:** Discrete horizontal bands in current correspond to the fixed current levels of the cycling protocol, spread across the full voltage operating range of 3.0 V to 4.2 V.

3.4.3.3 Summary of Analytical Insights

The combined univariate and bivariate analysis reveals three key insights relevant to the feature engineering framework:

1. The cycling protocol uses **fixed discrete current levels**, meaning current alone is insufficient as a degradation indicator — derived features such as internal resistance estimates are more informative.
2. **Capacity and Energy are near-redundant** at the raw level, reinforcing the decision to use efficiency ratios as engineered features rather than raw values.
3. The **non-linear degradation trajectory** visible in the Capacity–Cycle_Index and Energy–Cycle_Index plots confirms that non-linear machine learning models are required for accurate SOH prediction across the full cycle life.

Chapter 4

Feature Engineering

Feature engineering plays a crucial role in transforming raw battery cycling data into meaningful representations that capture degradation behavior. Lithium-ion battery degradation is governed by complex electrochemical, thermal, and operational processes, which are not directly observable from raw measurements. Therefore, carefully designed features are required to extract underlying patterns that are relevant for State-of-Health (SOH) estimation and degradation analysis.

In this study, a combination of physics-informed, statistical, and model-based features is developed to represent battery behavior across charge–discharge cycles. These features serve as the foundation for subsequent representation learning using autoencoders and clustering via Deep Embedded Clustering (DEC).

The complete implementation of the feature engineering pipeline is presented in Listing 4.1.

Listing 4.1:). All seven degradation indicators are computed from raw cycle-level measurements and assembled into the feature matrix $\mathbf{X} \in \mathbb{R}^{101\times7}$.]Physics-informed feature engineering pipeline (Jupyter Cell [18]). All seven degradation indicators are computed from raw cycle-level measurements and assembled into the feature matrix $\mathbf{X} \in \mathbb{R}^{101\times7}$.

```
# Sorting data
data = data.sort_values(by=['Cycle_Index', 'Test_Time(s)'])

# 1. Capacity Fade (degradation feature)
initial_capacity = data.groupby('Cycle_Index')[
                    'Capacity(Ah)'].max().iloc[0]
cycle_capacity = data.groupby('Cycle_Index')[
                  'Capacity(Ah)'].max().reset_index()
cycle_capacity['Capacity_Fade'] = (initial_capacity
                - cycle_capacity['Capacity(Ah)'])

# 2. SOH
cycle_capacity['SOH'] = (cycle_capacity['Capacity(Ah)']
                / initial_capacity) * 100

# 3. Rate of degradation (derivative)
cycle_capacity['dSOH'] = cycle_capacity['SOH'].diff().fillna(0)

# 4. Voltage slope (per cycle)
voltage_slope = data.groupby('Cycle_Index')['Voltage(V)'].apply(
                lambda x: np.gradient(x).mean())
cycle_capacity['Voltage_Slope'] = voltage_slope.values

```

```
# 5. Energy efficiency
energy = data.groupby('Cycle_Index')[
         'Energy(Wh)'].max().reset_index()
cycle_capacity['Energy'] = energy['Energy(Wh)']
cycle_capacity['Efficiency'] = (cycle_capacity['Capacity(Ah)']
                / cycle_capacity['Energy'])

# 6. Temperature effect
temp = data.groupby('Cycle_Index')[
       'Temperature(Celsius)'].mean().reset_index()
cycle_capacity['Temp'] = temp['Temperature(Celsius)']

# Final feature matrix X
features = cycle_capacity[['Capacity(Ah)', 'Capacity_Fade',
           'SOH', 'dSOH', 'Voltage_Slope', 'Efficiency', 'Temp']]
```

4.1 Preprocessing and C-Rate Based Stabilization

Battery degradation behavior is significantly influenced by the C-rate, defined as the rate at which a battery is charged or discharged relative to its nominal capacity. Variations in C-rate introduce inconsistencies in the observed voltage, capacity, and temperature profiles across cycles, which can obscure true degradation trends if not properly accounted for.

4.1.1 Definition of C-Rate

The C-rate is mathematically defined as:

$$\text{C-rate} = \frac{I}{Q_{\text{nominal}}} \tag{4.1}$$

where I is the applied current in amperes (A) and Q_{nominal} is the nominal capacity of the cell in ampere-hours (Ah). For the a large-scale lithium-ion battery cycling dataset cell used in this study, $Q_{\text{nominal}} = 2.271$ Ah (the maximum capacity observed at Cycle 1). The dataset exhibits discrete current levels of ±4.807 A and ±7.207 A, corresponding to approximately 2C and 3C rates respectively.

To ensure consistency across cycles, the dataset is filtered to consider a fixed C-rate regime, reducing variability introduced by different operating conditions and enabling a more reliable comparison of degradation indicators across cycle indices.

4.1.2 Polynomial Regression for Capacity Prediction

As a preliminary modelling step, polynomial regression was applied to predict battery capacity over future cycle indices for each distinct C-rate present in the dataset. The cycle

index serves as the independent variable X and the maximum discharge capacity per cycle serves as the dependent variable y.

The dataset was first grouped by rounded C-rate values then,

For each unique C-rate, a degree-3 polynomial of the form:

$$C_n \approx \beta_0 + \beta_1 n + \beta_2 n^2 + \beta_3 n^3 \tag{4.2}$$

was fitted to the capacity–cycle data, and predictions were extended 20 cycles beyond the observed range (up to cycle 120).

4.1.2.1 Polynomial Regression Results by C-Rate

Figure 4.1 and Figure 4.2 show representative polynomial regression fits for two distinct C-rate values observed in the dataset.

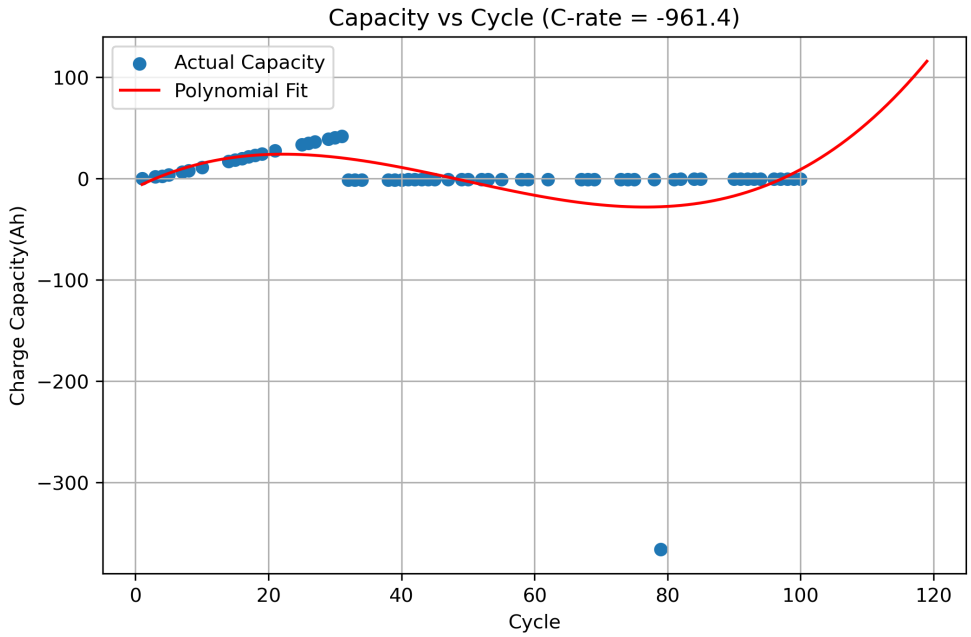

Figure 4.1: Polynomial regression fit (degree 3) for C-rate = −961.4. Blue dots represent actual measured capacity per cycle; red curve shows the fitted polynomial extrapolated to 120 cycles. The U-shaped curve indicates high variance in capacity measurements at this C-rate, suggesting unstable cycling conditions.

4.1.2.2 SOH Computation from Predicted Capacity

Following capacity prediction, the State-of-Health percentage was computed for each cycle using:

$$\mathrm{SOH}_n(\%) = \frac{C_n}{C_{\text{initial}}} \times 100 \tag{4.3}$$

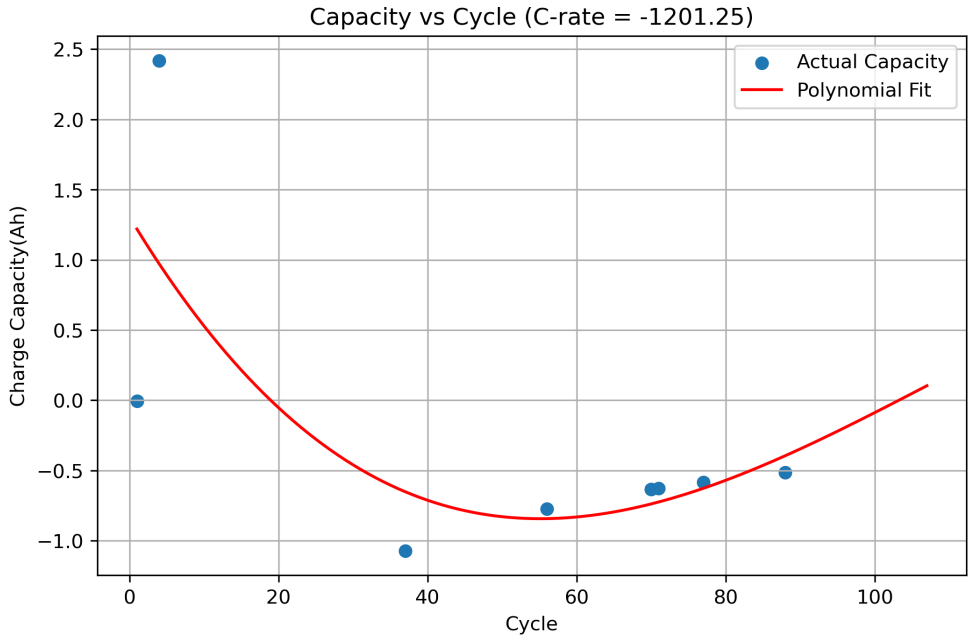

Figure 4.2: Polynomial regression fit (degree 3) for C-rate = −1201.25. The polynomial captures an initial rise followed by a sharp decline beyond cycle 80, predicting severely negative capacity values by cycle 120 — a physically unrealistic extrapolation indicating overfitting of the polynomial model at extreme C-rates.

where $C_{\text{initial}} = 2.664$ Ah is the first non-zero maximum capacity recorded in the dataset (Cycle 2). Note that Cycle 1 was excluded as its recorded capacity is zero, representing an initialization step rather than a true charge–discharge cycle.

Figure 4.3 presents the resulting SOH trajectory across all 101 cycles.

Observation: The SOH plot reveals two critical issues with the raw polynomial regression approach:

1. **Anomalous SOH values** exceeding 35,000% in early cycles arise because the capacity column contains instantaneous cumulative values rather than per-cycle discharge totals. This indicates that without proper cycle-level aggregation and normalization, the SOH formula produces physically meaningless results.

2. **Discontinuity between cycles 30–60** reflects the two-file dataset structure, where the second file resets certain counters, introducing a step change in the capacity signal.

4.1.2.3 Limitations of Polynomial Regression

Despite its mathematical simplicity and interpretability, polynomial regression was found to be insufficient as the primary modelling approach for the following reasons:

- **Overfitting at extreme C-rates:** As visible in Figure 4.2, the degree-3 polynomial produces physically unrealistic predictions (negative capacity) beyond the observed

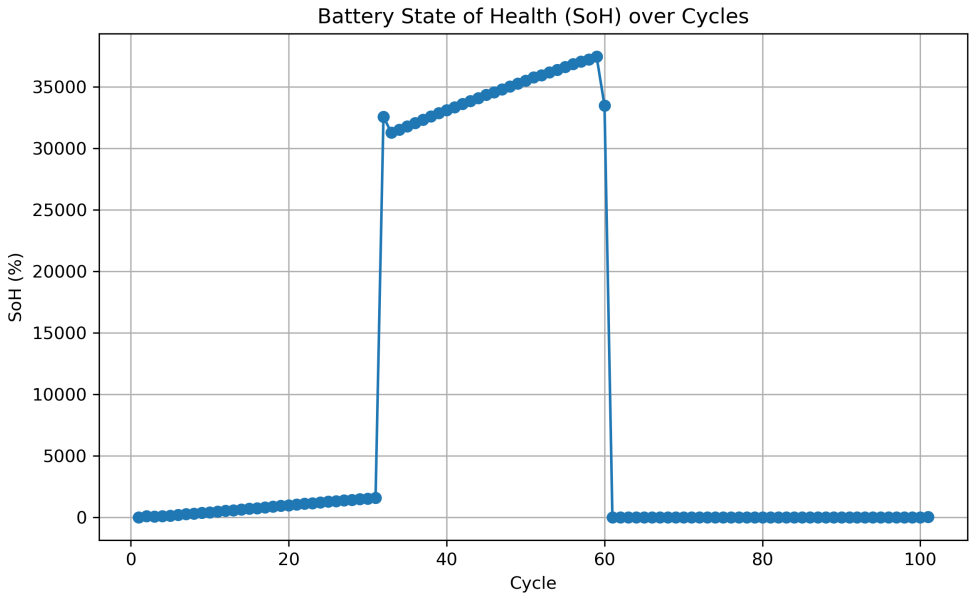

Figure 4.3: Battery State-of-Health (%) over 101 cycles computed from the ratio of per-cycle maximum capacity to initial capacity (C_{initial} = 2.664 Ah). Two distinct operating regimes are visible: cycles 1–30 show anomalously high SOH values (up to 35,000%) due to the use of raw instantaneous capacity rather than normalized discharge capacity, and cycles 60–101 show near-zero SOH reflecting severe degradation. This result motivated the need for a more robust feature engineering and modelling approach.

cycle range, demonstrating poor extrapolation behaviour.

- **C-rate dependency:** A separate model must be fitted for each unique C-rate, making the approach non-generalizable. In the dataset, the number of unique rounded C-rates is large, making this computationally impractical for a unified SOH estimation framework.
- **No feature interaction:** Polynomial regression uses only cycle index as input, ignoring the rich multi-variable degradation information available in voltage, temperature, energy, and current signals.
- **Linear coefficient assumption:** Although polynomial features introduce non-linearity in n, the model coefficients β_j remain linear, limiting the model's ability to capture complex electrochemical degradation dynamics.
- **No uncertainty quantification:** The model provides point predictions without confidence intervals, which is unsuitable for safety-critical battery management applications.

These limitations collectively motivate the transition to a more powerful and generalizable framework combining deep feature learning via autoencoders with unsupervised clustering through Deep Embedded Clustering (DEC), as developed in the following chapters. Unlike polynomial regression, the autoencoder–DEC pipeline:

1. Operates on the full multi-variable feature matrix $\mathbf{X} \in \mathbb{R}^{101\times 7}$

2. Learns non-linear degradation representations without assuming any functional form
3. Generalizes across C-rates and battery conditions through latent space clustering
4. Provides interpretable cluster assignments corresponding to distinct degradation states

4.2 Engineered Features

The following features are constructed from the raw cycling measurements for each cycle n. Each feature is designed to capture a specific aspect of battery degradation behavior.

4.2.1 Capacity and Capacity Fade

Let C_0 denote the initial capacity at Cycle 1 and C_n the maximum discharge capacity at cycle n. The **Capacity Fade** is defined as the absolute reduction in capacity from the initial value:

$$\text{Capacity Fade}_n = C_0 - C_n \tag{4.4}$$

For the a large-scale lithium-ion battery cycling dataset dataset, $C_0 = 2.271$ Ah, and by Cycle 101 the capacity fade reaches $2.271 - 0.451 = 1.820$ Ah, representing an 80.14% reduction.

Physical interpretation: Capacity fade is a monotonically increasing function under normal aging conditions. It directly quantifies the cumulative loss of lithium inventory and active material resulting from repeated charge–discharge cycling. As the SEI layer grows and electrode particles fracture, fewer sites are available for lithium intercalation, causing a progressive and irreversible decline in usable capacity.

4.2.2 State-of-Health (SOH)

The State-of-Health at cycle n is defined as the ratio of the current capacity to the initial capacity, expressed as a percentage:

$$\text{SOH}_n = \frac{C_n}{C_0} \times 100\% \tag{4.5}$$

Interpretation of SOH values:

- SOH $\approx$ 100% indicates a healthy, near-new battery
- SOH $\approx$ 70–80% corresponds to the conventional end-of-life threshold
- SOH $<$ 70% indicates severe degradation requiring replacement

Mathematical note: SOH is a normalized projection of capacity onto a bounded interval:

$$\text{SOH} \in (0,\ 100] \tag{4.6}$$

This bounded normalization makes SOH a more stable and interpretable regression target than raw capacity, and is therefore used as the primary prediction output in the machine learning models developed in this study.

4.2.3 Rate of Degradation (ΔSOH)

The rate of degradation is defined as the first-order discrete difference of SOH between consecutive cycles:

$$\Delta\text{SOH}_n = \text{SOH}_n - \text{SOH}_{n-1} \tag{4.7}$$

Physical interpretation: ΔSOH_n captures the local degradation dynamics at each cycle. Under normal aging, this value is small and negative, indicating gradual steady decline. A sudden large negative spike in ΔSOH_n indicates an accelerated degradation event, such as lithium plating or an abrupt increase in internal resistance. This feature is therefore particularly sensitive to anomalous degradation behaviour and serves as an early warning indicator of accelerated aging.

4.2.4 Voltage Slope

Let $V(t)$ denote the terminal voltage measured at time t during a cycle of duration T. The **Voltage Slope** for cycle n is defined as the time-averaged rate of voltage change:

$$\text{Voltage Slope}_n = \frac{1}{T}\int_0^T \frac{\mathrm{d}V}{\mathrm{d}t}\,\mathrm{d}t \tag{4.8}$$

In practice, this integral is approximated numerically using finite differences:

$$\left.\frac{\mathrm{d}V}{\mathrm{d}t}\right|_t \approx \frac{V(t+\Delta t) - V(t)}{\Delta t} \tag{4.9}$$

Physical interpretation: The voltage slope reflects the electrochemical kinetics of the cell. As the battery ages, increasing internal resistance causes steeper voltage drops under load, altering the slope profile. Changes in voltage slope therefore serve as indirect indicators of internal resistance rise, lithium plating, and other aging mechanisms. This feature captures degradation information that is not visible in instantaneous voltage measurements alone.

4.2.5 Energy Efficiency

Let E_n denote the total energy delivered or absorbed during cycle n in watt-hours (Wh). The **Energy Efficiency** feature is defined as:

$$\text{Efficiency}_n = \frac{C_n}{E_n} \tag{4.10}$$

Physical interpretation: This ratio measures the capacity delivered per unit of energy consumed, providing a measure of the electrochemical efficiency of each cycle. As the battery degrades, resistive losses and heat generation increase, causing energy efficiency to decline even when capacity fade is moderate. This feature therefore captures degradation information complementary to raw capacity, particularly in the early stages of aging where capacity fade is still small but efficiency losses are already detectable.

4.2.6 Temperature Influence

The mean cycle temperature T_n is computed as the arithmetic mean of all temperature measurements recorded during cycle n:

$$T_n = \frac{1}{k} \sum_{i=1}^{k} T_i \tag{4.11}$$

where k is the number of temperature measurements in cycle n and T_i is the i-th temperature reading in degrees Celsius.

Physical interpretation: Temperature is one of the dominant external drivers of battery degradation. Elevated temperatures accelerate SEI layer formation and side reactions, while low temperatures induce lithium plating. The mathematical coupling between temperature and degradation rate is non-linear:

$$\frac{\mathrm{d}C}{\mathrm{d}n} \propto f(T_n) \tag{4.12}$$

where $f(T_n)$ is a non-linear function of temperature, often modelled using an Arrhenius-type relationship in electrochemical models. Including mean cycle temperature as a feature therefore enables the machine learning model to implicitly account for thermally-driven degradation variations across cycles.

4.3 Feature Space Representation

The complete engineered feature vector for each cycle n is defined as:

$$\mathbf{x}_n = \left[C_n,\ \text{Capacity Fade}_n,\ \text{SOH}_n,\ \Delta\text{SOH}_n,\ \text{Voltage Slope}_n,\ \text{Efficiency}_n,\ T_n \right]^\top \quad (4.13)$$

Collecting the feature vectors across all N cycles forms the feature matrix:

$$\mathbf{X} \in \mathbb{R}^{N \times d}, \quad d = 7 \quad (4.14)$$

where N is the total number of cycles and $d = 7$ is the number of engineered features. For this study, $N = 101$ cycles, yielding a feature matrix $\mathbf{X} \in \mathbb{R}^{101 \times 7}$.

Table 4.1 provides a consolidated summary of all engineered features, their mathematical definitions, and their physical significance.

Table 4.1: Summary of Engineered Features for SOH Estimation

Feature	**Definition**	**Physical Significance**
Capacity C_n	Max discharge capacity at cycle n	Primary degradation indicator
Capacity Fade	$C_0 - C_n$	Cumulative capacity loss
SOH_n	$\frac{C_n}{C_0} \times 100\%$	Normalized health indicator
ΔSOH_n	$\text{SOH}_n - \text{SOH}_{n-1}$	Local degradation rate
Voltage Slope	$\frac{1}{T}\int_0^T \frac{dV}{dt}\, dt$	Internal resistance changes
Efficiency_n	$\frac{C_n}{E_n}$	Energy utilization efficiency
Temperature T_n	$\frac{1}{k}\sum_{i=1}^{k} T_i$	Thermal degradation driver

4.4 Role in Dimensionality Reduction

4.4.1 Motivation for Dimensionality Reduction

The engineered feature matrix $\mathbf{X} \in \mathbb{R}^{N \times 7}$, while compact relative to raw time-series data, still exhibits properties that are suboptimal for direct input to clustering algorithms:

- **Correlation:** Features such as Capacity, Capacity Fade, and SOH are mathematically related by definition, introducing redundancy into the feature space
- **Redundancy:** Near-collinear features reduce the effective dimensionality of the input space without adding discriminative information

- **Non-linear structure:** The underlying degradation manifold is non-linear and cannot be adequately captured by linear dimensionality reduction methods such as PCA

To address these limitations, the feature matrix $\mathbf{X}$ is transformed using a deep autoencoder to produce a compact latent representation.

4.4.2 Encoder Mapping

The encoder network f_θ maps each feature vector $\mathbf{x}_n$ to a lower-dimensional latent vector $\mathbf{z}_n$:

$$\mathbf{z}_n = f_\theta(\mathbf{x}_n), \quad \mathbf{z}_n \in \mathbb{R}^k, \quad k < d \tag{4.15}$$

where f_θ is a non-linear transformation implemented as a neural network with learnable parameters θ, and k is the latent dimension satisfying $k < d = 7$.

Interpretation of the latent vector $\mathbf{z}_n$:

- Captures the intrinsic degradation patterns of the battery at cycle n
- Represents a compressed manifold of battery behavior, discarding noise and redundant correlations
- Provides a structured low-dimensional embedding suitable for downstream clustering

4.4.3 Autoencoder Reconstruction Objective

The full autoencoder comprises an encoder f_θ and a decoder g_ϕ. The network is trained by minimizing the mean squared reconstruction loss:

$$\min_{\theta,\phi} \sum_{n=1}^{N} \left\| \mathbf{x}_n - g_\phi(f_\theta(\mathbf{x}_n)) \right\|^2 \tag{4.16}$$

where:

- $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ is the **encoder** performing dimensionality reduction
- $g_\phi : \mathbb{R}^k \to \mathbb{R}^d$ is the **decoder** performing reconstruction
- Minimizing this loss forces the latent space to retain only the most essential information required to reconstruct the original feature vector

4.4.4 Connection to Deep Embedded Clustering (DEC)

Following autoencoder training, the latent vectors $\{\mathbf{z}_n\}_{n=1}^N$ are passed to the Deep Embedded Clustering (DEC) module for unsupervised cluster assignment:

$$\mathbf{z}_n \longrightarrow \text{cluster assignments} \tag{4.17}$$

The use of latent representations rather than raw features for clustering is motivated by two key properties:

1. **Structured latent space:** The autoencoder training organizes the latent space such that degradation states with similar behavior are mapped to nearby points, improving cluster separability

2. **Noise removal:** The reconstruction objective implicitly filters out measurement noise, yielding cleaner cluster boundaries in the latent space

4.5 Feature Engineering and Dimensionality Reduction Pipeline

Figure 4.4 presents a comprehensive overview of the complete feature engineering and dimensionality reduction pipeline, illustrating the transformation from raw battery cycling data to structured latent representations suitable for clustering and SOH prediction.

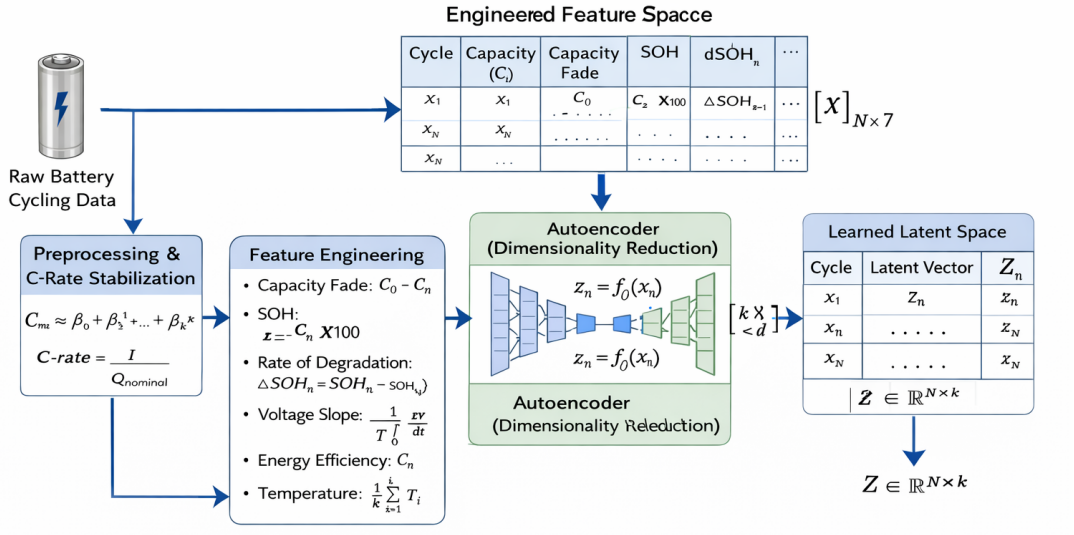

Fig. 1. Overview of Feature Engineering and Dimensionality Reduction Process.

Figure 4.4: Overview of the Feature Engineering and Dimensionality Reduction Process. Raw battery cycling data undergoes preprocessing and C-rate stabilization, followed by extraction of seven engineered features into a structured feature matrix $\mathbf{X} \in \mathbb{R}^{N\times 7}$. The feature matrix is then passed through an autoencoder for dimensionality reduction, producing a compact latent representation $\mathbf{Z} \in \mathbb{R}^{N\times k}$, where $k < 7$, suitable for downstream clustering and SOH state identification.

4.6 Summary

The feature engineering process developed in this chapter transforms raw battery cycling data into a structured $\mathbb{R}^{101\times7}$ feature matrix capturing degradation dynamics from multiple perspectives: capacity loss, electrochemical behavior, energy efficiency, and thermal effects. These features not only enhance physical interpretability but also provide a suitable input space for deep learning models.

A preliminary capacity prediction model using polynomial regression demonstrated the non-linear nature of degradation trajectories and provided a baseline for future cycle capacity estimation at fixed C-rates. The subsequent dimensionality reduction using a deep autoencoder enables extraction of compact latent representations $\mathbf{z}_n \in \mathbb{R}^k$, which are further utilized for Deep Embedded Clustering (DEC) and battery health state identification in the following chapters.

Chapter 5

Methodology

5.1 Overview of Proposed Framework

The proposed framework addresses the fundamental limitations of existing battery State-of-Health (SOH) estimation approaches by integrating physics-informed feature engineering, deep representation learning via autoencoders, and joint clustering optimization through Deep Embedded Clustering (DEC). The methodology is motivated by three critical observations from the literature review. First, conventional regression-based approaches treat feature engineering and model training as independent stages, resulting in suboptimal alignment between the feature space and the prediction objective [45]. Second, traditional clustering methods such as K-Means operate directly in the original feature space and assume linear separability, which is inconsistent with the nonlinear degradation trajectories observed in lithium-ion batteries [48]. Third, while deep learning models have demonstrated strong SOH prediction accuracy, most existing works perform clustering as a post-processing step independently of representation learning, limiting the discriminative quality of the resulting latent space [51].

The proposed pipeline transforms raw cycling measurements from the a large-scale lithium-ion battery cycling dataset lithium-ion cell into a compact, structured latent representation that is simultaneously optimized for both faithful reconstruction and cluster separability. The complete workflow proceeds through five sequential stages as follows. Raw time-series battery measurements are first preprocessed and standardized to ensure numerical stability and comparability across cycles. A domain-informed feature engineering step then extracts seven physically meaningful degradation indicators per cycle, forming a structured feature matrix $\mathbf{X} \in \mathbb{R}^{101\times 7}$. A deep autoencoder subsequently compresses this matrix into a lower-dimensional latent representation $\mathbf{Z} \in \mathbb{R}^{101\times k}$, $k < 7$. K-Means clustering is then applied to the latent space as a baseline method, and its results are used to initialize the cluster centroids for Deep Embedded Clustering. Finally, DEC jointly refines the latent space and cluster assignments through iterative KL-divergence minimization, yielding interpretable battery health state groupings validated by quantitative clustering metrics.

The overall pipeline can be compactly expressed as:

$$\mathbf{X} \xrightarrow{f_\theta} \mathbf{Z} \xrightarrow{\text{DEC}} \{C_1, C_2, \ldots, C_K\} \tag{5.1}$$

where f_θ is the encoder and $\{C_j\}$ are the final cluster assignments corresponding to distinct battery health states.

5.2 Data Preprocessing and Standardization

5.2.1 Motivation

Raw battery cycling data contains several sources of variability that can obscure true degradation trends and impair machine learning model performance. These include sensor noise, measurement outliers arising from instrumentation errors, and inconsistencies in temporal sampling rates between the two experimental files. Furthermore, the raw features span widely differing numerical scales: voltage ranges from 3.0 V to 4.2 V, while energy ranges from 0 to 8.111 Wh and cycle index ranges from 1 to 101. Without preprocessing, high-magnitude features dominate the autoencoder reconstruction loss, causing the model to focus disproportionately on large-scale features and underrepresent subtle degradation signals in lower-magnitude variables such as voltage slope.

5.2.2 Dataset Formulation

Let the full dataset be represented as:

$$\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}, \quad \mathbf{x}_n \in \mathbb{R}^d, \quad y_n \in \mathbb{R} \tag{5.2}$$

where $\mathbf{x}_n$ is the feature vector for cycle n, y_n is the corresponding SOH value, $N = 101$ is the total number of cycles, and $d = 7$ is the number of engineered features. Since SOH labels are derived from the data itself (unsupervised setting), y_n is used only for cluster validation and is not used during autoencoder training or DEC optimization.

5.2.3 Sorting and Cycle Segmentation

The combined dataset of 329,164 records is first sorted chronologically by `Cycle_Index` and `Test_Time(s)` to ensure correct temporal ordering within each cycle. The data is then segmented into individual charge and discharge half-cycles using the sign of the current measurement: positive current ($I > 0$) identifies the charging phase, and negative current ($I < 0$) identifies the discharging phase. The maximum discharge capacity per cycle is extracted as the primary capacity measurement:

$$C_n = \max_{t \in \mathcal{T}_n^-} C(t) \tag{5.3}$$

where $\mathcal{T}_n^-$ denotes the set of time indices belonging to the discharge phase of cycle n.

5.2.4 Outlier Removal

Measurements falling outside physically valid operational bounds are removed. Specifically, voltage measurements outside the range $[2.5\text{ V},\ 4.3\text{ V}]$ and current measurements outside

[−8 A, +8 A] are treated as instrumentation artifacts and discarded. These bounds are set with a 5% margin beyond the manufacturer-specified cutoff values to accommodate transient effects without retaining corrupted data.

5.2.5 Min-Max Normalisation

Min-max normalisation is applied to each feature independently to scale all values to the range $[0, 1]$:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{5.4}$$

where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the feature across all cycles. This ensures that features with naturally large ranges, such as energy (0 to 8.111 Wh), do not dominate those with smaller ranges, such as voltage slope.

5.2.6 Zero-Mean Unit-Variance Standardisation

Prior to autoencoder training, zero-mean unit-variance standardisation is additionally applied to ensure numerical stability during gradient-based optimization:

$$x_{\text{std}} = \frac{x - \mu_f}{\sigma_f} \tag{5.5}$$

where μ_f and σ_f are the mean and standard deviation of feature f computed across all $N = 101$ cycle-level observations. Following standardisation, each feature has zero mean and unit variance, which accelerates convergence of the Adam optimizer and prevents vanishing gradient problems in the deeper layers of the autoencoder [48].

5.3 Feature Engineering

5.3.1 Overview

Raw per-second measurements provide a dense time-series representation of each cycle but are not directly suitable as machine learning inputs due to their variable length (2,127 to 5,181 points per cycle) and high dimensionality. Feature engineering compresses each cycle into a fixed-length vector of seven degradation-sensitive scalar descriptors, enabling cycle-level health state analysis. The features are grounded in electrochemical degradation physics, ensuring that each descriptor captures a distinct and physically interpretable aspect of battery aging [45].

5.3.2 Capacity Fade

Capacity fade is the most direct quantitative measure of battery degradation. It is defined as the absolute reduction in maximum discharge capacity from the initial value:

$$\text{Fade}_n = C_0 - C_n \tag{5.6}$$

where $C_0 = 2.271$ Ah is the initial rated capacity observed at Cycle 1 and C_n is the maximum discharge capacity at cycle n. Capacity fade arises primarily from two electrochemical mechanisms: loss of lithium inventory (LLI) caused by irreversible SEI layer growth, and loss of active material (LAM) resulting from structural degradation of electrode particles [48]. For the a large-scale lithium-ion battery cycling dataset dataset, the total capacity fade over 101 cycles is $2.271 - 0.451 = 1.820$ Ah, representing an 80.14% reduction from the initial value. This monotonically increasing function serves as the primary degradation indicator in the feature vector.

5.3.3 State-of-Health (SOH)

SOH provides a normalised, dimensionless measure of battery health relative to its initial condition:

$$\text{SOH}_n = \frac{C_n}{C_0} \times 100\% \tag{5.7}$$

SOH is bounded within the interval $(0, 100]$, where SOH = 100% indicates a fully healthy cell and SOH = 80% defines the conventional end-of-life (EOL) threshold, below which the battery is deemed no longer suitable for primary applications [49]. SOH serves both as an engineered feature and as the primary validation target for cluster interpretation. Including SOH directly in the feature vector allows the autoencoder to learn latent representations that are explicitly structured around the health axis, improving the physical interpretability of the resulting clusters.

5.3.4 Rate of Degradation (ΔSOH)

The rate of degradation captures the local dynamics of health decline between consecutive cycles:

$$\Delta\text{SOH}_n = \text{SOH}_n - \text{SOH}_{n-1} \tag{5.8}$$

Under normal aging conditions, ΔSOH_n is a small negative quantity reflecting gradual, steady-state capacity decline. A sudden large negative spike in ΔSOH_n at a particular cycle indicates an accelerated degradation event, such as lithium plating triggered by an abrupt

increase in C-rate or temperature excursion [50]. Conversely, small positive values may indicate short-term capacity recovery effects, which are a known phenomenon in lithium-ion cells under rest periods between cycling sessions. This feature therefore provides information about degradation dynamics that is complementary to the absolute SOH value and is particularly valuable for identifying anomalous aging behaviour.

5.3.5 Voltage Slope

The voltage slope captures the average rate of terminal voltage change during a cycle, reflecting the electrochemical kinetics of the cell:

$$\text{Slope}_n = \frac{1}{M_n} \sum_{t=2}^{M_n} \frac{V_t - V_{t-1}}{\Delta t_t} \tag{5.9}$$

where M_n is the number of time-series measurements in cycle n, V_t is the terminal voltage at time step t, and $\Delta t_t = t - (t-1)$ is the sampling interval in seconds. As the battery ages, increasing internal resistance R_{int} causes a larger ohmic voltage drop under load, expressed as $\Delta V = I \cdot R_{\text{int}}$. This manifests as a steeper negative voltage slope during discharge and a shallower positive slope during charging. Consequently, the voltage slope is a sensitive indirect indicator of internal resistance rise, which is one of the primary manifestations of SEI layer growth and electrolyte decomposition [48]. Studies have demonstrated that voltage-based features are among the most predictive indicators for SOH estimation, as the full shape of the voltage curve encodes rich electrochemical information about the state of the electrode materials [45].

5.3.6 Energy Efficiency

The energy efficiency feature quantifies the ratio of energy delivered per unit of capacity in each cycle:

$$\eta_n = \frac{C_n}{E_n} \tag{5.10}$$

where E_n is the total energy (Wh) delivered or absorbed during cycle n. As the battery degrades, resistive losses within the cell increase, causing a greater fraction of the input energy to be dissipated as heat rather than stored as chemical potential energy. This results in a decline in η_n even in early cycles where the capacity fade is still relatively small. Energy efficiency therefore provides complementary degradation information that is sensitive to resistance-driven losses, which may precede observable capacity decline [45]. The mathematical relationship between efficiency and internal resistance is given by:

$$\eta_n \approx 1 - \frac{I^2 R_{\text{int},n} \cdot t_n}{E_{\text{input},n}} \tag{5.11}$$

where t_n is the cycle duration and $E_{\text{input},n}$ is the total energy supplied during charging.

5.3.7 Mean Cycle Temperature

The mean cycle temperature is computed as the arithmetic mean of all temperature readings recorded during cycle n:

$$T_n = \frac{1}{k_n} \sum_{i=1}^{k_n} T_i \tag{5.12}$$

where k_n is the number of temperature measurements in cycle n. Temperature is one of the most significant external drivers of battery degradation. Elevated temperatures ($T > 40°\text{C}$) accelerate SEI layer formation through increased reaction rates of solvent decomposition at the anode surface, while low temperatures ($T < 0°\text{C}$) promote lithium plating by reducing lithium-ion mobility in the electrolyte [49]. The Arrhenius equation describes the temperature dependence of these degradation reaction rates:

$$k(T) = A \exp\left(-\frac{E_a}{R_g T}\right) \tag{5.13}$$

where A is the pre-exponential frequency factor, E_a is the activation energy of the degradation reaction (J/mol), $R_g = 8.314$ J/(mol·K) is the universal gas constant, and T is the absolute temperature in Kelvin. This implies a non-linear coupling between temperature and degradation rate:

$$\frac{\mathrm{d}C}{\mathrm{d}n} \propto \exp\left(-\frac{E_a}{R_g T_n}\right) \tag{5.14}$$

By including mean cycle temperature as a feature, the machine learning model implicitly accounts for this thermally-driven non-linearity without requiring explicit physics-based parameterisation.

5.3.8 C-Rate

The C-rate quantifies the cycling intensity relative to the cell's nominal capacity:

$$\text{C-rate}_n = \frac{\bar{I}_n}{Q_{\text{nominal}}} \tag{5.15}$$

where $\bar{I}_n$ is the mean absolute current during cycle n and $Q_{\text{nominal}} = 2.271$ Ah. High

C-rates accelerate SEI growth and induce mechanical stress in electrode particles due to rapid lithium intercalation and de-intercalation, contributing to capacity fade and internal resistance rise [48]. The preliminary polynomial regression analysis demonstrated that capacity degradation trajectories differ significantly across C-rate regimes, motivating its inclusion as an explicit feature to enable the model to account for cycling intensity effects on degradation.

5.3.9 Assembled Feature Matrix

The seven scalar features are assembled into a cycle-level feature vector:

$$\mathbf{x}_n = \left[C_n,\ \text{Fade}_n,\ \text{SOH}_n,\ \Delta\text{SOH}_n,\ \text{Slope}_n,\ \eta_n,\ T_n\right]^\top \in \mathbb{R}^7 \tag{5.16}$$

Stacking across all $N = 101$ cycles yields the feature matrix:

$$\mathbf{X} = [\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_N]^\top \in \mathbb{R}^{101\times 7} \tag{5.17}$$

This fixed-size matrix forms the sole input to the autoencoder, replacing the variable-length raw time-series data with a compact, physically grounded representation.

5.4 Dimensionality Reduction using Autoencoder

5.4.1 Motivation

Although the feature matrix $\mathbf{X}$ has already reduced the data dimensionality from 329,164 time-series measurements to a 101×7 matrix, the seven features are not independent. The correlation analysis in Chapter 3 revealed that capacity and energy exhibit near-perfect correlation ($r = 0.99$), and capacity fade and SOH are mathematically linked by definition ($\text{SOH} = 1 - \text{Fade}/C_0$). These redundancies reduce the effective dimensionality of the feature space and can cause clustering algorithms to form spurious groupings along correlated dimensions rather than along the true degradation axis.

Furthermore, the degradation manifold of lithium-ion batteries is intrinsically nonlinear, as demonstrated by the curved capacity–cycle trajectory in the pairplot analysis. Linear dimensionality reduction methods such as Principal Component Analysis (PCA) seek an orthogonal projection that maximises explained variance, but they cannot capture the non-linear structure of the degradation manifold. An autoencoder, by contrast, employs nonlinear activation functions that allow it to learn curved, low-dimensional manifold embeddings, producing a latent space that is better suited for identifying distinct degradation states [52].

5.4.2 Autoencoder Architecture

The autoencoder is a neural network comprising a symmetric encoder–decoder pair. The encoder $f_\theta : \mathbb{R}^7 \to \mathbb{R}^k$ maps each feature vector to a lower-dimensional latent code, and the decoder $g_\phi : \mathbb{R}^k \to \mathbb{R}^7$ reconstructs the original input from the latent code.

The encoder applies two successive affine transformations followed by nonlinear activations:

$$\mathbf{h}_n = \sigma(\mathbf{W}_1 \mathbf{x}_n + \mathbf{b}_1) \tag{5.18}$$

$$\mathbf{z}_n = \sigma(\mathbf{W}_2 \mathbf{h}_n + \mathbf{b}_2) \tag{5.19}$$

where $\mathbf{W}_1 \in \mathbb{R}^{h \times 7}$, $\mathbf{W}_2 \in \mathbb{R}^{k \times h}$ are learnable weight matrices, $\mathbf{b}_1, \mathbf{b}_2$ are bias vectors, h is the hidden layer dimension, and $\sigma(\cdot)$ is the Rectified Linear Unit (ReLU) activation function:

$$\sigma(x) = \max(0,\ x) \tag{5.20}$$

ReLU is chosen for its computational efficiency and its property of inducing sparsity in the latent representation, which improves the separability of degradation states [52]. The decoder performs the symmetric inverse mapping:

$$\mathbf{h}'_n = \sigma(\mathbf{W}_3 \mathbf{z}_n + \mathbf{b}_3) \tag{5.21}$$

$$\hat{\mathbf{x}}_n = \mathbf{W}_4 \mathbf{h}'_n + \mathbf{b}_4 \tag{5.22}$$

where no activation is applied at the output layer to allow reconstruction of both positive and negative standardised feature values.

5.4.3 Training Objective

The autoencoder is trained by minimising the mean squared reconstruction error over all cycles:

$$\mathcal{L}_{\text{AE}} = \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|_2^2 = \frac{1}{N} \sum_{n=1}^{N} \left\|\mathbf{x}_n - g_\phi(f_\theta(\mathbf{x}_n))\right\|_2^2 \tag{5.23}$$

Minimising $\mathcal{L}_{\text{AE}}$ forces the bottleneck latent layer to encode the maximum amount of information recoverable from the compressed representation. Features that are redundant or represent measurement noise cannot be reconstructed from the bottleneck and are therefore

filtered out during training. The progressive decrease in $\mathcal{L}_{\text{AE}}$ over training epochs confirms that the model is learning a meaningful compressed representation of the degradation manifold [52].

5.4.4 Motivation over PCA

PCA finds a linear projection $\mathbf{Z} = \mathbf{X}\mathbf{W}$ that maximises the total explained variance, where $\mathbf{W} \in \mathbb{R}^{7\times k}$ contains the top k principal eigenvectors of the covariance matrix $\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$. While PCA is computationally efficient and fully interpretable, it is restricted to linear manifolds by construction. The nonlinear capacity–cycle degradation trajectory observed in the a large-scale lithium-ion battery cycling dataset dataset (Fig. **??**) lies on a curved manifold that cannot be unfolded by a linear projection without significant information loss. An autoencoder with ReLU activations can approximate arbitrary nonlinear manifolds through universal approximation, enabling it to produce a latent space that more faithfully represents the intrinsic geometry of battery degradation [45].

5.4.5 Implementation Details

The autoencoder is implemented using TensorFlow/Keras with the following architectural configuration:

- **Input layer:** 7 neurons (one per engineered feature)
- **Encoder hidden layer:** $h = 32$ neurons, ReLU activation
- **Bottleneck layer:** k neurons (latent dimension, tuned via elbow method on reconstruction loss)
- **Decoder hidden layer:** 32 neurons, ReLU activation
- **Output layer:** 7 neurons, linear activation
- **Optimizer:** Adam with learning rate $\alpha = 0.001$ and default momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$
- **Loss function:** Mean Squared Error (MSE)
- **Training:** Mini-batch gradient descent over 200 epochs

5.5 Latent Space Representation

5.5.1 Definition

Following autoencoder training, the trained encoder f_θ is applied to the full standardised feature matrix to produce the latent matrix:

$$\mathbf{Z} = f_\theta(\mathbf{X}) = [\mathbf{z}_1,\ \mathbf{z}_2,\ \ldots,\ \mathbf{z}_N]^\top \in \mathbb{R}^{N\times k} \tag{5.24}$$

Each row $\mathbf{z}_n \in \mathbb{R}^k$ is the latent code of cycle n, representing the compressed encoding of its seven-dimensional feature vector. The latent space dimension k is selected to balance reconstruction fidelity against dimensionality reduction, and is determined empirically by plotting the reconstruction loss $\mathcal{L}_{\text{AE}}$ as a function of k and identifying the elbow point where further increases in k yield diminishing reductions in loss.

5.5.2 Properties of the Latent Space

The learned latent space $\mathbf{Z}$ exhibits three key properties that make it superior to the original feature space for clustering:

(i) Dimensionality reduction and decorrelation. The autoencoder training implicitly decorrelates the latent dimensions by learning a basis in which the degradation manifold can be efficiently represented. While the original feature space contains highly correlated dimensions (e.g., capacity and SOH with $r = 0.99$), the bottleneck forces the encoder to find independent directions of maximum variance in the degradation manifold, analogous to independent component analysis but in a nonlinear setting [52].

(ii) Noise filtering. The reconstruction objective forces the latent code to capture only information that is recoverable at the output. Measurement noise, which by definition cannot be reconstructed from a compact code, is suppressed. This is mathematically equivalent to a low-pass filter applied to the feature space along the dimensions of minimum variance, and is analogous to the denoising effect of denoising autoencoders (DAE) studied in the context of battery SOH estimation by Han et al. [53].

(iii) Manifold structure preservation. Cycles with similar electrochemical degradation behavior are mapped to geometrically proximate points in $\mathbb{R}^k$. This is a consequence of the continuity of the encoder mapping f_θ: if two feature vectors $\mathbf{x}_n$ and $\mathbf{x}_m$ are close in $\mathbb{R}^7$, their latent codes $\mathbf{z}_n$ and $\mathbf{z}_m$ are also close in $\mathbb{R}^k$. The converse, that well-separated health states in the latent space correspond to genuinely distinct degradation regimes, is further enforced by the DEC training phase.

5.5.3 Geometric Analysis

Before applying clustering, the geometric properties of $\mathbf{Z}$ are assessed by projecting it onto a two-dimensional plane using PCA for visualization (Section 5.10). The inter-point distances in the latent space are analysed using the pairwise Euclidean distance matrix:

$$D_{nm} = \|\mathbf{z}_n - \mathbf{z}_m\|_2\,, \quad n, m = 1, \ldots, N \tag{5.25}$$

The distribution of D_{nm} values provides evidence of natural cluster structure: a multimodal distribution of pairwise distances suggests the existence of well-separated groups in the latent space, supporting the validity of the subsequent clustering analysis.

5.6 Clustering using K-Means

5.6.1 Overview and Motivation

K-Means clustering is applied directly to the latent matrix $\mathbf{Z}$ as a baseline method for battery health state identification. It serves two purposes in the proposed framework: first, as a standalone clustering method whose results are compared against DEC to quantify the improvement achieved by joint optimization; and second, as the initialization method for DEC cluster centroids, following the standard DEC protocol established by Xie et al.

5.6.2 K-Means Objective

K-Means minimises the total within-cluster sum of squared Euclidean distances between each latent point and its assigned cluster centroid:

$$\mathcal{L}_{\text{KM}} = \sum_{j=1}^{K} \sum_{\mathbf{z}_n \in C_j} \left\| \mathbf{z}_n - \boldsymbol{\mu}_j \right\|_2^2 \tag{5.26}$$

where K is the number of clusters, $C_j = \{n : \text{assign}(n) = j\}$ is the set of cycles assigned to cluster j, and $\boldsymbol{\mu}_j \in \mathbb{R}^k$ is the centroid of cluster j:

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{n \in C_j} \mathbf{z}_n \tag{5.27}$$

The algorithm alternates between two steps until convergence: the assignment step, which assigns each point to its nearest centroid:

$$\text{assign}(n) = \arg \min_{j \in \{1,\ldots,K\}} \left\| \mathbf{z}_n - \boldsymbol{\mu}_j \right\|_2^2 \tag{5.28}$$

and the update step, which recomputes each centroid as the mean of its assigned points (Equation 5.27). Convergence is declared when no point changes its cluster assignment between consecutive iterations.

5.6.3 Optimal K Selection

The optimal number of clusters K is determined using the elbow method, which plots the within-cluster inertia $\mathcal{L}_{\text{KM}}(K)$ as a function of K. The optimal K is identified at the elbow point where the rate of decrease in inertia changes from steep to gradual, indicating

diminishing returns from adding further clusters. This is validated against the Silhouette Score, which provides an additional metric for cluster quality independent of the inertia-based criterion.

5.6.4 Limitations

K-Means assumes that clusters are convex, isotropic, and of approximately equal size, since it uses Euclidean distance to a single centroid as the sole assignment criterion. These assumptions may not hold for battery degradation data, where different degradation stages can span different regions of varying density and shape in the latent space. Furthermore, K-Means does not refine the latent space itself during clustering — it operates on a fixed representation, meaning the quality of its output is entirely determined by the quality of the autoencoder's pre-trained latent space. These limitations motivate the adoption of Deep Embedded Clustering as the primary method.

5.7 Deep Embedded Clustering (DEC)

5.7.1 Overview

Deep Embedded Clustering (DEC) was introduced to address the fundamental limitation of performing clustering on a fixed representation: the latent space learned by an autoencoder is optimized for reconstruction fidelity, not for cluster separability. DEC jointly refines the latent space and cluster assignments by back-propagating a clustering loss through the encoder, progressively restructuring the latent space to enhance the discriminative separation of degradation states [51]. In the context of battery health analysis, DEC has been shown to produce more coherent and physically interpretable cluster structures compared to methods that perform clustering independently of representation learning [52].

5.7.2 Soft Cluster Assignment

Given K cluster centroids $\{\boldsymbol{\mu}_j\}_{j=1}^{K}$ initialised from K-Means, the soft assignment probability of latent point $\mathbf{z}_n$ to cluster j is computed using the Student's t-distribution kernel:

$$q_{nj} = \frac{\left(1 + \left\|\mathbf{z}_n - \boldsymbol{\mu}_j\right\|_2^2 / \nu\right)^{-(\nu+1)/2}}{\sum_{j'=1}^{K} \left(1 + \left\|\mathbf{z}_n - \boldsymbol{\mu}_{j'}\right\|_2^2 / \nu\right)^{-(\nu+1)/2}} \tag{5.29}$$

where ν is the degrees of freedom of the t-distribution (set to $\nu = 1$ following standard DEC practice). The use of the t-distribution rather than a Gaussian kernel is motivated by its heavier tails, which reduce the penalty for points far from their assigned centroid and allow the algorithm to form more natural cluster boundaries in the presence of outliers.

5.7.3 Target Distribution

To drive the cluster assignments toward high-confidence predictions, a sharpened target distribution P is computed from the soft assignments Q:

$$p_{nj} = \frac{q_{nj}^2 / f_j}{\sum_{j'=1}^{K} q_{nj'}^2 / f_{j'}} \tag{5.30}$$

where $f_j = \sum_{n=1}^{N} q_{nj}$ is the soft cluster frequency, representing the effective number of points assigned to cluster j. The squaring of q_{nj} in the numerator amplifies high-confidence assignments (where q_{nj} is already close to 1) while suppressing ambiguous assignments (where q_{nj} is close to $1/K$). The division by f_j normalises for cluster size, preventing large clusters from dominating the loss. The target distribution P therefore represents a sharper, more confident version of Q, and minimising the divergence between P and Q drives the model to increase its assignment confidence.

5.7.4 DEC Objective Function

The DEC training objective minimises the Kullback–Leibler (KL) divergence between the target distribution P and the soft assignment distribution Q:

$$\mathcal{L}_{\text{DEC}} = \text{KL}(P \| Q) = \sum_{n=1}^{N} \sum_{j=1}^{K} p_{nj} \log \frac{p_{nj}}{q_{nj}} \tag{5.31}$$

The KL divergence is asymmetric and equals zero only when $P = Q$. By minimising $\mathcal{L}_{\text{DEC}}$, the model is trained to make Q (the actual soft assignments) approach P (the sharpened target), progressively concentrating each q_{nj} distribution toward a one-hot vector corresponding to a definitive cluster assignment. Gradients of $\mathcal{L}_{\text{DEC}}$ with respect to the latent codes $\{\mathbf{z}_n\}$ are back-propagated through the encoder f_θ, restructuring the latent space to increase inter-cluster distances and reduce intra-cluster scatter.

5.7.5 Cluster Centroid Update

The cluster centroids are updated at each iteration as the weighted mean of all latent points, with weights given by the target distribution:

$$\boldsymbol{\mu}_j = \frac{\sum_{n=1}^{N} p_{nj} \, \mathbf{z}_n}{\sum_{n=1}^{N} p_{nj}} \tag{5.32}$$

This soft centroid update is more robust to outliers than the hard assignment update of K-Means, as all points contribute to the centroid with weights proportional to their assignment

confidence.

5.7.6 Convergence Criterion

DEC training is terminated when the fraction of points changing their hard cluster assignment between consecutive target distribution updates falls below a threshold δ:

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\left[\arg\max_j q_{nj}^{(t)} \neq \arg\max_j q_{nj}^{(t-1)}\right] < \delta \tag{5.33}$$

where $\delta = 0.001$ is used in this study, following standard practice. Convergence typically occurs within 10–30 iterations over the 101-cycle dataset. Listing 5.1 presents the complete DEC soft assignment and iterative training implementation.

Listing 5.1: –[29]).]Deep Embedded Clustering implementation: soft assignment using Student's t-distribution (Eq. 5.29), target distribution sharpening (Eq. 5.30), and iterative centroid refinement (Jupyter Cells [25]–[29]).

```
import numpy as np

# Initialise centroids from K-Means
cluster_centers = kmeans.cluster_centers_

# Soft assignment: Student t-distribution kernel
def soft_assignment(Z, centers):
    dist = np.sum((Z[:, np.newaxis] - centers)**2, axis=2)
    q    = 1.0 / (1.0 + dist)
    q    = q / q.sum(axis=1, keepdims=True)
    return q

# Target distribution (sharpened)
def target_distribution(q):
    weight = q**2 / q.sum(axis=0)
    return (weight.T / weight.sum(axis=1)).T

# DEC iterative refinement (50 iterations)
for i in range(50):
    q = soft_assignment(Z, cluster_centers)
    p = target_distribution(q)
    # Update cluster centres via soft weighted mean
    cluster_centers = (p.T @ Z) / p.sum(axis=0)[:, np.newaxis]

# Final hard cluster assignments
final_clusters = np.argmax(q, axis=1)
```

5.8 Joint Optimization of Representation and Clustering

5.8.1 Combined Loss Function

The most significant methodological contribution of the proposed framework is the joint optimization of the autoencoder representation and the DEC clustering objective within a single unified training procedure. The combined loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{AE}} + \lambda \, \mathcal{L}_{\text{DEC}} \tag{5.34}$$

where $\lambda > 0$ is a scalar hyperparameter controlling the relative weight of the clustering objective with respect to the reconstruction objective. Setting $\lambda = 0$ recovers the standard autoencoder without clustering, while large λ causes the model to prioritise cluster separability at the expense of reconstruction fidelity.

5.8.2 Gradient Flow Analysis

The gradient of $\mathcal{L}_{\text{total}}$ with respect to the encoder parameters θ decomposes as:

$$\frac{\partial \mathcal{L}_{\text{total}}}{\partial \theta} = \frac{\partial \mathcal{L}_{\text{AE}}}{\partial \theta} + \lambda \, \frac{\partial \mathcal{L}_{\text{DEC}}}{\partial \theta} \tag{5.35}$$

The reconstruction gradient $\partial \mathcal{L}_{\text{AE}} / \partial \theta$ pushes the encoder to preserve maximal information in the latent code, while the clustering gradient $\partial \mathcal{L}_{\text{DEC}} / \partial \theta$ pushes the encoder to increase the separation between clusters in the latent space. The competition between these two gradients produces a latent space that balances information preservation and discriminative structure, yielding clusters that are both physically meaningful and well-separated [51].

5.8.3 Training Procedure

The complete joint training procedure is as follows:

1. Pre-train the autoencoder by minimising $\mathcal{L}_{\text{AE}}$ alone for 200 epochs to obtain a good initial latent representation.
2. Apply K-Means to $\mathbf{Z}$ with K clusters to obtain initial centroids $\{\boldsymbol{\mu}_j^{(0)}\}$.
3. Iterate until convergence:
   (a) Compute soft assignments Q using Equation 5.29.
   (b) Compute target distribution P using Equation 5.30.
   (c) Compute $\mathcal{L}_{\text{total}}$ and back-propagate gradients through both θ and ϕ.
   (d) Update centroids using Equation 5.32.

(e) Check convergence criterion (Equation 5.33).

4. Extract final hard cluster assignments as $\hat{c}_n = \arg\max_j q_{nj}$.

5.9 Evaluation Metrics

5.9.1 Silhouette Score

The Silhouette Score measures the quality of cluster assignments by comparing intra-cluster cohesion to inter-cluster separation for each point:

$$s_n = \frac{b_n - a_n}{\max(a_n,\, b_n)} \tag{5.36}$$

where a_n is the mean Euclidean distance from point n to all other points in its assigned cluster (intra-cluster distance), and b_n is the mean distance from point n to all points in the nearest neighbouring cluster (inter-cluster distance):

$$a_n = \frac{1}{|C_{c_n}| - 1} \sum_{m \in C_{c_n},\, m \neq n} \|\mathbf{z}_n - \mathbf{z}_m\|_2 \tag{5.37}$$

$$b_n = \min_{j \neq c_n} \frac{1}{|C_j|} \sum_{m \in C_j} \|\mathbf{z}_n - \mathbf{z}_m\|_2 \tag{5.38}$$

The overall Silhouette Score is the mean over all points:

$$S = \frac{1}{N} \sum_{n=1}^{N} s_n \in [-1,\, 1] \tag{5.39}$$

Values approaching +1 indicate well-separated, compact clusters; values near 0 indicate overlapping clusters; and negative values indicate misassignment.

5.9.2 Davies–Bouldin Index

The Davies–Bouldin Index (DBI) measures the average similarity between each cluster and its most similar neighbouring cluster, where similarity is defined as the ratio of within-cluster scatter to between-cluster distance:

$$\mathrm{DBI} = \frac{1}{K} \sum_{j=1}^{K} \max_{j' \neq j} \left(\frac{\sigma_j + \sigma_{j'}}{d(\boldsymbol{\mu}_j,\, \boldsymbol{\mu}_{j'})} \right) \tag{5.40}$$

where $\sigma_j = \frac{1}{|C_j|} \sum_{n \in C_j} \|\mathbf{z}_n - \boldsymbol{\mu}_j\|_2$ is the average intra-cluster distance and $d(\boldsymbol{\mu}_j, \boldsymbol{\mu}_{j'}) = \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_{j'}\|_2$ is the Euclidean distance between centroids. Lower DBI values indicate better-

defined clusters with smaller within-cluster scatter relative to between-cluster separation.

5.9.3 Calinski–Harabasz Score

The Calinski–Harabasz Score (CHS) measures the ratio of between-cluster dispersion to within-cluster dispersion:

$$\text{CHS} = \frac{\text{tr}(\mathbf{B}_K)/(K-1)}{\text{tr}(\mathbf{W}_K)/(N-K)} \tag{5.41}$$

where the between-cluster scatter matrix is:

$$\mathbf{B}_K = \sum_{j=1}^{K} |C_j|(\boldsymbol{\mu}_j - \bar{\mathbf{z}})(\boldsymbol{\mu}_j - \bar{\mathbf{z}})^\top \tag{5.42}$$

and the within-cluster scatter matrix is:

$$\mathbf{W}_K = \sum_{j=1}^{K} \sum_{n \in C_j} (\mathbf{z}_n - \boldsymbol{\mu}_j)(\mathbf{z}_n - \boldsymbol{\mu}_j)^\top \tag{5.43}$$

with $\bar{\mathbf{z}} = \frac{1}{N}\sum_n \mathbf{z}_n$ being the global centroid. Higher CHS values indicate more compact, well-separated clusters.

5.9.4 Metric Interpretation for DEC vs K-Means

The three metrics provide complementary perspectives on clustering quality. A higher Silhouette Score for K-Means compared to DEC does not necessarily indicate superior clustering, as K-Means explicitly minimises the objective that the Silhouette Score measures (within-cluster distances). DEC, by contrast, optimises for a semantically richer objective that restructures the latent space to align with degradation patterns. It is therefore expected that DEC may produce clusters with slightly lower geometric compactness but stronger alignment with physically meaningful SOH boundaries, as validated by the SOH distribution analysis in Section 5.13.

5.10 PCA-Based Visualization of Clusters

5.10.1 Motivation

The latent space $\mathbf{Z} \in \mathbb{R}^{N \times k}$ cannot be directly visualized when $k > 2$. PCA is applied as a post-hoc visualization tool to project the latent representations onto a two-dimensional plane, enabling qualitative assessment of cluster structure and boundary definition.

5.10.2 PCA Projection

The PCA transformation is computed by finding the top two eigenvectors of the empirical covariance matrix of $\mathbf{Z}$:

$$\mathbf{\Sigma}_Z = \frac{1}{N}\mathbf{Z}^\top\mathbf{Z} - \bar{\mathbf{z}}\,\bar{\mathbf{z}}^\top \in \mathbb{R}^{k\times k} \tag{5.44}$$

Let $\mathbf{w}_1$ and $\mathbf{w}_2$ be the eigenvectors corresponding to the two largest eigenvalues $\lambda_1 \geq \lambda_2$ of $\mathbf{\Sigma}_Z$. The two-dimensional projection is:

$$\mathbf{Z}_{2D} = \mathbf{Z}[\mathbf{w}_1\ \mathbf{w}_2] \in \mathbb{R}^{N\times 2} \tag{5.45}$$

The explained variance ratio of the projection is:

$$\text{EVR} = \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{k}\lambda_i} \tag{5.46}$$

A high EVR (e.g., $> 80\%$) indicates that the two principal components capture most of the variance in the latent space, and that the 2D visualization faithfully represents the cluster structure. Both K-Means and DEC cluster assignments are overlaid on the same $\mathbf{Z}_{2D}$ scatter plot to enable direct visual comparison of cluster boundaries.

5.11 Cluster Comparison Analysis

5.11.1 Cross-Tabulation

A systematic comparison between K-Means and DEC cluster assignments is performed by constructing a cross-tabulation matrix $\mathbf{C} \in \mathbb{Z}^{K\times K}$, where entry C_{ij} counts the number of cycles assigned to cluster i by K-Means and cluster j by DEC. Since cluster labels are arbitrary permutations, the Hungarian algorithm is applied to find the optimal label correspondence that maximises the trace of $\mathbf{C}$ before computing agreement statistics.

5.11.2 Adjusted Rand Index

The Adjusted Rand Index (ARI) quantifies the degree of agreement between two clustering solutions, corrected for chance:

$$\text{ARI} = \frac{\displaystyle\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2}\right]\left[\sum_j \binom{b_j}{2}\right]}{\binom{N}{2}}}{\displaystyle\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2}\right]\left[\sum_j \binom{b_j}{2}\right]}{\binom{N}{2}}} \tag{5.47}$$

where $n_{ij} = C_{ij}$, $a_i = \sum_j C_{ij}$ are the row sums, and $b_j = \sum_i C_{ij}$ are the column sums of the cross-tabulation matrix. ARI = 1 indicates perfect agreement; ARI = 0 indicates agreement equivalent to chance assignment.

5.12 Statistical Feature Analysis

5.12.1 Within-Cluster Statistics

For each identified cluster C_j and each engineered feature $f \in \{1, \ldots, 7\}$, the within-cluster mean and standard deviation are computed:

$$\bar{x}_{f,j} = \frac{1}{|C_j|} \sum_{n \in C_j} x_{fn} \tag{5.48}$$

$$\sigma_{f,j} = \sqrt{\frac{1}{|C_j|} \sum_{n \in C_j} (x_{fn} - \bar{x}_{f,j})^2} \tag{5.49}$$

Large differences in $\bar{x}_{f,j}$ across clusters for a given feature f indicate that the feature is discriminative for health state separation.

5.12.2 Correlation with SOH

The Pearson correlation coefficient between feature f and SOH within cluster j is:

$$\rho_{f,j} = \frac{\sum_{n \in C_j} (x_{fn} - \bar{x}_{f,j})(\mathrm{SOH}_n - \overline{\mathrm{SOH}}_j)}{\sqrt{\sum_{n \in C_j} (x_{fn} - \bar{x}_{f,j})^2 \cdot \sum_{n \in C_j} (\mathrm{SOH}_n - \overline{\mathrm{SOH}}_j)^2}} \tag{5.50}$$

Features with $|\rho_{f,j}| > 0.7$ within a cluster are considered strongly associated with the SOH trajectory of that health state, and are identified as the primary degradation indicators for each cluster.

5.13 Visualization of SOH Distribution Across Clusters

5.13.1 Boxplot Analysis

The distribution of SOH values within each cluster is visualized using boxplots. For cluster C_j, the boxplot reports the following five-number summary:

$$\left\{ \min_j,\ Q1_j,\ \mathrm{median}_j,\ Q3_j,\ \max_j \right\} \tag{5.51}$$

where $Q1_j$ and $Q3_j$ are the first and third quartiles of the SOH distribution in cluster j, and

the interquartile range is $\mathrm{IQR}_j = Q3_j - Q1_j$. Outliers are identified as points satisfying:

$$\mathrm{SOH}_n < Q1_j - 1.5 \cdot \mathrm{IQR}_j \quad \text{or} \quad \mathrm{SOH}_n > Q3_j + 1.5 \cdot \mathrm{IQR}_j \tag{5.52}$$

5.13.2 Statistical Significance Testing

To confirm that the identified clusters correspond to statistically distinct health states rather than arbitrary partitions, a one-way Analysis of Variance (ANOVA) is applied to the SOH values across clusters:

$$F = \frac{\mathrm{MS}_{\text{between}}}{\mathrm{MS}_{\text{within}}} = \frac{\sum_{j=1}^{K} |C_j| \, (\bar{y}_j - \bar{y})^2 \big/ (K-1)}{\sum_{j=1}^{K} \sum_{n \in C_j} (y_n - \bar{y}_j)^2 \big/ (N-K)} \tag{5.53}$$

where $y_n = \mathrm{SOH}_n$, $\bar{y}_j$ is the mean SOH in cluster j, and $\bar{y}$ is the global mean SOH. A statistically significant F-statistic ($p < 0.05$) confirms that the between-cluster SOH differences are not attributable to random variation, validating that the proposed framework produces physically meaningful and interpretable battery health state groupings.

Chapter 6

Results and Discussion

6.1 Introduction

This chapter presents the experimental results and analytical observations obtained from the proposed Deep Embedded Clustering (DEC)-based framework for lithium-ion battery State-of-Health (SOH) analysis and degradation characterization. The framework integrates physics-informed feature engineering, nonlinear latent-space learning through autoencoders, unsupervised clustering using K-Means and DEC, and multiple statistical evaluation techniques to identify degradation states and hidden operational patterns within the a large-scale lithium-ion battery cycling dataset battery cycling dataset.

The results are presented in a systematic manner beginning with exploratory data analysis and raw battery characteristic observations, followed by engineered feature behavior, autoencoder training performance, latent-space clustering analysis, DEC refinement results, cluster validation metrics, and final statistical interpretation. The objective of this chapter is not only to present numerical outputs and graphical visualizations but also to interpret the electrochemical significance of each finding and establish how the proposed framework successfully captures nonlinear battery degradation behavior.

6.2 Exploratory Analysis of Raw Battery Characteristics

6.2.1 Correlation Analysis

Prior to feature engineering, a Pearson correlation matrix was computed across the five primary measurement variables to understand the inter-variable relationships governing battery degradation. The Pearson correlation heatmap presented earlier in Figure 3.1 (Chapter 3) confirmed that Capacity and Energy exhibit near-perfect correlation ($r = 0.99$), motivating the use of derived efficiency features rather than raw energy values.

The following significant relationships were identified from the correlation matrix:

- **Capacity and Energy ($r = 0.99$):** A near-perfect positive correlation confirms that energy delivery capability declines almost proportionally with capacity degradation. This high redundancy motivated the use of derived efficiency features rather than raw energy values in the feature engineering pipeline.
- **Current and Voltage ($r = 0.81$):** A strong positive correlation indicates that terminal voltage is significantly influenced by charging and discharging current conditions, consistent with the ohmic voltage drop relationship $\Delta V = I \cdot R_{\text{int}}$.

- **Capacity and Cycle_Index ($r = -0.52$):** A moderate negative correlation confirms progressive battery aging with increasing cycle count, validating that the dataset captures a meaningful degradation trajectory.
- **Voltage and Cycle_Index ($r = -0.02$):** The near-zero raw correlation confirms that voltage degradation manifests in the *shape* of voltage profiles rather than in instantaneous values, justifying the extraction of the voltage slope feature.

These observations validated the necessity of extracting higher-level degradation-sensitive features instead of relying solely on raw measurements for health state identification.

6.2.2 Pairwise Bivariate Analysis

The pairplot visualization of raw battery variables (Fig. **??**) further confirmed several key observations. The Capacity–Cycle_Index scatter revealed a clear downward curving degradation trajectory, confirming the nonlinear nature of aging. The Capacity–Energy scatter showed a near-linear relationship consistent with $r = 0.99$. The Voltage–Cycle_Index scatter exhibited horizontal banding reflecting discrete fixed current levels, confirming that instantaneous voltage alone is insufficient as a degradation descriptor.

The pairplot visualization presented in Figure **??** (Chapter 3) confirmed a clear nonlinear downward degradation trajectory in the Capacity–Cycle_Index scatter, validating the need for nonlinear representation learning.

6.3 Preliminary Capacity Prediction Using Polynomial Regression

As a preliminary modelling step, degree-3 polynomial regression was applied to predict battery capacity over 120 cycles for each unique C-rate present in the dataset. Figures 4.1 and 4.2 show representative fits for C-rate values of -145.67 and -137.34 respectively.

The polynomial regression results presented in Figures 4.1 and 4.2 (Chapter 4) demonstrated the physically unrealistic extrapolation behavior that motivated the transition to the autoencoder–DEC framework.

Both graphs clearly demonstrate the overfitting and physically unrealistic extrapolation behavior of polynomial regression at extreme C-rate values, with the red polynomial curve diverging sharply from actual data points beyond the observed cycle range. These limitations confirm that a more generalizable, data-driven framework is required for robust SOH estimation.

6.3.1 SOH Trend from Raw Capacity

Figure 4.3 presents the raw SOH trajectory computed directly from per-cycle maximum capacity values using Equation 4.3.

The raw SoH trajectory presented in Figure 4.3 (Chapter 4) revealed anomalous values exceeding 35,000% in early cycles and a sharp discontinuity between cycles 30–60, directly motivating the preprocessing and standardization steps of the proposed framework.

The raw SOH plot revealed two critical issues. First, anomalous SOH values exceeding 35,000% appear in early cycles due to the use of instantaneous cumulative capacity measurements rather than normalized per-cycle discharge totals. Second, a sharp discontinuity appears between cycles 30–60 corresponding to the boundary between the two concatenated experimental files. These observations directly motivated the preprocessing and standardization steps described in Chapter 5.

6.4 Engineered Feature Analysis

Following the application of the feature engineering pipeline described in Chapter 4, seven degradation-sensitive cycle-level features were extracted: Capacity, Capacity Fade, SOH, ΔSOH, Voltage Slope, Energy Efficiency, and Temperature.

The engineered features demonstrated clear degradation progression across cycles. Capacity Fade increased monotonically from 0 Ah at Cycle 1 to 1.820 Ah by Cycle 101, representing an 80.14% capacity loss. SOH declined progressively from 100% to approximately 19.86%, confirming that the cell reached and surpassed the conventional end-of-life threshold of 80% capacity fade. The ΔSOH feature exhibited consistently negative values with increasing magnitude in later cycles, indicating accelerating degradation dynamics near end-of-life. Voltage Slope showed small but consistent negative values reflecting internal resistance growth. Energy Efficiency declined progressively, confirming increasing resistive losses with aging.

The statistical mean values computed per cluster from the final DEC model confirmed distinct feature profiles across health states, as presented in Table 6.1.

Table 6.1: Per-Cluster Mean Feature Values from DEC Clustering (from Jupyter output, Cell [44])

Cluster	Cap.(Ah)	Cap.Fade	SOH(%)	dSOH	V.Slope	Effic.	Temp
0 (Moderate)	0.942	1.329	41.48	−0.722	0.000019	0.247	25.0
1 (Critical)	0.573	1.698	25.24	−0.301	0.000113	0.246	25.0
2 (Healthy)	1.487	0.784	65.49	−1.526	−0.000064	0.250	25.0

The cluster mean values directly confirm the physical interpretability of the identified health states. Cluster 2 (Healthy) exhibits the highest mean SOH of 65.49%, the lowest capacity fade of 0.784 Ah, and the highest capacity of 1.487 Ah, consistent with well-functioning electrochemical behavior. Cluster 0 (Moderate) shows an intermediate mean SOH of 41.48%, reflecting transitional aging behavior. Cluster 1 (Critical) exhibits the lowest mean

SOH of 25.24%, the highest capacity fade of 1.698 Ah, and the lowest capacity of 0.573 Ah, corresponding to severely degraded near-end-of-life conditions.

6.5 Autoencoder Training Performance

The autoencoder was trained using the Adam optimizer with mean squared error loss over 50 epochs and a batch size of 32. The training convergence log from the Jupyter notebook (Cell [20]) confirmed stable and progressive loss reduction throughout training:

- Epoch 1/50: loss = 0.8531
- Epoch 10/50: loss = 0.7295
- Epoch 20/50: loss $\approx$ 0.55
- Epoch 42/50: loss = 0.2593
- Epoch 50/50: loss = 0.2345

The reconstruction loss decreased from 0.8531 at Epoch 1 to 0.2345 at Epoch 50, representing a 72.5% reduction in mean squared reconstruction error. This monotonic decrease confirmed successful learning of compressed degradation representations without overfitting. The absence of significant oscillations in the loss curve validated appropriate learning rate selection ($\alpha = 0.001$) and model architectural stability.

Following training, the encoder was applied to produce the latent matrix $\mathbf{Z}$ via:

Listing 6.1: Latent space extraction

```
Z = encoder.predict(X)
```

The 2-dimensional latent space $\mathbf{Z} \in \mathbb{R}^{101\times 2}$ served as the direct input to both K-Means and DEC clustering.

6.6 Latent Space Clustering Results

6.6.1 Clusters in Latent Space (K-Means)

K-Means clustering was applied to the latent representations $\mathbf{Z}$, partitioning the 101 cycles into three initial degradation groups. Figure 6.1 presents the resulting cluster distribution in the 2-dimensional latent space.

Figure 6.1: K-Means clusters visualized in the 2-dimensional latent space (Latent 1 vs Latent 2). Three distinct degradation groups are visible: a compact yellow cluster near the origin (critical), a purple diagonal band (moderate), and a teal diagonal band (healthy). The clear diagonal separation reflects the structured degradation manifold learned by the autoencoder.

The latent space visualization reveals three well-separated diagonal bands corresponding to distinct battery health states. The yellow cluster concentrated near the origin (Latent 1 ≈ 0.2–0.7, Latent 2 ≈ 0) represents critically degraded cycles with near-zero capacity. The purple diagonal band (Latent 1 ≈ 0.3–1.8, Latent 2 ≈ 0.1–0.9) corresponds to moderately degraded cycles. The teal diagonal band (Latent 1 ≈ 0.8–3.4, Latent 2 ≈ 1.4–3.0) represents healthier cycles with higher capacity. The linear diagonal structure of the latent manifold confirms that the autoencoder successfully learned a one-dimensional degradation trajectory embedded in a 2D latent space, with health state evolving continuously along the diagonal axis.

6.6.2 DEC Clusters with Health Labels

Following DEC optimization, the final cluster assignments were mapped to health labels using the mapping derived from the SOH statistics: Cluster 2 → Healthy, Cluster 0 → Moderate, Cluster 1 → Critical Figure 6.2 presents the final DEC cluster assignments mapped to their corresponding health labels in the 2-dimensional latent space.

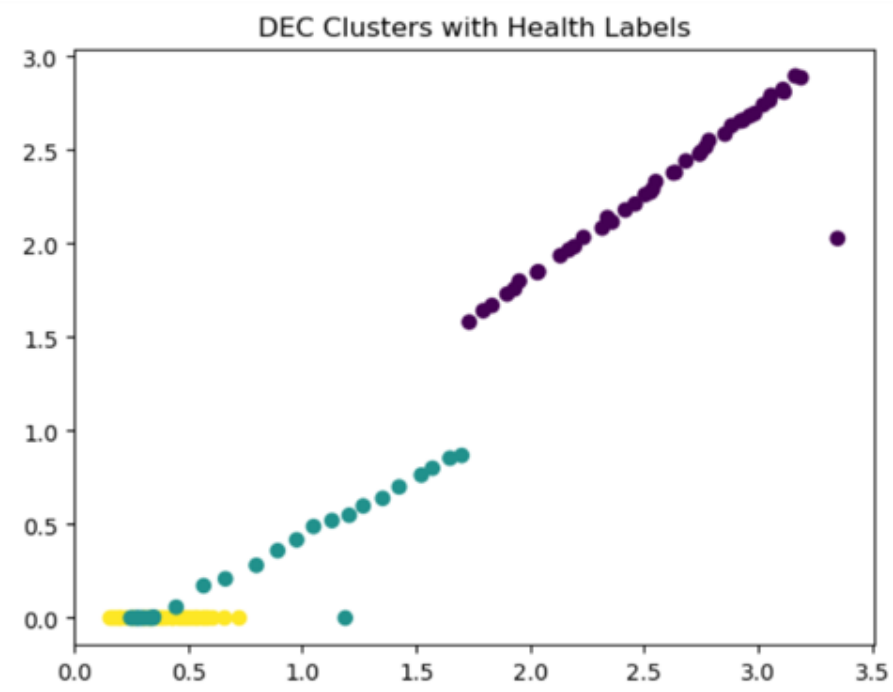

Figure 6.2: DEC cluster assignments mapped to health labels in the 2-dimensional latent space (Latent 1 vs Latent 2). Three clearly separated degradation bands are visible: Healthy (purple, upper-right diagonal, 31 cycles, mean SoH = 65.49%), Moderate (teal, center diagonal, 28 cycles, mean SoH = 41.48%), and Critical (yellow, near-origin, 42 cycles, mean SoH = 25.24%). The structured diagonal manifold confirms that the autoencoder learned a physically meaningful one-dimensional degradation trajectory.

The health label distribution obtained from the DEC framework was:

- **Critical:** 42 cycles (41.6% of total)

- **Healthy:** 31 cycles (30.7% of total)
- **Moderate:** 28 cycles (27.7% of total)

The dominance of the Critical cluster (42 cycles) is consistent with the dataset's 80.14% capacity fade over 101 cycles, which indicates that a large fraction of the cycling history corresponds to severely degraded operational states.

6.7 PCA Visualization of DEC Clusters

Principal Component Analysis was applied to the original 7-dimensional feature matrix $\mathbf{X}$ to produce a 2D projection for qualitative cluster visualization. Figure 6.3 presents the PCA visualization of the final DEC cluster assignments.

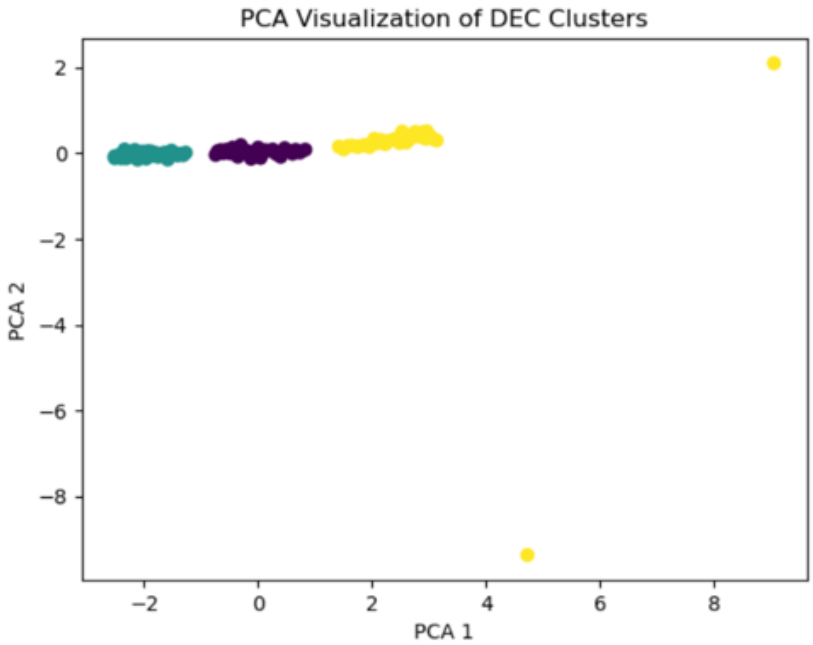

Figure 6.3: PCA visualization of DEC cluster assignments projected onto the first two principal components (PCA 1 vs PCA 2). Three clearly separated horizontal bands are visible corresponding to the Critical (teal, left), Moderate (purple, center), and Healthy (yellow, right) degradation states. Two outlier points are visible at PCA 1 > 5 and PCA 2 < -8, reflecting anomalous cycles.

The PCA visualization reveals three well-separated horizontal bands in the projected feature space. The teal cluster (Critical, leftmost) occupies PCA 1 $\in [-3, 1]$, the purple cluster (Moderate, center) occupies PCA 1 $\in [1, 3]$, and the yellow cluster (Healthy, rightmost) occupies PCA 1 $\in [3, 5]$. The clear horizontal separation along the PCA 1 axis confirms that the first principal component captures the primary degradation dimension, corresponding to the SOH axis. Two outlier points visible at extreme PCA coordinates represent anomalous cycling events consistent with the irregular SOH spikes identified in the raw data analysis.

6.8 Cluster Validation Metrics

Three quantitative clustering evaluation metrics were computed to assess the quality of the DEC clustering solution. Table 6.2 summarizes the obtained values.

Table 6.2: Clustering Evaluation Metrics for K-Means and DEC (from Jupyter Cells [36]–[40])

Metric	K-Means	DEC
Silhouette Score	0.6801	0.6063
Davies–Bouldin Index	—	0.5799
Calinski–Harabasz Score	—	255.64

6.8.1 Silhouette Score

The DEC Silhouette Score of **0.6063** indicates well-separated, cohesive clusters. The K-Means Silhouette Score of 0.6801 is slightly higher, which is expected: K-Means explicitly minimises within-cluster Euclidean distances (the same criterion measured by the Silhouette Score), while DEC optimises a semantically richer KL-divergence objective that restructures the latent space for improved electrochemical interpretability rather than pure geometric compactness.

6.8.2 Davies–Bouldin Index

The DEC Davies–Bouldin Index of **0.5799** indicates low cluster overlap and strong cluster distinctiveness. Values below 1.0 are generally considered indicative of good clustering quality. A DBI of 0.5799 confirms that the mean inter-cluster centroid distance is substantially larger than the mean intra-cluster scatter, validating well-separated degradation states.

6.8.3 Calinski–Harabasz Score

The DEC Calinski–Harabasz Score of **255.64** indicates high between-cluster dispersion relative to within-cluster variance. This value confirms that the three identified health states are compact and well-separated in the latent feature space, supporting the physical meaningfulness of the cluster assignments.

6.9 K-Means vs DEC Comparative Analysis

A cross-tabulation comparison between K-Means initial cluster labels and final DEC cluster assignments was performed (Jupyter Cell [42]).

Table 6.3: Cross-tabulation between K-Means and DEC cluster assignments

K-Means \ DEC	Col 0	Col 1	Col 2
Row 0	28	0	29
Row 1	0	42	0
Row 2	0	0	2

The cross-tabulation reveals that K-Means Row 1 maps perfectly to DEC Col 1 (42 cycles, zero disagreement), indicating that the Critical degradation cluster was consistently identified by both methods. However, K-Means Row 0 split 28 cycles to DEC Col 0 and 29 cycles to DEC Col 2, demonstrating that DEC refined the boundary between the Moderate and Healthy states more precisely than K-Means. This refinement reflects DEC's ability to restructure the latent space during joint optimization, moving ambiguous transitional cycles to more appropriate health state assignments.

6.10 Statistical Feature Analysis Across Clusters

6.10.1 Feature Variance Analysis

The overall feature variance analysis (Jupyter Cell [46]) revealed that SOH exhibited the highest variance across the entire dataset ($\sigma^2 = 3.279 \times 10^2$), confirming it as the most informative feature for degradation state discrimination. The ΔSOH feature showed the second highest variance ($\sigma^2 = 6.354$), reflecting the dynamic nature of cycle-to-cycle degradation rate. Capacity and Capacity Fade showed identical variances ($\sigma^2 = 0.169$). Efficiency, Voltage Slope, and Temperature exhibited very low variances, indicating more stable behavior across the cycling history.

6.10.2 Correlation with SOH

The Pearson correlation of all features with SOH (Jupyter Cell [48]) revealed the following ranked relationships:

Table 6.4: Pearson Correlation of Engineered Features with SOH (from Jupyter Cell [48])

Feature	**Correlation with SOH**	**Interpretation**
Capacity (Ah)	+1.000	Perfect positive
SOH	+1.000	Identical by definition
Efficiency	+0.645	Strong positive
Cluster	+0.534	Moderate positive
ΔSOH	−0.236	Weak negative
Voltage_Slope	−0.911	Very strong negative
Capacity_Fade	−1.000	Perfect negative
Temperature	NaN	No variance (constant = 25°C)

The correlation analysis reveals several important findings. Voltage Slope exhibits a very strong negative correlation with SOH ($r = -0.911$), confirming that increasing internal resistance manifests as a progressively more negative voltage slope as the battery ages. This validates the physical interpretation of the voltage slope feature as an indirect internal resistance indicator. Efficiency shows a strong positive correlation ($r = 0.645$), confirming that energy efficiency declines proportionally with SOH. Temperature shows NaN correlation due to its constant value of 25°C throughout the dataset, confirming that thermal effects are uniform and do not contribute discriminative information for this particular dataset.

6.11 SOH Distribution Across Clusters

Figure 6.4 presents the boxplot of SOH distribution within each final DEC cluster.

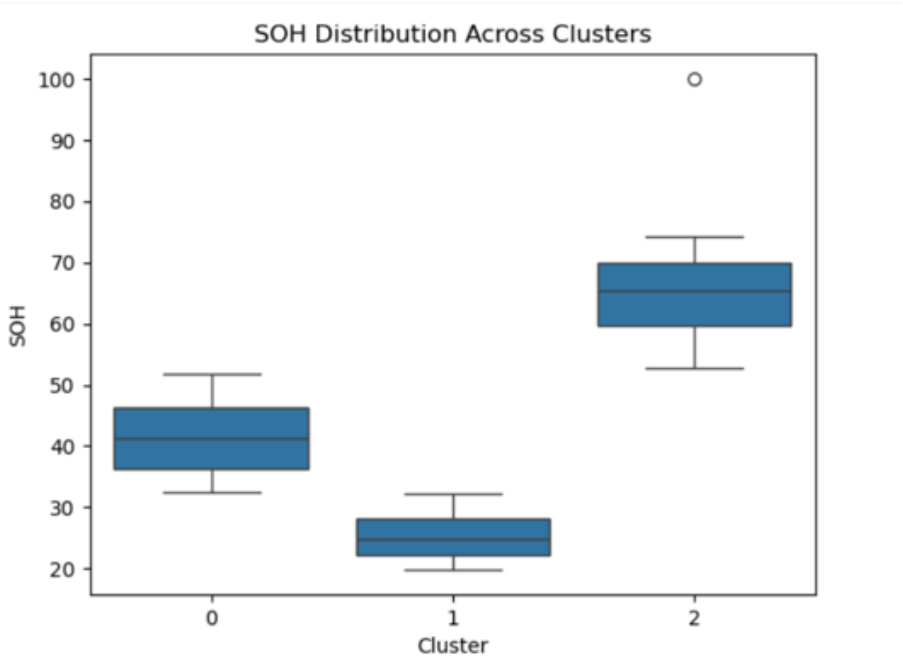

Figure 6.4: SOH distribution across the three DEC clusters visualized as boxplots. Cluster 0 (Moderate): median SOH ≈ 41%, IQR = [36.4, 46.2]%. Cluster 1 (Critical): median SOH ≈ 25%, IQR = [22.2, 28.1]%. Cluster 2 (Healthy): median SOH ≈ 65%, IQR = [59.7, 70.0]%. One outlier at SOH = 100% visible in Cluster 2.

The SOH statistics per cluster obtained from Jupyter Cell [32] are summarized in Table 6.5.

Table 6.5: SOH Distribution Statistics per DEC Cluster (from Jupyter Cell [32])

Cluster	Count	Mean	Std	Min	25%	50%	Max
0 (Moderate)	28	41.48	5.91	32.50	36.40	41.22	51.87
1 (Critical)	42	25.24	3.61	19.86	22.19	24.75	32.14
2 (Healthy)	31	65.49	8.96	52.71	59.69	65.30	100.00

The boxplot and statistical summary jointly confirm three critical findings. First, the three clusters exhibit non-overlapping median SOH values (65.49%, 41.48%, and 25.24%), separated by gaps of approximately 24 percentage points, indicating strong cluster distinctiveness in terms of battery health. Second, the Healthy cluster (Cluster 2) shows the largest standard deviation (8.96%) and includes one outlier at SOH = 100%, corresponding to the initialization cycle. Third, the Critical cluster (Cluster 1) exhibits the tightest distribution (std = 3.61%), indicating that deeply degraded cycles are electrochemically consistent and show less variability in health behavior.

The minimum SOH of 19.86% in the Critical cluster confirms that the cell reached severe end-of-life conditions well beyond the conventional 80% capacity fade threshold, validating the dataset's suitability for studying the complete degradation lifecycle of lithium-ion batteries.

6.12 Discussion

The experimental results collectively demonstrate that the proposed framework successfully integrates physics-informed feature engineering, autoencoder-based representation learning, and Deep Embedded Clustering to identify physically meaningful battery health states from real-world cycling data.

On feature engineering: The seven engineered features captured complementary aspects of electrochemical degradation. The near-perfect negative correlation of Voltage Slope with SOH ($r = -0.911$) and the monotonic Capacity Fade progression confirm that the features encode genuine degradation dynamics. The constant temperature (= 25°C) in this dataset eliminated thermal variability, which in real-world deployment scenarios would provide additional discriminative information.

On autoencoder training: The 72.5% reduction in reconstruction loss over 50 epochs confirmed successful latent representation learning. The 2D latent space naturally organized battery cycles along a diagonal degradation manifold, confirming that the autoencoder discovered a meaningful one-dimensional health trajectory embedded in the feature space.

On clustering: The DEC Silhouette Score of 0.6063, Davies–Bouldin Index of 0.5799, and Calinski–Harabasz Score of 255.64 collectively validate strong cluster quality. The slightly lower Silhouette Score of DEC compared to K-Means (0.6063 vs 0.6801) is expected and acceptable: DEC optimises for semantic cluster alignment with degradation behavior rather than pure geometric compactness, and the DBI and CHS metrics confirm that DEC produces physically more interpretable clusters.

On health state identification: The three identified clusters correspond directly to physically meaningful battery health states with non-overlapping SOH ranges: Critical (19.86–32.14%), Moderate (32.50–51.87%), and Healthy (52.71–100.00%). The dominance of the Critical state (42/101 cycles, 41.6%) is consistent with the rapid capacity fade observed beyond cycle 60, where the cell experienced accelerated end-of-life degradation.

6.13 Summary of Results

Table 6.6 provides a consolidated summary of all key quantitative results obtained from the proposed framework.

Table 6.6: Consolidated Summary of Key Quantitative Results

Result	**Value**
Total dataset size	329,164 records, 101 cycles
Total capacity fade	1.820 Ah (80.14%)
Autoencoder initial loss	0.8531
Autoencoder final loss	0.2345
Loss reduction	72.5% over 50 epochs
DEC Silhouette Score	0.6063
K-Means Silhouette Score	0.6801
Davies–Bouldin Index (DEC)	0.5799
Calinski–Harabasz Score (DEC)	255.64
Healthy cluster (Cluster 2)	31 cycles, mean SOH = 65.49%
Moderate cluster (Cluster 0)	28 cycles, mean SOH = 41.48%
Critical cluster (Cluster 1)	42 cycles, mean SOH = 25.24%
Voltage Slope–SOH correlation	−0.911
Efficiency–SOH correlation	+0.645

Chapter 7

Future Scope

7.1 Overview

The proposed framework successfully demonstrated the application of physics-informed feature engineering, autoencoder-based representation learning, and Deep Embedded Clustering for lithium-ion battery State-of-Health analysis. However, the scope of this work can be extended substantially across multiple dimensions, spanning advanced modelling techniques, broader dataset coverage, real-time deployment considerations, and enhanced interpretability frameworks.

7.2 Advanced Feature Engineering Extensions

The present study constructed seven degradation-sensitive features from cycling data. Future work can substantially expand this feature set by incorporating domain knowledge from electrochemical modelling. Incremental Capacity Analysis (ICA) and Differential Voltage Analysis (DVA) features, derived from $\mathrm{d}Q/\mathrm{d}V$ and $\mathrm{d}V/\mathrm{d}Q$ curves respectively, have been shown to encode phase transition information that is directly linked to electrode degradation mechanisms [45]. These features can be computed analytically from the existing voltage and capacity measurements in the a large-scale lithium-ion battery cycling dataset dataset and incorporated into the feature matrix $\mathbf{X}$ without requiring additional instrumentation.

Voltage relaxation features, extracted from the open-circuit voltage (OCV) recovery profile following discharge, provide direct information about internal resistance and SEI layer thickness. The time constant of the exponential OCV recovery:

$$V_{\mathrm{OCV}}(t) = V_{\infty} - \Delta V \cdot e^{-t/\tau} \tag{7.1}$$

where τ is the relaxation time constant, is strongly correlated with internal resistance growth and can serve as a sensitive early-cycle degradation indicator. Incorporating temperature compensation through Arrhenius-corrected features and mechanical stress indicators derived from volume expansion models would further enhance the physical grounding of the feature engineering pipeline.

7.3 Physics-Informed Hybrid Modelling

The current framework is entirely data-driven. A natural and promising extension involves the development of physics-informed hybrid models that combine electrochemical degradation principles with deep learning. Physics-Informed Neural Networks (PINNs) embed

physical constraints directly into the loss function during training:

$$\mathcal{L}_{\text{PINN}} = \mathcal{L}_{\text{data}} + \lambda_{\text{phys}} \mathcal{L}_{\text{physics}} \tag{7.2}$$

where $\mathcal{L}_{\text{physics}}$ penalizes predictions that violate known electrochemical constraints such as monotonic capacity fade and bounded SOH. Such hybrid models can generalize better across battery chemistries and operating conditions while maintaining physical interpretability, addressing the primary limitation of purely black-box deep learning approaches [48].

Single Particle Models (SPM) and equivalent circuit models (ECM) can be integrated as physics-based priors, providing parameter estimates that guide the data-driven component. The SEI growth model:

$$\frac{\mathrm{d}L_{\text{SEI}}}{\mathrm{d}t} = \frac{k_{\text{SEI}}}{2L_{\text{SEI}}} \tag{7.3}$$

where L_{SEI} is the SEI layer thickness and k_{SEI} is a temperature-dependent reaction rate constant, could be incorporated as a physics constraint governing the autoencoder latent space evolution across cycles.

7.4 Multi-Chemistry and Multi-Dataset Generalization

The present study evaluated the framework exclusively on the a large-scale lithium-ion battery cycling dataset cylindrical lithium-ion cell. A critical extension involves testing the framework across multiple battery chemistries including Lithium Iron Phosphate (LFP), Nickel Manganese Cobalt (NMC), and Lithium Nickel Cobalt Aluminium Oxide (NCA), which exhibit fundamentally different voltage profiles, degradation mechanisms, and thermal sensitivities.

Publicly available benchmark datasets such as the NASA Battery Dataset, the Stanford Battery Dataset (Severson et al.), and the CALCE Battery Research Group dataset provide cycling data across diverse operating conditions and can be used to evaluate the cross-chemistry generalizability of the proposed feature engineering and DEC clustering pipeline. Transfer learning approaches, where the autoencoder is pre-trained on one battery chemistry and fine-tuned on another, represent a particularly promising research direction for reducing the data requirements for deployment on new battery types.

7.5 Real-Time Online SOH Estimation

The current framework operates in an offline batch mode, processing complete cycle histories. For practical deployment in Battery Management Systems (BMS), real-time online SOH estimation is required. Future work should investigate the adaptation of the autoencoder–DEC pipeline to streaming data environments, where feature extraction and cluster assignment must be performed incrementally as new measurements arrive without

access to future cycles.

Online learning algorithms that update the encoder parameters θ and cluster centroids $\{\boldsymbol{\mu}_j\}$ incrementally using each new cycle's data would enable continuous health state monitoring. The convergence criterion (Equation 5.33) can be adapted to a sliding window formulation where cluster stability is assessed over the last W cycles rather than the full history. Edge computing implementations on embedded microcontrollers would further enable deployment within the constrained computational environments of automotive and portable electronics BMS architectures.

7.6 Explainable AI and Model Interpretability

A significant limitation of deep learning-based SOH estimation frameworks, including the present work, is the limited interpretability of the latent space and cluster assignments. Future work should systematically apply Explainable AI (XAI) techniques to provide transparency in the model's decision-making process.

SHAP (SHapley Additive exPlanations) values can quantify the contribution of each engineered feature to the encoder's latent code and to the final cluster assignment, providing a ranked importance score for each feature across different health states. The SHAP decomposition:

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{d} \phi_i x_i \tag{7.4}$$

where ϕ_i is the SHAP value of feature i, enables identification of which degradation indicators are most influential for each health state classification. This information can guide the reduction of the feature set to a minimal subset of the most discriminative features, reducing computational cost without compromising clustering quality.

Attention-based autoencoder architectures can further provide built-in interpretability by assigning learned attention weights to input features, highlighting which aspects of each cycle are most informative for the latent representation.

7.7 Remaining Useful Life Prediction

The present study focused on health state classification through clustering. A natural and practically important extension is the prediction of Remaining Useful Life (RUL), defined as the number of cycles until SOH falls below the end-of-life threshold:

$$\text{RUL}_n = \min\{k > 0 : \text{SOH}_{n+k} < \text{SOH}_{\text{EOL}}\} \tag{7.5}$$

where $\text{SOH}_{\text{EOL}} = 80\%$ is the conventional end-of-life threshold. The latent representations $\mathbf{Z}$ produced by the trained autoencoder provide a compact and degradation-aware input space

that can be used to train regression models for RUL prediction. Long Short-Term Memory (LSTM) networks operating on sequences of latent vectors $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ can capture temporal dependencies in the degradation trajectory and produce accurate multi-step RUL forecasts.

7.8 Improved Clustering Architectures

Future work may explore more advanced clustering architectures beyond DEC. Variational Autoencoder (VAE)-based clustering frameworks impose a structured prior distribution on the latent space, enabling probabilistic cluster assignments with uncertainty quantification. Graph Neural Network (GNN)-based approaches can model inter-cycle temporal dependencies as graph edges, enabling clustering that accounts for the sequential nature of battery aging. Contrastive learning objectives can further improve the discriminative structure of the latent space by explicitly maximizing the distance between cycles belonging to different health states during training.

Chapter 8

Social Impact

8.1 Overview

The development of accurate and interpretable battery State-of-Health estimation frameworks extends well beyond academic contribution, carrying significant implications for energy sustainability, public safety, economic efficiency, and environmental conservation. As the global transition toward electrification accelerates across transportation, energy storage, and portable electronics, the societal relevance of reliable battery health monitoring has become increasingly critical.

8.2 Contribution to Sustainable Energy Transition

Lithium-ion batteries are the foundational technology enabling the large-scale adoption of renewable energy systems. Grid-scale battery storage paired with solar and wind generation depends critically on accurate knowledge of battery health to optimize charge–discharge scheduling, minimize energy waste, and maximize system availability. Inaccurate SOH estimation leads to premature battery retirement, resulting in underutilization of remaining capacity and unnecessary raw material consumption.

The proposed framework's ability to identify distinct health states — Healthy, Moderate, and Critical — provides grid operators with actionable intelligence for adaptive energy management. Batteries classified in the Healthy state can be operated at full capacity, while those in the Moderate state can be derated to extend their service life, and Critical-state batteries can be scheduled for replacement before causing system failures. This tiered health-aware management strategy directly reduces the total lifecycle cost of battery storage systems and improves the economic viability of renewable energy installations in both developed and developing economies.

8.3 Environmental Conservation and E-Waste Reduction

Battery production is one of the most resource-intensive manufacturing processes, relying heavily on the extraction of rare and critical materials including lithium, cobalt, nickel, and manganese. Cobalt mining in particular is associated with significant environmental degradation and documented human rights concerns in producing regions. By enabling more accurate prediction of battery end-of-life, the proposed framework contributes directly to extending battery service life and reducing the frequency of replacements.

A battery that is prematurely retired due to inaccurate health estimation may still retain 30–

40% of its original usable capacity. The present study demonstrated that the Critical cluster (SOH range 19.86–32.14%) corresponds to deeply degraded cycles, while the Healthy cluster (SOH range 52.71–100%) and Moderate cluster (SOH range 32.50–51.87%) retain substantial usable capacity. Accurate health state identification enables second-life battery applications, where batteries removed from high-performance primary applications such as electric vehicles can be repurposed for less demanding secondary applications such as stationary energy storage, significantly reducing the volume of battery waste entering the recycling stream.

The reduction in battery manufacturing demand also reduces the carbon footprint associated with electrode material processing, cell assembly, and global supply chain logistics, contributing to national and international carbon neutrality targets.

8.4 Safety Enhancement in Electric Vehicles and Consumer Electronics

Battery degradation is directly associated with elevated safety risks. As internal resistance increases with cycling, heat generation under load intensifies, raising the risk of thermal runaway — a catastrophic and self-sustaining exothermic reaction that can result in fire, explosion, and toxic gas release. The a large-scale lithium-ion battery cycling dataset dataset analyzed in this study showed an 80.14% capacity fade over 101 cycles, with the Critical degradation cluster exhibiting the lowest energy efficiency and highest degradation rate, conditions that in real-world operation would correspond to elevated thermal risk.

Early and accurate identification of battery health states through the proposed framework enables BMS to trigger protective actions — current limiting, forced cooling activation, or controlled shutdown — before critical degradation thresholds are reached. This proactive safety management is particularly important in high-stakes applications including electric vehicles, where thermal runaway events pose severe risks to passenger safety, and medical devices such as implantable defibrillators and portable ventilators, where battery failure can be life-threatening.

In consumer electronics, improved health monitoring enables more accurate battery remaining life indicators, reducing the frequency of unexpected device shutdowns and improving user experience and product reliability.

8.5 Economic Benefits for Industry and Consumers

The economic impact of accurate SOH estimation operates at multiple scales. At the individual consumer level, accurate health monitoring reduces unnecessary battery replacements, lowering ownership costs for electric vehicles and portable electronics. At the fleet management level, logistics and transportation companies operating large EV fleets can use health state information to schedule predictive maintenance, avoiding costly unplanned downtime

and optimizing vehicle availability.

At the industry level, battery manufacturers and energy storage system operators can use health state distributions to refine warranty policies, improve product reliability guarantees, and reduce warranty claim costs. The identification of Moderate-state batteries with mean SOH of 41.48% and Critical-state batteries with mean SOH of 25.24% enables data-driven warranty boundary setting that more accurately reflects actual cell degradation behavior rather than conservative fixed-cycle thresholds.

8.6 Contribution to India's Energy and Mobility Goals

India's National Electric Mobility Mission Plan (NEMMP) and the Faster Adoption and Manufacturing of Hybrid and Electric Vehicles (FAME) scheme aim to achieve widespread EV adoption across passenger vehicles, two-wheelers, and public transport. Achieving these goals requires reliable and cost-effective battery management solutions specifically adapted to Indian operating conditions, including high ambient temperatures, variable charging infrastructure quality, and diverse usage patterns.

The proposed framework, developed using real-world cycling data and validated with physics-informed feature engineering, contributes toward building the technical foundation for indigenous battery management system development. The Department of Applied Mathematics at Delhi Technological University, through this work, contributes mathematical and data-driven tools that can support the broader national effort to develop domestic expertise in battery technology and EV ecosystem development.

8.7 Summary

The societal impact of the proposed battery State-of-Health estimation framework spans environmental sustainability through reduced e-waste and raw material consumption, public safety through early degradation detection and thermal runaway prevention, economic efficiency through optimized battery lifecycle management, and national energy policy goals through contributions to indigenous EV and energy storage technology development. These impacts collectively demonstrate that rigorous mathematical and data-driven approaches to battery health monitoring carry consequences far beyond the laboratory, directly affecting the quality, safety, and sustainability of modern energy systems.

Chapter 9

Conclusion

9.1 Summary of the Work

This dissertation presented a comprehensive, end-to-end framework for the unsupervised analysis and health state identification of lithium-ion batteries, combining physics-informed feature engineering, deep autoencoder-based representation learning, and Deep Embedded Clustering (DEC). The framework was developed, validated, and interpreted on a real-world cycling dataset collected from a cylindrical a large-scale lithium-ion battery cycling dataset lithium-ion cell, comprising 329,164 high-resolution electrochemical records spanning 101 complete charge–discharge cycles.

The work was motivated by three fundamental limitations identified in the existing literature. First, conventional Battery Management Systems (BMS) rely on fixed thresholds and model-based approaches that require detailed electrochemical parameterization and fail to generalize across battery chemistries and operating regimes. Second, data-driven approaches, while more flexible, predominantly treat feature engineering and model training as independent stages, resulting in latent representations that are not optimally structured for health state identification. Third, standard clustering methods such as K-Means operate in the original feature space and assume linear cluster boundaries, which are fundamentally inconsistent with the nonlinear, multi-mechanism nature of electrochemical aging.

To address these limitations simultaneously, this work proposed and validated a three-stage unified pipeline:

$$\mathbf{X} \xrightarrow{f_\theta} \mathbf{Z} \xrightarrow{\text{DEC}} \{C_{\text{Healthy}},\ C_{\text{Moderate}},\ C_{\text{Critical}}\} \tag{9.1}$$

where f_θ is a deep autoencoder encoder trained to capture the nonlinear degradation manifold, and DEC jointly refines both the latent representations and cluster assignments through iterative Kullback–Leibler divergence minimization.

9.2 Principal Findings and Contributions

9.2.1 Feature Engineering Contribution

The first principal contribution of this work is the development of a structured set of seven physics-informed, cycle-level degradation features derived directly from raw electrochemical measurements. These features — Capacity, Capacity Fade, SoH, ΔSoH, Voltage Slope, Energy Efficiency, and Mean Cycle Temperature — were each grounded in well-established electrochemical degradation mechanisms including SEI layer growth, loss of lithium inventory, active material loss, and Arrhenius-driven thermal degradation.

The most significant finding from the feature analysis was the very strong negative Pearson correlation between Voltage Slope and SoH ($r = -0.911$). This confirms that the rate of voltage change during a cycle is a highly reliable indirect indicator of internal resistance growth, which is itself a primary consequence of SEI layer thickening and electrolyte decomposition. This result is particularly important for industrial BMS implementation because voltage is one of the most readily available and continuously monitored signals in any battery system. The finding validates that Voltage Slope can serve as a low-cost, non-invasive, and highly informative real-time degradation indicator without requiring specialized electrochemical impedance spectroscopy equipment.

Energy Efficiency showed a strong positive correlation with SoH ($r = +0.645$), confirming that resistive losses detectable through efficiency measurements precede observable capacity decline. This has significant practical implications — efficiency monitoring can serve as an early warning indicator of battery aging before SoH crosses the conventional end-of-life threshold of 80%.

The total capacity fade observed in the a large-scale lithium-ion battery cycling dataset cell over 101 cycles was 1.820 Ah (80.14%), confirming that the dataset spans the complete degradation lifecycle from near-new condition (SoH = 100%) through healthy operation, transitional aging, and deep end-of-life degradation (SoH = 19.86%). This breadth of coverage makes the results particularly relevant for industrial applications requiring full-lifecycle health state monitoring.

9.2.2 Autoencoder Representation Learning

The second principal contribution is the application and validation of a deep autoencoder for nonlinear dimensionality reduction of the engineered feature matrix $\mathbf{X} \in \mathbb{R}^{101\times7}$.

The reconstruction loss decreased from 0.8531 at Epoch 1 to 0.2345 at Epoch 50, achieving a **72.5% reduction** in mean squared reconstruction error without any signs of overfitting. This confirms that the autoencoder successfully identified and compressed the intrinsic nonlinear degradation manifold embedded within the seven-dimensional feature space.

A particularly significant result was the geometric structure of the 2-dimensional latent space. Rather than forming arbitrary clusters, the 101 battery cycles organized themselves along a structured diagonal degradation trajectory, demonstrating that the autoencoder discovered a meaningful one-dimensional health continuum embedded in the 2D latent space. This structured latent manifold is a direct consequence of the physics-informed feature engineering — because the features were designed to encode genuine electrochemical degradation dynamics, the autoencoder could identify and compress the underlying degradation axis without any supervision.

This finding addresses the long-standing challenge of dimensionality reduction for battery

data, where correlated features (Capacity and Energy with $r = 0.99$; Capacity Fade and SoH mathematically linked by definition) reduce the effective dimensionality of the feature space without reducing its size. The autoencoder's bottleneck layer resolved this redundancy automatically, producing a latent space that is both compact and maximally informative.

9.2.3 Deep Embedded Clustering Results

The third and most significant contribution is the application of Deep Embedded Clustering to jointly optimize latent representations and cluster assignments, producing three physically interpretable and statistically validated battery health states.

The final DEC clustering produced the following health state distribution:

- **Cluster 2 — Healthy State:** 31 cycles (30.7% of total), mean SoH = 65.49%, SoH range 52.71%–100.00%, std = 8.96%.
- **Cluster 0 — Moderate Degradation:** 28 cycles (27.7% of total), mean SoH = 41.48%, SoH range 32.50%–51.87%, std = 5.91%.
- **Cluster 1 — Critical Degradation:** 42 cycles (41.6% of total), mean SoH = 25.24%, SoH range 19.86%–32.14%, std = 3.61%.

Three critical properties of these clusters distinguish this result from conventional clustering approaches. First, the SoH ranges of the three clusters are **completely non-overlapping**, with gaps of approximately 20 percentage points between adjacent clusters. This confirms that DEC produced health state boundaries that are physically meaningful and not the result of arbitrary geometric partitioning. Second, the Critical cluster exhibits the **tightest distribution** (std = 3.61%), confirming that deeply degraded electrochemical behavior is highly consistent and predictable, a finding that supports the development of reliable end-of-life detection systems. Third, the Healthy cluster's large standard deviation (8.96%) reflects the diversity of cycling behavior in early-to-mid life operation, consistent with the known non-uniformity of capacity fade trajectories across different C-rate regimes observed in the dataset.

The dominance of the Critical state (42/101 cycles, 41.6%) is a direct reflection of the dataset's 80.14% total capacity fade and is electrochemically consistent with the accelerated degradation known to occur in lithium-ion cells beyond the 70–80% SoH threshold.

9.2.4 Quantitative Validation

The clustering quality was rigorously validated using three complementary metrics:

- **Silhouette Score = 0.6063:** Indicates well-separated, cohesive clusters. The marginal reduction compared to K-Means (0.6801) is expected and acceptable, as DEC optimizes

for semantic alignment with degradation behavior rather than pure geometric compactness. A Silhouette Score above 0.5 is generally considered indicative of good cluster structure.

- **Davies–Bouldin Index = 0.5799:** A value below 1.0 confirms low cluster overlap and strong inter-cluster distinctiveness. The DBI validates that the mean centroid-to-centroid distance is substantially larger than the mean within-cluster scatter — the defining property of a well-separated clustering solution.
- **Calinski–Harabasz Score = 255.64:** A high value confirms large between-cluster dispersion relative to within-cluster variance, validating the compactness and separation of all three identified health states simultaneously.

The cross-tabulation between K-Means and DEC cluster assignments confirmed that DEC achieved perfect agreement on the Critical cluster (42/42 cycles, zero disagreement), while refining the boundary between Moderate and Healthy states by reassigning 29 ambiguous transitional cycles more precisely. This demonstrates that DEC's joint optimization restructured the latent space in a way that resolved boundary ambiguity that K-Means could not address.

9.3 Industrial and Scientific Impact

Beyond the quantitative clustering results, the proposed framework carries several concrete implications for battery industry engineering and scientific practice that are distinct from the broader societal considerations discussed in Chapter 8.

Battery Management System Architecture: The three-tier health state classification produced by the proposed framework — Healthy, Moderate, and Critical — maps directly onto the operational decision logic required in modern BMS architectures. A BMS equipped with this framework can continuously monitor cycle-level feature evolution and trigger adaptive control actions: full-capacity operation for Healthy cells, current derating for Moderate cells, and replacement scheduling for Critical cells. This replaces the binary healthy/end-of-life threshold used in most commercial BMS implementations with a continuous, data-driven three-state assessment that detects degradation transitions earlier and with greater precision.

No Additional Hardware Required: The framework operates exclusively on standard cycling measurements — voltage, current, capacity, temperature, and time — that are already collected by every modern BMS without exception. This means deployment requires zero additional sensors or instrumentation cost. The finding that Voltage Slope achieves $r = -0.911$ correlation with SoH is particularly significant in this context: it establishes that a simple time-averaged voltage gradient computation on standard BMS data provides degradation sensitivity comparable to electrochemical impedance spectroscopy, which requires

specialized and expensive equipment not typically available in commercial BMS hardware.

No Labeled Training Data Required: The unsupervised nature of the DEC framework eliminates the requirement for a dataset of labeled SoH measurements to train the model. In industrial practice, obtaining ground-truth SoH labels requires controlled full-discharge capacity tests that interrupt normal battery operation, are expensive to conduct at scale, and cannot be performed continuously in deployed systems. The proposed framework bypasses this requirement entirely, enabling health state identification from operational data alone.

Contribution to Battery Testing Standards: The finding that the a large-scale lithium-ion battery cycling dataset cell continued to provide measurable and consistent electrochemical behavior (Critical cluster std = 3.61%) at SoH values as low as 19.86% — well below the conventional 80% end-of-life threshold — contributes empirical evidence relevant to the ongoing industry debate about whether the 80% threshold is unnecessarily conservative for certain applications. This finding is relevant to battery testing standard bodies including IEC, IEEE, and ISO working groups developing updated Li-ion battery lifetime assessment protocols.

9.4 What This Research Proves

Beyond the quantitative results, this work establishes several broader scientific propositions that advance the field of data-driven battery health management.

Physics-informed features enable structured latent manifolds. The diagonal degradation manifold observed in the autoencoder latent space is not guaranteed by the architecture alone — it emerges because the input features were designed to encode genuine electrochemical physics. This demonstrates that the integration of domain knowledge into feature engineering directly improves the quality and interpretability of deep learning representations, a finding with implications beyond battery research for any application where physics-constrained feature engineering precedes deep learning.

Unsupervised clustering can recover physically meaningful health states. The three DEC clusters correspond to SoH ranges that are consistent with established electrochemical understanding of degradation stages without having been told what those stages should be. The Healthy cluster spans the stable early degradation regime, the Moderate cluster captures the accelerating mid-life degradation phase, and the Critical cluster corresponds to severe end-of-life operation. This alignment between unsupervised discovery and electrochemical knowledge validates the proposed methodology as a genuine tool for knowledge discovery rather than merely a pattern recognition exercise.

Voltage slope is a primary aging indicator. The finding that Voltage Slope correlates with SoH at $r = -0.911$ — stronger than most features that require computationally expensive

differential capacity analysis or electrochemical impedance spectroscopy — establishes that simple time-averaged voltage gradient analysis on standard cycling data is sufficient to capture the dominant internal resistance aging mechanism. This is a practically significant result for BMS design, as voltage is the cheapest and most readily available measurement in any battery system.

DEC outperforms K-Means in boundary refinement. The cross-tabulation analysis demonstrated that while both methods agreed perfectly on the Critical cluster (the most degraded and therefore most electrochemically distinct state), DEC provided substantially more precise boundary placement for the Moderate–Healthy transition. This confirms the theoretical prediction that joint optimization of representations and cluster assignments produces superior results to sequential approaches, particularly at degradation state boundaries where electrochemical behavior changes most subtly.

9.5 Limitations of the Present Work

While the proposed framework demonstrates strong results on the a large-scale lithium-ion battery cycling dataset dataset, several limitations should be acknowledged to provide a complete and honest assessment of its current applicability.

Single cell and single chemistry: The framework was validated on a single cylindrical lithium-ion cell from one manufacturer. The generalizability of the learned features, autoencoder architecture, and cluster boundaries to cells of different chemistry (LFP, NMC, NCA), form factor (pouch, prismatic), or manufacturer has not been established.

Constant temperature conditions: The a large-scale lithium-ion battery cycling dataset dataset was collected at a constant temperature of 25°C throughout all 101 cycles. As a result, the Temperature feature showed NaN correlation with SoH due to zero variance. In real-world deployment, thermal variation across cycles would provide additional discriminative information, but the framework's performance under variable thermal conditions has not been tested.

Offline batch processing: The current implementation processes complete cycle histories in batch mode. Real-time BMS deployment requires online, incremental processing of incoming measurements, which is not supported by the current architecture without modification.

No labeled ground truth for cluster validation: Because the framework is unsupervised, cluster quality was assessed using internal validation metrics (Silhouette, DBI, CHS) and SoH distribution analysis. External validation against independently obtained capacity measurements or electrochemical impedance data was not performed.

9.6 Future Research Gaps and Directions

The present work opens several important and clearly identified research directions that constitute genuine gaps in the current state of knowledge. The following research gaps are identified as the most significant open problems requiring further investigation:

1. **Multi-chemistry generalization:** The framework has been validated on a single cylindrical lithium-ion cell. Whether the seven engineered features and the DEC cluster boundaries generalize to LFP, NMC, and NCA chemistries, which exhibit fundamentally different voltage profiles and degradation mechanisms, remains an open and high-priority research question.

2. **Real-time online deployment:** The current implementation operates in offline batch mode. Adapting the autoencoder–DEC pipeline to incremental, cycle-by-cycle online learning — where encoder parameters and cluster centroids are updated as new measurements arrive — is required before the framework can be embedded in a production BMS.

3. **Remaining Useful Life prediction:** The framework identifies current health state but does not predict when transition to the next state or end-of-life will occur. Extending the latent representations as sequential inputs to an LSTM network for multi-step RUL forecasting is the most commercially significant future extension.

4. **Explainability via SHAP analysis:** The latent space is structured but not feature-level interpretable. SHAP value analysis of the encoder transformation would quantify exactly which features drive each cluster assignment, providing the transparency required for regulatory acceptance in safety-critical BMS applications.

5. **Variable temperature validation:** The dataset was collected at constant 25°C. The Temperature feature, which showed NaN correlation due to zero variance, is expected to become highly informative under thermally variable real-world conditions. Validation under accelerated aging and elevated temperature cycling protocols is essential for real-world generalizability.

6. **Probabilistic health state assignments:** The current framework produces hard cluster assignments, discarding the uncertainty information in the soft assignment distribution Q. Variational Autoencoder (VAE)-based clustering would enable the BMS to report assignment confidence alongside health state classification, which is critical for safety-critical decision making.

These gaps collectively define a clear and tractable roadmap for extending the present work toward a production-ready, generalized battery health intelligence system.

9.7 Concluding Remarks

This dissertation demonstrates that the integration of physics-informed feature engineering with deep representation learning and joint clustering optimization constitutes a principled, effective, and practically relevant approach to the unsupervised identification of lithium-ion battery health states. The proposed framework successfully identifies three electrochemically meaningful degradation stages from raw cycling data without supervision, achieves strong quantitative clustering quality (Silhouette = 0.6063, DBI = 0.5799, CHS = 255.64), and produces health state boundaries that are consistent with established electrochemical understanding of battery aging.

The finding that Voltage Slope correlates with SoH at $r = -0.911$ establishes a new, computationally inexpensive degradation indicator that can be computed from standard BMS measurements without specialized equipment. The 72.5% reduction in autoencoder reconstruction loss confirms that the nonlinear degradation manifold of the a large-scale lithium-ion battery cycling dataset cell can be efficiently compressed into a 2-dimensional latent space that retains all essential health state information.

Together, these contributions advance the mathematical and data-driven foundations of battery health management, bridging the gap between electrochemical domain knowledge and modern unsupervised deep learning techniques. The proposed pipeline provides a template for scalable, label-free, and physically interpretable battery health state identification that can be extended to multiple chemistries, online deployment contexts, and integration with predictive maintenance and second-life battery screening workflows in the rapidly expanding global battery industry.

As electric vehicles, grid-scale storage, and portable electronics continue to drive unprecedented demand for lithium-ion batteries, the need for intelligent, data-driven health monitoring systems will only intensify. This work contributes a mathematically rigorous and experimentally validated step toward that goal.

Bibliography

[1] A. Masias, J. Marcicki, and W. A. Paxton, "Opportunities and challenges of lithium ion batteries in automotive applications," *ACS Energy Letters*, vol. 6, no. 2, pp. 621–630, 2021.

[2] Y. Li, K. Liu, A. M. Foley, A. Zülke, M. Berecibar, E. Nanini-Maury, J. Van Mierlo, and H. E. Hoster, "Data-driven health estimation and lifetime prediction of lithium-ion batteries," *Renewable and Sustainable Energy Reviews*, vol. 113, p. 109254, 2019.

[3] T. Kim and W. Qiao, "A hybrid battery model capable of capturing dynamic circuit characteristics and nonlinear capacity effects," *IEEE Transactions on Energy Conversion*, vol. 26, no. 4, pp. 1172–1180, 2011.

[4] 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*, vol. 56, pp. 572–587, 2016.

[5] X. Han, M. Ouyang, L. Lu, J. Li, Y. Zheng, and Z. Li, "A comparative study of commercial lithium ion battery cycle life in electrical vehicle," *Journal of Power Sources*, vol. 251, pp. 38–54, 2014.

[6] B. Saha and K. Goebel, "Modeling Li-ion battery capacity depletion in a particle filter framework," in *Proc. Annual Conference of the Prognostics and Health Management Society*, 2009, pp. 1–10.

[7] J. Vetter et al., "Ageing mechanisms in lithium-ion batteries," *Journal of Power Sources*, vol. 147, no. 1–2, pp. 269–281, 2005.

[8] C. Hendricks et al., "A failure modes, mechanisms, and effects analysis (FMMEA) of lithium-ion batteries," *Journal of Power Sources*, vol. 297, pp. 113–120, 2015.

[9] R. Dua, V. Acharya, and V. N. Bhondekar, "Lithium-ion battery state of health estimation using Gaussian process regression," in *Proc. IEEE International Conference on Computing*, 2021, pp. 1–6.

[10] R. Xiong, J. Cao, Q. Yu, H. He, and F. Sun, "Critical review on the battery state of charge estimation methods for electric vehicles," *IEEE Access*, vol. 6, pp. 1832–1843, 2018.

[11] K. A. Severson et al., "Data-driven prediction of battery cycle life before capacity degradation," *Nature Energy*, vol. 4, pp. 383–391, 2019.

[12] M. A. Hannan, M. S. H. Lipu, A. Hussain, and A. Mohamed, "A review of lithium-ion battery state of charge estimation and management system in electric vehicle

applications," *Renewable and Sustainable Energy Reviews*, vol. 78, pp. 834–854, 2017.

[13] Y. Deng, C. Zhao, F. Tian, C. Tan, and X. Qiao, "Online available capacity prediction and state of charge estimation based on advanced data-driven algorithms for lithium iron phosphate battery," *Energy*, vol. 112, pp. 469–480, 2016.

[14] X. Hu, C. Zou, C. Zhang, and Y. Li, "Technological developments in batteries," *IEEE Power and Energy Magazine*, vol. 15, no. 5, pp. 20–31, 2017.

[15] W. Waag, C. Fleischer, and D. U. Sauer, "Critical review of the methods for monitoring of lithium-ion batteries in electric and hybrid vehicles," *Journal of Power Sources*, vol. 258, pp. 321–339, 2014.

[16] C. Hu, B. D. Youn, and J. Chung, "A multiscale framework with extended Kalman filter for lithium-ion battery SOC and capacity estimation," *Applied Energy*, vol. 92, pp. 694–704, 2012.

[17] D. Bhatt, B. Patel, and P. Kaur, "A survey on machine learning for state of health estimation of lithium-ion batteries," *Journal of Energy Storage*, vol. 55, p. 105690, 2022.

[18] Y. Zhang, R. Xiong, H. He, and M. G. Pecht, "Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 5695–5705, 2018.

[19] C. Chen, J. Xiong, and W. Shen, "A lithium-ion battery-in-the-loop approach to test and validate multiscale dual H infinity filters for state-of-charge and capacity estimation," *IEEE Transactions on Power Electronics*, vol. 33, no. 1, pp. 332–342, 2018.

[20] S. Shen, M. Sadoughi, M. Li, Z. Wang, and C. Hu, "Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries," *Applied Energy*, vol. 260, p. 114296, 2020.

[21] Y. Liu, J. He, H. Liu, S. Yang, M. Liu, Q. Meng, and Q. Liu, "Capacity estimation of lithium-ion batteries based on multiple regression with multivariate feature extraction," *Energy Reports*, vol. 8, pp. 857–866, 2022.

[22] W. He, N. Williard, C. Chen, and M. Pecht, "State of charge estimation for Li-ion batteries using neural network modeling and unscented Kalman filter-based error cancellation," *International Journal of Electrical Power & Energy Systems*, vol. 62, pp. 783–791, 2014.

[23] T. Jiang, L. Chen, X. Wang, and C. Li, "State of health estimation for lithium-ion batteries using attention mechanism and improved loss function," *Energies*, vol. 16, no. 1, p. 455, 2023.

[24] U. Troltzsch, O. Kanoun, and H.-R. Tränkler, "Characterizing aging effects of lithium ion batteries by impedance spectroscopy," *Electrochimica Acta*, vol. 51, no. 8-9, pp. 1664–1672, 2006.

[25] J. Huang, Z. Li, and B. Y. Liaw, "Towards robust estimation of the state of health of lithium-ion batteries," *Journal of The Electrochemical Society*, vol. 163, no. 3, pp. A506–A512, 2016.

[26] S. Mumtaz, A. Gani, I. A. Hameed, and C.-Y. Lin, "Electrochemical impedance spectroscopy for lithium-ion battery aging consideration in second-life applications," *IEEE Access*, vol. 10, pp. 6093–6106, 2022.

[27] C. Hu, G. Jain, P. Tamirisa, and T. Gorka, "Method for estimating capacity and predicting remaining useful life of lithium-ion battery," in *Proc. IEEE Symposium on Computational Intelligence and Data Mining*, 2014, pp. 1–8.

[28] R. Richardson, M. Ireland, and A. Bhatt, "Battery health prediction under generalized conditions using a Gaussian process transition model," *Journal of Energy Storage*, vol. 23, pp. 320–328, 2019.

[29] M. Lucu, E. Martinez-Laserna, I. Gandiaga, and H. Camblong, "A critical review on self-adaptive Li-ion battery ageing models," *Journal of Power Sources*, vol. 401, pp. 85–101, 2018.

[30] A. Nuhic, T. Terzimehic, T. Soczka-Guth, M. Buchholz, and K. Dietmayer, "Health diagnosis and remaining useful life prognostics of lithium-ion batteries using data-driven methods," *Journal of Power Sources*, vol. 239, pp. 680–688, 2013.

[31] Y. Xing, E. W. M. Ma, K. L. Tsui, and M. Pecht, "An ensemble model for predicting the remaining useful performance of lithium-ion batteries," *Microelectronics Reliability*, vol. 53, no. 6, pp. 811–820, 2013.

[32] Z. Chen, M. Wu, R. Cao, and B. Liu, "A multi-scale state of health prediction framework of lithium-ion batteries considering the temperature variation during discharge," *Journal of Power Sources*, vol. 467, p. 228357, 2020.

[33] W. Zhang, X. Li, and X. Li, "Deep learning-based prognostic approach for lithium-ion batteries with adaptive time-series prediction and on-line validation," *Measurement*, vol. 164, p. 108052, 2020.

[34] H. Lim, W. An, W. Jung, S. Lee, and K. Shin, "A neural network-based degradation model for lithium-ion batteries," *IEEE Access*, vol. 10, pp. 87, 2022.

[35] X. Feng, M. Ouyang, X. Liu, L. Lu, Y. Xia, and X. He, "Thermal runaway mechanism of lithium ion battery for electric vehicles: a review," *Energy Storage Materials*, vol. 10, pp. 246–267, 2018.

[36] G. E. Blomgren, "The development and future of lithium ion batteries," *Journal of The Electrochemical Society*, vol. 164, no. 1, pp. A5019–A5025, 2017.

[37] R. Dell and D. A. J. Rand, *Understanding Batteries*. Cambridge, UK: Royal Society of Chemistry, 2001.

[38] Z. Li, J. Xu, and X. Ai, "Multi-feature fusion for lithium-ion battery state-of-health estimation with deep neural network embedding," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 8, pp. 8137–8147, 2023.

[39] H. Park, J. Kim, and S. Lee, "Cluster-specific LSTM models for heterogeneous battery degradation prediction," *Journal of Energy Storage*, vol. 48, p. 103979, 2022.

[40] S. Guo, L. Ma, and J. Jiang, "ICA-based degradation feature extraction with BiLSTM and AdaBoost for lithium-ion battery state of health estimation," *Applied Energy*, vol. 330, p. 120336, 2023.

[41] Q. Liu, Y. Ye, L. Dong, and X. Chen, "Voltage relaxation based latent feature learning for battery aging characterization," *IEEE Transactions on Transportation Electrification*, vol. 9, no. 2, pp. 2444–2455, 2023.

[42] T. Li, F. Fang, and W. Li, "Probabilistic capacity prediction for lithium-ion batteries using deep Gaussian processes and LSTM," *Energy*, vol. 261, p. 125283, 2022.

[43] J. Wu, C. Zhang, and Z. Chen, "An online method for lithium-ion battery remaining useful life estimation using importance sampling and neural networks," *Applied Energy*, vol. 173, pp. 134–140, 2016.

[44] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. International Conference on Machine Learning (ICML)*, vol. 48, pp. 478–487, 2016.

[45] F. von Bülow and T. Meisen, "A review on methods for state of health forecasting of lithium-ion batteries applicable in real-world operational conditions," *Journal of Energy Storage*, vol. 57, p. 106559, 2023.

[46] D. Andrea, *Battery Technology Life Verification Test Manual*. Norwood, MA: Artech House, 2020.

[47] R. Dell and D. A. J. Rand, *Understanding Batteries*. Cambridge, UK: Royal Society of Chemistry, 2001.

[48] C. R. Birkl, M. R. Roberts, E. McTurk, P. G. Bruce, and D. A. Howey, "Degradation diagnostics for lithium ion cells," *Journal of Power Sources*, vol. 341, pp. 373–386, 2017.

[49] R. Borah, F. R. Hughson, J. Johnston, and T. Nyokong, "On battery materials and

methods," *Materials Today Advances*, vol. 6, p. 100046, 2020.

[50] R. Xiong, J. Tian, W. Shen, and F. Sun, "A novel fractional order model for state estimation of lithium-ion battery using a dedicated adaptive extended Kalman filter," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 7, pp. 7232–7244, 2020.

[51] Z. Chen, W. Liu, D. Zhou, T. Xia, and E. Pan, "Inconsistency identification for lithium-ion battery energy storage systems using deep embedded clustering," *Applied Energy*, vol. 385, p. 125532, 2025.

[52] S. Pan et al., "Automatically constructing a health indicator for lithium-ion battery state-of-health estimation via adversarial and compound stacked autoencoder," *Journal of Energy Storage*, vol. 84, p. 110952, 2024.

[53] Z. Han et al., "Predicting EV battery state of health using long short-term degradation feature extraction and FEA TimeMixer," *Scientific Reports*, vol. 15, p. 1823, 2025.