

Article

State of Charge Estimation for Both Electric Bus and Passenger Vehicles with Different Battery Types Using Multi-Instance Learning

Ibrahim Atakan Kubilay and Derya Birant *

Department of Computer Engineering, Dokuz Eylul University, Izmir 35390, Turkey; atakan@cs.deu.edu.tr

* Correspondence: derya@cs.deu.edu.tr

Abstract

State of charge (SoC) estimation for an electric vehicle (EV) is critical for range estimation, preventing overcharging or undercharging, energy management, and battery lifespan. The current studies are typically based on single-instance learning, focus on a specific battery chemistry or a single vehicle type, rely on controlled Lab conditions, and often lack explainability mechanisms. To overcome all these limitations, this paper proposes a more practically applicable and explainable SoC estimation framework that jointly addresses vehicle diversity (bus and passenger EVs), battery chemistry variability (Nickel Cobalt Manganese and Lithium Iron Phosphate), collective battery dynamics, and real-world on-road operational uncertainties within a unified learning architecture. This study introduces MIL-LGBM, a specialized method that integrates Multiple Instance Learning with Light Gradient Boosting Machine to effectively capture the complex nonlinear behavior of battery SoC while maintaining a low computational footprint. An experimental study conducted on a real-world dataset demonstrated that the proposed framework achieved a 0.799 MAE for passenger vehicles across a wide range of on-road driving conditions.

Keywords: machine learning; state of charge estimation; electric vehicles; regression; multi-instance learning

1. Introduction

State of Charge (SoC) estimation is defined as the process of determining the ratio of a battery's remaining usable capacity to its total or nominal capacity, essentially serving as a high-precision fuel gauge for electric vehicles (EVs) [1]. This estimation is fundamentally important for safety, as it enables the battery management system (BMS) to mitigate the risks of overcharging or over-discharging, which can cause thermal runaway, electrolyte decomposition, or catastrophic battery failure [2]. Beyond safety, accurate SoC estimation is the cornerstone of effective energy management and performance optimization, underpinning critical vehicle functions such as range prediction, torque and speed control, and the management of regenerative braking limits [3]. Furthermore, maintaining precise SoC monitoring is necessary to extend battery lifespan and enhance consumer trust by providing stable and reliable information that reduces range anxiety [1,4]. Because SoC is an internal state that cannot be directly measured by physical sensors, developing robust algorithms that can infer it from measurable quantities like voltage, current,

Academic Editor: Sijia Yang

Received: 4 June 2026

Revised: 16 July 2026

Accepted: 20 July 2026

Published: 21 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

and temperature is a critical problem for the long-term, efficient operation of EV propulsion systems [5].

Robust SoC estimation is inherently difficult due to the complex non-linear electrochemical properties of batteries and phenomena like hysteresis and voltage relaxation, which traditional linear models often fail to capture accurately [4]. Furthermore, estimation accuracy is severely compromised by environmental and operational fluctuations, particularly temperature variations, varying discharge rates, and sensor measurement noise [5]. As batteries age, degradation of State of Health (SoH) causes shifts in nominal capacity and internal resistance, leading to model drift or divergence if parameters are not updated in real time [3].

The current studies are typically based on single-instance learning, focus on a single vehicle type or a specific battery chemistry, rely on controlled experimental Lab conditions, and often lack explainability mechanisms. The present study adopts a more comprehensive and realistic approach to SoC estimation by addressing all these issues together. Additionally, many existing algorithms depend on large-cell simplifications that overlook cell-to-cell inconsistencies within a battery pack [6]. While advanced deep learning models have been investigated, they frequently impose a high computational burden and memory requirements that may exceed the capabilities of standard on-board vehicle hardware, while simultaneously lacking the interpretability of physics-based models [7].

The primary contributions of this work are outlined below:

- i. Unlike conventional single-instance learning (SIL)-based SoC estimation studies, this work represents the first attempt that introduces a Multi-Instance Learning (MIL)-based SoC estimation framework, enabling the model to learn collective EV battery behavioral patterns rather than relying solely on isolated instantaneous measurements.
- ii. This study introduces MIL-LGBM, an innovative methodology designed to process aggregated bag representations through a gradient boosting approach, ensuring both high predictive precision and real-time operational efficiency.
- iii. Whereas conventional research is typically restricted to a single vehicle category, the proposed framework performs a comprehensive cross-domain analysis involving both electric buses and electric passenger vehicles, allowing a broader evaluation across heterogeneous driving characteristics.
- iv. In contrast to existing approaches that are generally limited to single battery chemistry, this study investigates both nickel-cobalt-manganese (NCM) and lithium-iron-phosphate (LFP) battery technologies, thereby improving the generalizability of the proposed methodology across different electrochemical behaviors and battery dynamics.
- v. Unlike studies relying on laboratory-generated or simulation-based datasets, the proposed framework is developed and validated using real-world driving data, enabling the model to capture realistic on-road conditions.
- vi. This study further stands out by integrating Explainable Artificial Intelligence (XAI) techniques into the SoC estimation framework, enhancing model transparency and interpretability beyond conventional black-box predictions.

Consequently, the proposed methodology distinguishes itself by proposing a more practically applicable, explainable, and robust SoC estimation framework that jointly addresses vehicle diversity (bus and passenger EVs), battery chemistry variability (NCM and LFP), collective battery dynamics (MIL), and real-world operational uncertainties within a unified learning architecture.

The remainder of this article is structurally mapped as follows. Section 2 provides a meticulous review of existing studies on battery SoC estimation, highlighting the limitations of traditional methodologies and identifying the scientific gaps that this work aims

to bridge. Section 3 outlines the theoretical foundations and structural design of the proposed MIL-LGBM method, providing an in-depth explanation of how multi-instance learning principles are integrated with the gradient boosting paradigm. Section 4 details the data preprocessing stages, experimental setups, and comprehensive validation scenarios conducted on real-world driving datasets to evaluate the accuracy and robustness of the developed paradigm that effectively handles diverse battery types across both electric buses and passenger vehicles. Section 5 synthesizes the core findings and delineates prospective directions for future investigations.

2. Related Work

Battery SoC estimation has been addressed in the literature due to its importance for ensuring safety and extending the life cycle of electric vehicle energy storage systems. Various studies have examined this problem from different perspectives, ranging from the development of traditional physical models to the deployment of advanced machine learning architectures. The aim of each is to address the nonlinear and complex behavior of battery cells under diverse operating conditions. Previous studies have explored different approaches under varying assumptions, datasets, and application scenarios. They differ in terms of vehicle type, battery chemistry, feature design, and evaluation methodology. Table 1 provides a comprehensive overview of the existing literature [8–29], summarizing recent advancements by their methods, applicability to electric buses or passenger vehicles, dataset type (Lab or on-road), battery characteristics, performance metrics, feature sets, integration of explainability, and reported limitations.

In the literature, Single-Instance Learning (SIL) has been adopted for SoC estimation [8–29], in which each data instance is directly associated with a single SoC label. However, battery-related sensor signals, such as voltage, temperature, and current, typically exhibit frequency fluctuations and transient variations, whereas SoC represents a slowly varying internal battery state. Consequently, learning from isolated instances may cause the model to overemphasize instantaneous operating conditions, which can sometimes be misleading and may reduce generalization capability under dynamic driving environments. In contrast, Multi-Instance Learning (MIL) can provide an alternative perspective by grouping sequential instances into bags that represent short-term system behavior rather than isolated measurements. This enables the model to capture physically meaningful battery dynamics more effectively. By learning from aggregated behavioral patterns, MIL-based approaches can reduce overfitting, improve robustness against noisy fluctuations, and provide a more realistic approximation of battery dynamics during real-world operation.

A significant portion of current research [8–19,24,25] in the field focuses on passenger electric vehicles, where data availability and standardized testing conditions are relatively more accessible. In contrast, a more limited body of work [20–23,26–29] has investigated SoC estimation specifically for electric buses, despite their increasing adoption and distinct operational characteristics, such as heavy-duty load requirements and more dynamic cycles. Although these two domains have been explored independently, there is a notable lack of comprehensive studies that evaluate both vehicle types within a unified framework. In this context, the present study distinguishes itself by offering a cross-platform analysis that encompasses both vehicle types, thereby providing a more comprehensive evaluation.

Another important aspect in the literature is the type of dataset used for model development and evaluation. State-of-the-art studies heavily rely on laboratory-collected data or high-fidelity simulated battery models under controlled conditions [8–17,19,23–25,27]. While such settings enable systematic analysis and reproducibility, they may fail to fully reflect the multifaceted complexity and variability of on-road driving conditions, such as erratic regenerative braking and varying road gradients. These idealized datasets generally omit operational noise, data transmission inconsistencies, complex aging pat-

terns, and anomalies inherent in on-road vehicle operations, which can lead to accuracy degradation when models are deployed in the field. Furthermore, lab data often ignores the dynamic interactions between the battery and other vehicle components, such as the motor and powertrain, as well as external environmental factors like aerodynamic drag and vibration. In contrast, on-road data inherently includes dynamic driving behaviors, environmental variations, and measurement uncertainties, which can substantially impact battery performance and SoC estimation accuracy. Field data also presents challenges, including incomplete discharge–charge cycles, limited sampling frequency, and high data heterogeneity. The key difference between laboratory and real-life usage data lies in user variability, as driving habits and behaviors vary among EV drivers. Even for identical vehicle types, the battery charge patterns can vary considerably across users. In this context, the use of real-world on-road data in the present study enhances the practical relevance of the proposed approach. Due to its closer representation of actual application conditions, our study provides a more realistic assessment.

Current research on battery state estimation is often limited by a narrow focus on a single battery chemistry, commonly either Nickel Cobalt Manganese (NCM, NiCoMn) [8,10,11,13,14,16–18,22,25] or Lithium Iron Phosphate (LFP, LiFePO₄) [20,29]. However, these materials exhibit distinct electrochemical characteristics and internal dynamics, including differences in voltage profiles, degradation patterns, thermal behavior, and sensitivity to operating conditions. Such variations can influence the generalizability and performance of battery SoC estimation models. In this context, our research offers a distinct advantage by analyzing both NCM and LFP battery types, enabling a broader evaluation of model performance across diverse battery technologies.

The literature has explored various methodologies to capture the nonlinear and complex dynamics of battery cells under varying load and environmental conditions. SoC estimation methods can generally be categorized into four primary groups based on their underlying principles. The first group consists of Conventional (Direct) methods [29], which rely on straightforward measurable physical parameters, such as terminal voltage and current, and their empirical relationships. The second group includes Model-Based approaches, further subdivided into Filter-based (e.g., Kalman Filter variants) [13–17,24,27,28] and Observer-based (e.g., sliding mode observers) approaches [13] that utilize mathematical representations of the battery, such as equivalent circuit models. The third group comprises Data-Driven methods [8–11,18–22], where machine learning, deep learning, and transformer techniques are employed to learn complex nonlinear mappings between historical sensor data and battery SoC. Finally, Hybrid approaches combine these paradigms, such as integrating physical filters with neural networks, aiming to synergize their strengths. In the following paragraphs, studies in each of these four categories are discussed in more detail.

Conventional (direct) methods have been explored in various studies for SoC estimation due to their straightforward implementation and low computational complexity. Common examples include open-circuit voltage (OCV) lookup tables and Coulomb counting. For instance, Srirattanawichaikul et al. [29] employed the discrete Coulomb counting method for a battery pack-level management system. Their study was validated using an electric bus operating within a real-world campus transit system for over a year. While conventional techniques offer simplicity and ease of implementation, they have inherent limitations, such as the requirement for an accurate initial SoC, their inability to account for battery aging, and their sensitivity to measurement noise and sensor errors, which can restrict their long-term accuracy under varying operating conditions.

Table 1. Summary and comparison of recent studies on battery SoC estimation with key differences.

Ref.	Year	Description	Methods	Dataset (Lab/ On-Road)	SIL or MIL	Bus?	Private Vehicle?	Bat- tery Type	Perfor- mance Metrics	Features	XAI?	Limitation
[8]	2026	Embedded de- ployment	NN, CNN	Lab	SIL	No	Yes	NCM	RMSE, MAE, MSE	V, I, T (3)	No	Precision loss
[9]	2026	Parallel archi- tecture	BiLSTM, TCN,	Lab	SIL	No	Yes	NCA	RMSE, MAE, MaxE, R ²	V, I, T (3)	No	Standalone model with complex dy- namics
[10]	2026	Low temp. en- vironments	CNN-BiLSTM- AM	Lab	SIL	No	Yes	NCM	RMSE, MAE, MSE	V, I, T (3)	No	Hybrid model complexity
[11]	2026	Spatiotemporal features	TCN-Trans- former	Lab	SIL	No	Yes	NCM	RMSE, MAE, R ²	V, I, T (3)	No	TCN-related limitations
[12]	2026	SoC and SoH estimation	μ-Analysis	Simula- tion	SIL	No	Yes	LMO	RMSE	V, I, Speed (3)	No	Model de- pendency
[13]	2026	Equivalent cir- cuit model	SMO-EKF	Lab	SIL	No	Yes	NCM	RMSE, MAE	V, I, oth- ers (+7)	No	Parameter identification load
[14]	2026	Online param- eter identifica- tion	QPSO-AUKF	Lab	SIL	No	Yes	NCM	RMSE, MAE	V, I (2)	No	Linearization errors
[15]	2026	Correntropy criterion	GMMCC-AEKF	Lab	SIL	No	Yes	LMO	RMSE, MAE	V, I, T, Noise (4+)	No	Complex noise distri- butions
[16]	2026	An online SoC estimation	AEKF-SVM	Lab	SIL	No	Yes	NCM	RMSE, MAE, MaxE	V, I (2)	No	Matrix calcu- lation load, long training time
[17]	2026	Digital twin technology	KF-LSTM	Lab	SIL	No	Yes	NCM	RMSE, MAE, R ²	V, I, T, dQ/dV (4)	No	Hybrid model complexity
[18]	2026	Online estima- tion	MLP, NARX, CNN	Lab/On- Road	SIL	No	Robot	NCM	RMSE, MAE	V, I, T (3)	No	Limited accu- racy
[19]	2026	Regression ver- sus classifica- tion	RF, ET	Lab	SIL	No	Yes	NMA	Accuracy, MAE	DRT peaks (5+)	No	Dependency on expert knowledge
[20]	2026	GPS-assisted prediction	LSTM, GRU, Transformer	On-Road	SIL	Yes	No	LPF	RMSE, MAE	V, I, T, GPS (32)	No	Accuracy can drop on new/unseen routes
[21]	2026	Feature selec- tion method	SVM, DT, LR	On-Road	SIL	Yes	No	Li-ion	RMSE, R ² , R	Varia- bles (36)	No	Data discreti- zation leads to info loss
[22]	2025	Using only kin- ematic data	LR, DT, SVM, NN	On-Road	SIL	Yes	No	NCM	RMSE	Speed, Accel, T, Wind (4+)	No	Lack of elec- trical battery data
[23]	2025	Triple forget- ting factor	TFF-AEKF	Simula- tion	SIL	Yes	No	Li-ion	RMSE, MAE, MaxE	V, I, T (3)	No	High compu- tational load
[24]	2025	ML versus Kal- man	FNN, NARX, LSTM, EKF, UKF	Lab	SIL	No	Yes	NCA	RMSE	V, I, T, Payload (4)	No	High memory usage

[25]	2024	ML vs. DL comparison	LR, SVM, RF, XGBoost, ET, KNN, NN, CNN, LSTM	Simulation (Tesla S)	SIL	No	Yes	NCM	MAE, MSE, MaxE, R ²	Vehicle variables (9)	No	Synthetic data
[26]	2023	Energy consumption estimation	LSTM	On-Road	SIL	Yes	No	Li-ion	RMSE, MAE	Speed, accel, motor, etc. (11)	No	Overfitting in very long training cycles
[27]	2023	Equivalent circuit modeling	FOAUKF, FH-ECM	Lab	SIL	Yes	No	NCM	RMSE	V, I (2)	No	High complexity of fractional calculations
[28]	2023	First-order equivalent circuit.	DAKF	On-Road	SIL	Yes	No	Li-ion	Error	V, I, T (3)	No	Poorly defined parameters ruin status
[29]	2020	Campus transit system	Coulomb counting	On-Road	SIL	Yes	No	LFP	Error	V, I, T (3)	No	A math model without any learning
Proposed		Vehicle types, battery types, MIL-based, and explainable	MIL-LGBM	On-Road	MIL	Yes	Yes	NCM, LFP	RMSE, MAE, R ²	V, I, T, others (8)	Yes	User interface-driven accessibility

Model-based methods represent another category for SoC estimation methodology in the literature, primarily relying on mathematical representations, typically through equivalent circuit models to simulate internal battery dynamics. Within this category, filter-based studies have widely employed Kalman Filter (KF) variants, such as Extended KF (EKF) [16,27], Unscented KF (UKF) [27], Adaptive UKF [B6], and Adaptive EKF [20, 24, B4]. Additionally, observer-based approaches, such as sliding mode observers (SMO), have also been explored [16] to enhance robustness against system disturbances. However, model-based methods generally suffer from high computational overhead due to complex matrix calculations, strong reliance on accurate model formulation and parameter identification, and sensitivity to linearization errors in nonlinear environments, which can limit their robustness under varying real-world driving conditions.

Data-driven methods have also been investigated as a powerful alternative to mathematical and physical models by directly learning the nonlinear complex mapping between sensor data and charge states. Within this category, traditional machine learning techniques have been applied, including neural networks (NN) [8,18,22], linear regression (LR) [21,25], decision trees (DT) [21,22], support vector machine (SVM) [16,21,22,25], k-nearest neighbors (KNN) [25], and random forest (RF) [19,25]. Additionally, deep learning (DL) methods have emerged, including Long Short-Term Memory (LSTM) [20,24–26], Bidirectional LSTM (BiLSTM) [9,10], Convolutional Neural Networks (CNN) [8,10,18,25], and Gated Recurrent Units (GRU) [20], due to their ability to model temporal dependencies. More recently, transformer-based architectures have also been explored for SoC estimation tasks [11,20]. However, despite their high predictive accuracy, DL and Transformer models face critical drawbacks: heavy reliance on high-quality, massive labeled data, substantial computational and memory requirements, and limited interpretability due to their black-box nature.

Hybrid approaches aim to integrate the strengths of multiple paradigms. For instance, Li et al. [24] proposed a hybrid approach that combines an adaptive EKF with an SVM battery model to provide online estimation. Similarly, Zhao et al. [25] introduced a hybrid Kalman Filter–Long Short-Term Memory (KF–LSTM) model within a synchronized digital twin

framework, enabling collaborative estimation of SoC and State of Health (SoH). However, hybrid methods face critical challenges, primarily including increased system complexity, the need for careful tuning of multiple components, higher computational cost, and reduced interpretability compared to single-paradigm approaches. These may limit practical deployment in safety-critical or resource-constrained environments.

The selection of input features plays a significant role in the accuracy of ML model estimates. A large majority of studies in the literature [8–11,15,17,18,23,28,29] primarily rely on fundamental electrical and thermal measurements, specifically voltage (V), current (I), and temperature (T). These signals are directly related to battery behavior and are readily available from battery management systems. Studies using real-world on-road data [20–22] often expand the feature space to include vehicle-related characteristics, such as speed, acceleration, and motor power, to capture interactions between battery dynamics and operational conditions. Some predictive frameworks have also incorporated specialized contextual features, such as GPS-derived location [20] and environmental factors, such as wind speed [22].

Another important aspect is the selection of performance metrics to quantify accuracy and robustness. A vast majority of research employed Mean Absolute Error (MAE) [8–11,13–20] and Root Mean Square Error (RMSE) [8–24,26,27] as the primary indicators to assess sensitivity to large deviations and average prediction performance. To further evaluate model performance, some studies also integrated the Mean Squared Error (MSE) [10,25] and Maximum Absolute Error (MaxE) [9,16,24,25] into their evaluation frameworks. While less common, the Coefficient of Determination (R^2) [9,11,17,21,25] was reported to evaluate the model's explanatory power. A limited number of studies utilized alternative metrics, such as the Correlation Coefficient (R) [21], to evaluate the strength of the relationship between predicted and actual SoC values. Furthermore, in cases where both regression and classification perspectives were considered, studies reported a combination of metrics, such as MAE alongside classification accuracy [19], to validate model performance from multiple viewpoints, covering both continuous and discrete charge-state levels.

Unlike previous studies, which typically are based on single-instance learning, focus on a single vehicle type or a specific battery chemistry, rely on controlled experimental Lab conditions, and often lack explainability mechanisms, the present study adopts a more comprehensive and realistic approach to SoC estimation. Specifically, the proposed methodology incorporates a Multi-Instance Learning (MIL)-based representation, enabling the model to learn collective battery behavior trends rather than relying solely on isolated instantaneous measurements. It introduces a methodology, MIL-LGBM, that leverages gradient boosting in a multi-instance setting to enhance the accuracy of EV battery SoC estimation. Moreover, this work distinguishes itself by offering a cross-platform analysis of both electric buses and electric passenger vehicles, enabling a broader evaluation across different usage profiles. In addition, another pivotal contribution of this work is the analysis of multiple battery chemistries (NCM and LFP), addressing the variability in electrochemical characteristics that are often overlooked in prior research. Furthermore, rather than relying on idealized laboratory or simulated datasets, our methodology is built upon on-road, real-life usage data, reflecting the inherent characteristics of actual driving cycles, regenerative braking, and environmental disturbances. This study is also distinctive in that it incorporates XAI techniques, enhancing model transparency and interpretability by moving beyond “black-box” predictions. Through this combination of diverse vehicle types, battery materials, real-world data, and explainability, the proposed approach provides a more robust and practically relevant solution compared to existing studies.

3. Materials and Methods

3.1. Proposed Methodology

This paper proposes a SoC estimation framework designed to simultaneously accommodate vehicle diversity (electric buses and passenger cars), distinct cell chemistries (NCM and LFP), collective battery characteristics via MIL, and the uncertain nature of real-world driving conditions within a single unified learning architecture. It introduces MIL-LGBM, a methodology that harmonizes multiple-instance learning with the gradient-boosting paradigm to accurately model the intricate, nonlinear dynamics of battery state-of-charge without imposing heavy computational demands on processing resources.

The entire process of the proposed methodology is systematically outlined in Figure 1. The raw data that comes from individual vehicles is combined into distinct Bus and Passenger Vehicle datasets. As part of the feature engineering phase, three features (acceleration, power, and temperature difference) are introduced to enhance the predictive capacity of the model. The datasets are subsequently subjected to rigorous data preprocessing and feature selection procedures to improve data quality and eliminate redundant variables. Following this stage, a Multi-Instance Data Transformation is executed to restructure the data into bags. After that, the transformed dataset is then split in the usual manner into training, validation, and test subsets. Model training is carried out directly on this generated MIL dataset. Based on the initial performance outcomes, hyperparameter tuning is implemented for improvement. The optimized model is comprehensively evaluated on the dedicated test set to compute the corresponding evaluation metrics. Subsequently, the finalized model is deployed on entirely unseen operational data to successfully generate the ultimate SoC predictions. Decision-making processes, such as battery health monitoring, vehicle routing, or energy management, can finally be executed with high precision based on these outputs.

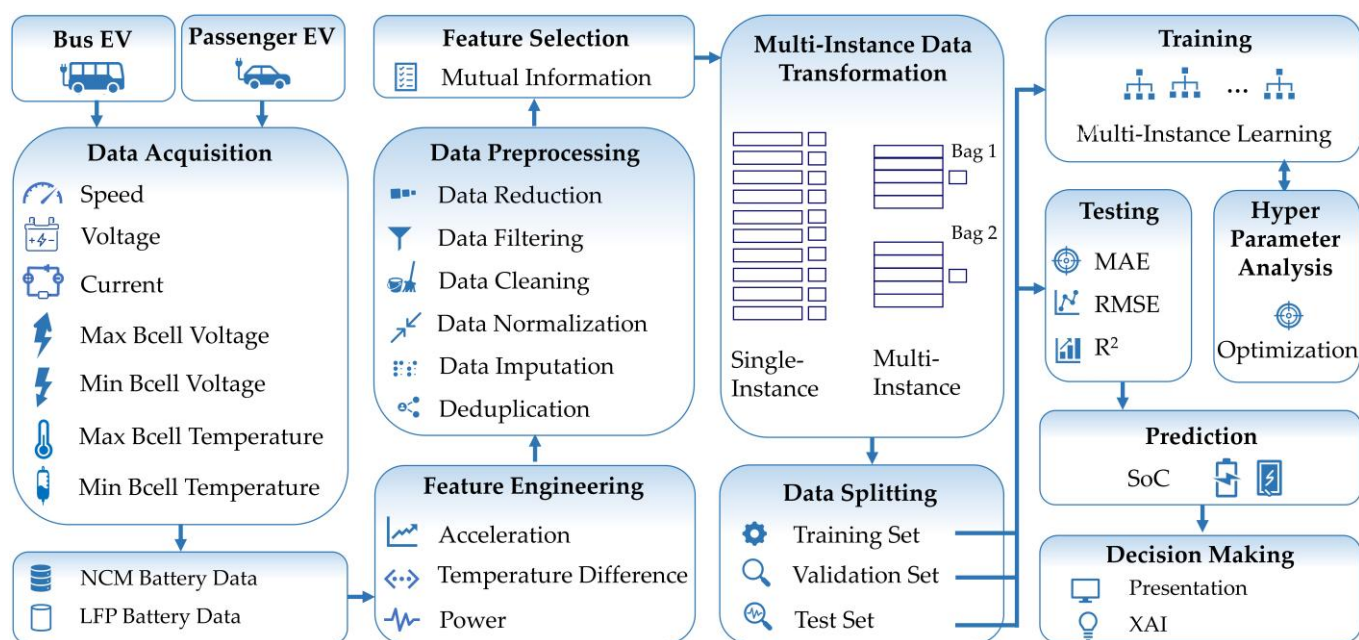


Figure 1. Overall workflow of the proposed MIL-LGBM framework for EV battery SoC estimation, including data acquisition, data preprocessing, feature engineering and selection, multi-instance data generation, MIL-based learning, testing, prediction, and decision-making with XAI analysis.

The target variable, SoC, represents the ratio of remaining energy within the battery relative to its reference capacity. This relationship is defined as:

$$\text{SoC}(t) = \frac{Q(t)}{Q_n} \quad (1)$$

where $Q(t)$ represents the instantaneous capacity of the battery, and Q_n refers to the nominal capacity provided by the manufacturer, reflecting the peak charge amount that the battery can hold.

3.2. Proposed Multi-Instance Learning Paradigm

Multi-Instance Learning (MIL) is a theoretical paradigm in which data are organized into sets of instances, referred to as bags, rather than isolated individual samples. Unlike conventional Single-Instance Learning (SIL), where each instance is independently associated with a target value, MIL enables the model to learn from grouped observations that collectively represent underlying system behavior.

MIL is particularly suitable for battery SoC estimation since battery-related sensor measurements, such as voltage, current, and temperature, often exhibit rapid fluctuations, whereas SoC evolves more gradually as an internal system state. Therefore, constructing bags from sequential observations allows the model to learn more representative battery behavior patterns, thereby improving robustness and generalization while reducing sensitivity to measurement noise.

Let the dataset be represented as:

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \quad (2)$$

where $x_i \in \mathbb{R}^d$ denotes the feature vector of the i -th instance and $y_i \in \mathbb{R}$ refers to the corresponding SoC value. Each instance is defined in the feature space:

$$x_i = \{x_{i1}, x_{i2}, \dots, x_{id}\} \quad (3)$$

where d denotes the number of features and x_{ij} is the j -th feature value in the i -th bag.

In the proposed MIL formulation, consecutive instances sharing the same SoC value are grouped into bags. Let w represent the maximum bag size. The number of instances within a bag may vary depending on the continuity of the SoC sequence. For example, if 18 sequential instances belong to the same SoC value and $w = 5$, the generated bag sizes become {5,5,5,3}.

The constructed multi-instance dataset is represented as:

$$MD = \{(B_1, Y_1), (B_2, Y_2), \dots, (B_M, Y_M)\} \quad (4)$$

where B_j is called a *bag*, Y_j denotes the SoC label associated with bag B_j , and M represents the total number of bags, which is approximately defined as $M \approx N/w$. Each bag B_j consists of sequential instances:

$$B_j = \{x_1, x_2, \dots, x_c\}, \quad c \leq w \quad (5)$$

Essentially, a standard (single instance) learner is a special case in which every bag consists of only one instance $B_i = \{x_{i1}\}$. Since all instances within the same bag share an identical SoC value, the bag label is directly assigned as:

$$Y_j = y_i, \quad \forall x_i \in B_j \quad (6)$$

For each bag, a bag-level representation is obtained through mean-based feature aggregation:

$$\bar{x}_j = \frac{1}{|B_j|} \sum_{x_i \in B_j} x_i \quad (7)$$

The resulting aggregated bag representations are thereafter utilized to train a nonlinear regression model. The objective is to construct a multiple instance learner F capable of accurately predicting unseen test instances.

$$F(\bar{x}_j) \rightarrow Y_j \quad (8)$$

The learner operates at the bag level and produces a decision for each specific test input. During the testing phase, each unseen test instance is processed individually using the trained regression model to estimate its corresponding SoC value.

3.3. Formal Description of the Proposed Approach

Following the construction of the MIL-based dataset, the proposed EV battery SoC estimation framework employs the light gradient boosting machine (LGBM) as the regression model, which is named as MIL-LGBM. LGBM is a specialized gradient-boosting decision tree method, particularly suited to learning complex nonlinear relationships while maintaining low computational complexity. This property makes it particularly suitable for MIL-based SoC estimation, where nonlinear interactions exist among voltage, current, temperature, power demand, acceleration, and other operational variables under real-world driving conditions.

The objective of the proposed MIL-based model is to learn the nonlinear mapping $F(\bar{x}_i) \rightarrow Y_i$ where $F(\cdot)$ denotes the LGBM-based regression function.

The final prediction model is constructed as an additive ensemble of regression trees:

$$\hat{Y}_i = \sum_{t=1}^T f_t(\bar{x}_i) \quad (9)$$

where T denotes the total number of decision trees and $f_t(\cdot)$ represents the t -th regression tree.

For MIL-based EV battery SoC estimation, LGBM iteratively minimizes the following objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^M l(Y_i, \hat{Y}_i^{(t)}) + \sum_{t=1}^T \Omega(f_t) \quad (10)$$

where $l(\cdot)$ denotes the regression loss function and $\Omega(f_t)$ represents the regularization term controlling model complexity.

At each boosting iteration, the prediction is updated as:

$$\hat{Y}_i^{(t)} = \hat{Y}_i^{(t-1)} + f_t(\bar{x}_i) \quad (11)$$

where the newly generated tree attempts to minimize the residual estimation error from the previous iteration.

The optimization process is performed using first-order and second-order gradient statistics:

$$g_i = \frac{\partial l(Y_i, \hat{Y}_i^{(t-1)})}{\partial \hat{Y}_i^{(t-1)}} \quad (12)$$

$$h_i = \frac{\partial^2 l(Y_i, \hat{Y}_i^{(t-1)})}{\partial (\hat{Y}_i^{(t-1)})^2}$$

where g_i and h_i denote the gradient and Hessian values, respectively, for MIL-based SoC estimation samples.

To improve computational efficiency for real-time SoC estimation, LGBM employs histogram-based learning. Continuous feature values are discretized into histogram bins,

significantly reducing split-search complexity. Let G_b and H_b denote the accumulated gradients and Hessians of histogram bin b :

$$G_b = \sum_{i \in b} g_i \quad H_b = \sum_{i \in b} h_i \quad (13)$$

The split gain for a candidate node partition is computed as:

$$\text{Gain} = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma \quad (14)$$

where G_L and G_R denote the gradient sums of the left and right child nodes, H_L and H_R denote the corresponding Hessian sums, λ is the regularization parameter, and γ represents the penalty applied for generating a new split.

For the proposed MIL-based SoC prediction framework, this split-gain strategy enables the model to identify the most informative feature partitions associated with battery operating behavior and driving conditions. The regularization function used in LGBM is defined as:

$$\Omega(f_t) = \gamma V + \frac{1}{2} \lambda \sum_{j=1}^V s_j^2 \quad (15)$$

where V denotes the total number of leaves in the regression tree, s_j represents the prediction score assigned to leaf j , γ controls the penalty for generating additional leaves, and λ is the L2 regularization parameter applied to leaf weights.

Unlike traditional level-wise tree construction strategies, LGBM adopts a leaf-wise tree growth mechanism. In the proposed MIL-based SoC estimation framework, this approach allows the model to concentrate its learning process on the most informative regions within the input feature space associated with driving behavior and battery dynamics. At each iteration, the algorithm identifies and chooses the leaf node that yields the highest split gain:

$$\text{Leaf}^* = \arg \max_{l \in \text{Leaves}} \text{Gain}(l) \quad (16)$$

where *Leaves* denotes the set of candidate leaf nodes.

The pseudo-code of the proposed MIL-LGBM method is presented in Algorithm 1. The algorithm begins by grouping sequential instances into bags based on their SoC labels, while respecting the predefined maximum bag size w . Following this, a mean-based aggregation technique is applied to generate bag-level feature representations, resulting in a multi-instance dataset (*MD*). These aggregated bags are then used to train the machine learning model. In the training phase, the algorithm iteratively constructs T regression trees by utilizing gradient and Hessian statistics, histogram-based split selection, and leaf-wise tree growth. At each boosting iteration, a new regression tree is added to minimize the residual estimation error of the current ensemble. In the final stage, the trained model is used to estimate the SoC value of each instance in the test set independently. By combining bag-level learning with gradient boosting, the proposed method is capable of capturing collective battery operating behavior.

Algorithm 1 Proposed MIL-LGBM Method

Inputs: D : dataset such that $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$

E : example test set such that $E = \{E_1, E_2, \dots, E_p\}$

w : bag size

T : total number of decision trees

Output: $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_p\}$ (a set of predictions that are assigned to instances in E)

$MD \leftarrow \emptyset$

$i \leftarrow 1$

```

while  $i \leq N$  do
    Create empty bag  $B$ 
     $soc \leftarrow y_i$ 
     $counter \leftarrow 0$ 
    while  $i \leq N$  and  $y_i = soc$  and  $counter < w$  do
         $B \leftarrow B \cup \{x_i\}$ 
         $i \leftarrow i + 1$ 
         $counter \leftarrow counter + 1$ 
    end while
     $\bar{x} \leftarrow (1/|B|) \sum_{x \in B} x$ 
     $Y \leftarrow soc$ 
     $MD \leftarrow MD \cup \{(\bar{x}, Y)\}$ 
end while
for  $t = 1$  to  $T$  do
    Compute gradients  $g$ 
    Compute Hessians  $h$ 
    Construct histogram bins
    Compute split gains using  $(G_L, G_R, H_L, H_R, \lambda, \gamma)$ 
    Select the node with the maximum gain
    Train regression tree  $f_t$ 
     $\hat{Y}^{(t)} = \hat{Y}^{(t-1)} + f_t(\bar{x})$ 
end for
 $F \leftarrow \sum_{t=1}^T f_t$ 
 $\hat{Y} \leftarrow \emptyset$ 
for  $i = 1$  to  $P$  do
     $\hat{y}_i = F(E_i)$ 
     $\hat{Y} = \hat{Y} \cup \hat{y}_i$ 
end for

```

The proposed MIL-LGBM methodology has several advantages for battery state estimation:

- First, the leaf-wise optimization approach enhances the ability of the framework to capture nonlinear patterns while maintaining computational efficiency.
- Second, the utilization of a histogram-based technique and regularization during tree growth notably reduces memory consumption and prediction latency. This advantage directly ensures the model is appropriate for real-time battery management units.
- Moreover, an important benefit is its seamless integration with XAI integration. In contrast to various deep learning architectures that operate as black-box models, the tree-based nature of MIL-LGBM enables a transparent assessment of feature importance and model logic.

4. Experimental Studies

4.1. Dataset Description

This study uses an open-access dataset [30] collected from the Guangzhou Intelligent Vehicle Monitoring Platform. The original dataset contains approximately 12.83 million records acquired over nearly one year under real-world on-road driving conditions and from diverse driving environments. However, a subset of the dataset is currently available, and this subset was utilized during the development and evaluation of the proposed framework.

Table 2 presents an overview of the vehicle specifications, including their types, battery chemistries, cell structures, cumulative mileage, initial rated capacities, and data acquisition details for the vehicles analyzed in this study. The dataset includes operational

records obtained from 10 fully electric vehicles, consisting of 7 passenger vehicles and 3 electric buses, thereby enabling cross-platform analysis under different driving characteristics. In addition, the dataset incorporates two distinct battery chemistries, comprising 6 vehicles equipped with Nickel Cobalt Manganese (NCM) batteries and 4 utilizing Lithium Iron Phosphate (LFP) technology. This diversity allows the proposed MIL-based SoC estimation framework to be evaluated under heterogeneous battery behaviors and electrochemical characteristics. Furthermore, the dataset also involves different battery pack structures, thereby increasing the variability of the analyzed battery systems and supporting a more comprehensive evaluation of the proposed MIL-based SoC estimation framework. For instance, Vehicles #1–#6 consist of 91 battery cells connected in series, Vehicle #9 contains a total of 360 battery cells, and Vehicle #10 contains 324 battery cells. Across the 10 monitored vehicles, the initial rated battery capacity ranges from 150 Ah to 645 Ah, depending on the vehicle type and chemistry.

Table 2. Technical specifications and data structural characteristics of the monitored electric vehicles, showing the diversity of battery capacity (120–645 Ah), dataset size, sampling frequency (0.1–0.5 Hz), vehicle type (Passenger and Bus), and battery chemistry (NCM and LFP). (NA: Not Available).

Vehicle ID	Vehicle Type	Battery Chemistry	Number of Instances	Initial Rated Capacity (Ah)	Cumulative Mileage (km)	Sampling Frequency (Hz)	Battery Pack Structure
1	Passenger	NCM	954,754	150	69,043	0.1	91
2	Passenger	NCM	998,243	150	73,950	0.1	91
3	Passenger	NCM	997,098	160	79,440	0.1	91
4	Passenger	NCM	1,150,999	160	96,279	0.1	91
5	Passenger	NCM	1,096,073	160	114,431	0.1	91
6	Passenger	NCM	501,031	160	27,318	0.1	91
7	Passenger	LFP	5,304,111	120	32,496	0.5	NA
8	Bus	LFP	675,236	645	82,668	0.1	NA
9	Bus	LFP	443,806	645	43,988	0.1	360
10	Bus	LFP	715,956	505	27,677	0.1	324

A strict data partitioning strategy was implemented based on the vehicle categories presented in Table 2. First, the data were grouped according to vehicle type, where Vehicles #1–#7 were used to construct the Passenger dataset and Vehicles #8–#10 were utilized to form the Bus dataset. Due to the distinct statistical characteristics and correlation structures of the two datasets, separate MIL-LGBM models were constructed and evaluated independently. This separation was essential to prevent cross-vehicle-type interference due to the different physical and operational nature of the vehicles; for instance, buses are equipped with substantially larger battery capacities (505–645 Ah compared to 120–160 Ah for passenger cars) and are typically subjected to high-frequency current transients resulting from intensive stop-and-go driving cycles. Similar to the data partitioning strategy established in the study utilizing the same dataset [30], an instance-based split was applied to each dataset independently with a fixed random seed (set to 42) to ensure reproducibility, assigning 80% of the instances to the training partition and reserving the remaining 20% as an unseen holdout test set for final model evaluation. The training subset was further assessed using 10-fold cross-validation to evaluate model stability and robustness. Therefore, throughout the cross-validation phase, one fold was systematically isolated as a validation set while the remaining folds were used for training during model development. Consequently, to prevent information leakage, the train-test split was performed before any model development stage, and the test subset remained completely unseen throughout training, hyperparameter optimization, and cross-validation.

According to the original dataset description in [30], during data acquisition, the vehicle onboard sensor signals were transmitted to the Battery Management System (BMS) through the Controller Area Network (CAN) bus and subsequently forwarded by the onboard telematics unit (T-BOX) to the signal base station at fixed sampling intervals. The vehicles were primarily sampled at 0.1 Hz, whereas a higher rate of 0.5 Hz was employed for one vehicle. The collected data were then transferred to the Original Equipment Manufacturer (OEM) and the Guangzhou New Energy Intelligent Vehicle Big Data Monitoring Platform for further processing. The SoC values were provided as one of the operational variables in the released dataset. In commercial EV fleet operations, onboard BMS units generally provide proprietary SoC values by integrating multiple sources of information rather than relying solely on open-loop counting. These localized micro-scale SoC values serve as reasonable and suitable industry-standard benchmarks for macro-scale data analytics. The objective of the proposed framework is therefore not to reproduce the BMS-reported SoC estimation procedure, but to develop a robust and explainable data-driven model capable of accurately predicting reference SoC values by handling potential signal noise and learning multi-instance block patterns present in remote telematics data streams. To further improve label reliability, the adopted preprocessing pipeline handles invalid records, outliers, and missing values before model training, as described in Section 4.2.

Compared with laboratory-controlled datasets, the utilized on-road dataset inherently captures realistic driving characteristics, including regenerative braking effects, user-dependent driving behaviors, fluctuating traffic conditions, and environmental disturbances. Consequently, the proposed MIL-based SoC estimation framework was evaluated under considerably more complex and practical operating conditions than many existing laboratory-based studies.

The original dataset comprised over 70 sensor signals related to vehicle dynamics, BMS measurements, and driving characteristics. Nevertheless, only the variables most relevant to battery SoC estimation were retained to reduce unnecessary complexity and improve computational efficiency. Specifically, Table 3 presents the utilized input features, including speed, pack voltage, current, cell maximum voltage, cell minimum voltage, cell maximum temperature, cell minimum temperature, temperature difference, acceleration, and power. The maximum and minimum cell voltage and temperature variables represent the highest and lowest values measured across different battery cells within the battery pack at each sampling instant. Among the variables in Table 3, the last three were generated through feature engineering, whereas the remaining were collected directly from onboard EV sensors via the CAN-BMS infrastructure.

Table 3. Description of the input features and target variable used for battery SoC estimation, including their physical meanings, measurement units, and data types.

Category	Feature Name	Description	Unit	Data Type
Vehicle Dynamics	speed	Instantaneous vehicle speed	km/h	Float
	acceleration	Vehicle acceleration	m/s ²	Float
	hv_voltage	Overall battery pack voltage	V	Integer
	hv_current	Overall battery pack current	A	Float
Electrical Characteristics	bcell_maxVoltage	Maximum voltage among different battery cells within the battery pack at the current sampling instant	V	Float
	bcell_minVoltage	Minimum voltage among different battery cells within the battery pack at the current sampling instant	V	Float

Thermal Characteristics	bcell_maxTemp	Maximum temperature among different battery cells within the battery pack at the current sampling instant	°C	Integer
	bcell_minTemp	Minimum temperature among different battery cells within the battery pack at the current sampling instant	°C	Integer
	dtemp	Temperature difference between battery cells	°C	Integer
Power Characteristics	power	Instantaneous battery power	kW	
Target Variable	State of Charge (SoC)	Remaining charge level relative to its capacity.	%	Integer

It should be noted that although explicit aging-related variables, such as state of health (SoH), are not direct inputs, the proposed MIL-LGBM framework implicitly accounts for varying battery states through high-fidelity voltage and current couplings (bcell_maxVoltage, bcell_minVoltage, hv_voltage, and hv_current). These features can manifest the impacts of internal resistance growth and capacity limitations dictated by the battery's current aging state. Furthermore, when the dataset spans nearly one year of real-world operation, potential aging effects can be inherently reflected in the measured electrical and thermal signals. Moreover, the use of long-term data can be beneficial as it enables the model to capture a broad range of operating conditions, battery behaviors, seasonal variations, and diverse ambient environments, thereby improving the robustness and generalization capability of the framework. For instance, the internal resistance and open-circuit voltage dynamics vary significantly between sub-zero winter conditions and high-temperature summer periods.

Figure 2 presents the Pearson correlation matrix computed across the features within the bus dataset to uncover dependencies. The critical observation from the heatmap is the strong positive correlation between SoC and both the maximum voltage ($r = 0.66$) and the minimum voltage ($r = 0.62$). These results suggest that cell voltage measurements contain valuable information for estimating battery charge levels in electric buses. Conversely, battery temperature-related variables exhibit moderate negative correlations with SoC, as evidenced by the scores for maximum temperature ($r = -0.46$) and minimum temperature ($r = -0.42$). This behavior aligns with electrochemical principles, where higher temperatures are typically associated with lower charge levels during operation. The results also indicate that vehicle speed and current show relatively weak linear relationships with SoC, implying that their influence may be nonlinear or dependent on interactions with other variables. Therefore, these feature interactions justify the necessity of a non-linear machine learning architecture capable of resolving dense interdependencies.

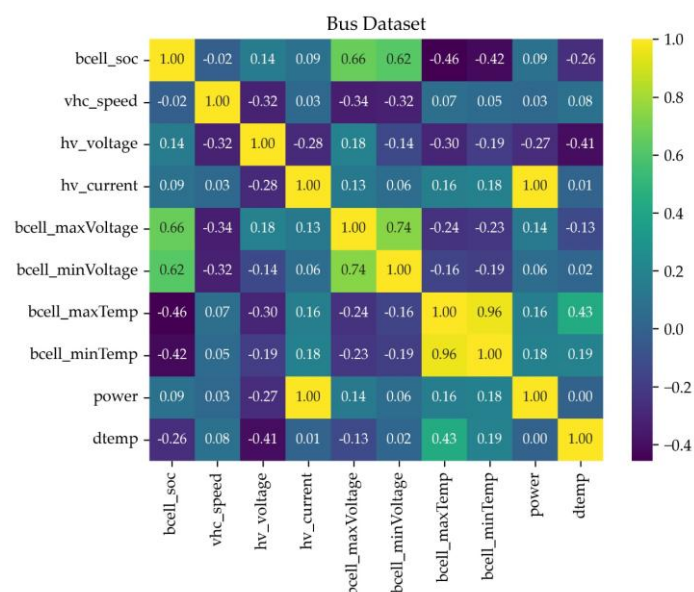


Figure 2. Heatmap of the bus dataset showing relationships among feature pairs, with yellow indicating high correlation (i.e., battery cell min-max voltages and SoC) and blue signifying low correlation (i.e., speed and SoC).

Figure 3 illustrates the feature correlation profile specifically evaluated for the passenger vehicle dataset. Unlike the bus dataset, substantially stronger positive correlations (around $r = 0.97$) between SoC and voltage-related features are observed, displaying the dominant role of voltage measurements in passenger vehicle battery charge estimation. In contrast to the bus dataset, thermal features such as maximum temperature (-0.02) and minimum temperature (-0.07) exhibit almost negligible linear correlation with SoC under on-road driving conditions. Overall, the low correlation blocks highlight a high degree of redundancy, which underpins the strategic integration of feature selection procedures prior to model deployment. The noticeable differences between Figures 2 and 3 indicate that the relationships between battery measurements and SoC may vary across vehicle platforms and battery systems. This further motivates the need for developing separate learning models capable of specifically capturing the unique characteristics of heterogeneous electric vehicles.

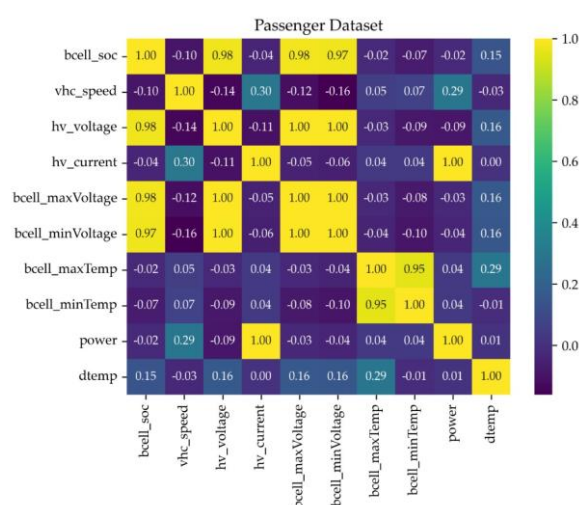


Figure 3. Heatmap of the passenger vehicle dataset showing relationships among feature pairs, with yellow indicating high correlation (i.e., battery voltage and SoC) and blue signifying low correlation (i.e., battery current and SoC).

Table 4 illustrates two sample bags to give an idea of how the features appear in the dataset. It demonstrates how individual instances are grouped under the multi-instance learning (MIL) paradigm. Each bag contains multiple consecutive instances that share the same SoC value and therefore receive a common bag label. For example, Bag 1 contains five instances associated with an SoC value of 88, while Bag 2 contains five instances associated with an SoC value of 80. This grouping strategy enables the model to collectively learn battery behavior from a collection of related operating conditions rather than from isolated records. This structure perfectly allows the model to absorb operational uncertainties and high-frequency sensor noise, converting localized instance variability into a stable, packet-level state-of-charge estimation. This MIL representation can reduce the influence of instantaneous fluctuations, resulting in a more informative representation for SoC estimation.

Table 4. Representative examples of MIL bags with a size of 5, constructed from sequential observations sharing the same SoC label.

Bag	Speed	Acceleration	hv Voltage	hv Current	bcell Max Voltage	bcell Min Voltage	bcell Max Temp	bcell Min Temp	dtemp	Power	SoC
1	63.0	0.6	373	18.2	4.110	4.088	24	22	2	6.79	88
1	62.4	0.6	374	21.9	4.117	4.097	24	22	2	8.19	88
1	56.2	6.2	373	41.5	4.139	4.092	24	22	2	15.48	88
1	45.3	10.9	373	42.4	4.116	4.089	24	22	2	15.82	88
1	49.5	4.2	372	62.3	4.119	4.095	24	22	2	23.18	88
2	66.5	7.8	366	23.5	4.038	4.012	22	21	1	8.60	80
2	68.6	2.1	365	33.1	4.029	4.007	22	21	1	12.08	80
2	76.2	7.6	364	32.3	4.022	3.997	22	21	1	11.76	80
2	76.2	0.0	365	27.3	4.026	3.999	22	21	1	9.96	80
2	76.7	0.5	365	16.0	4.024	4.007	22	21	1	5.84	80

Table 5 presents the descriptive statistics of the features used in the bus dataset. As shown, all continuous variables were normalized within a strict [0, 1] boundary prior to model development. The average SoC value is 75.173 with a standard deviation of 15.932, indicating that the dataset covers a broad range of battery charge levels. The data reveal that electric buses experience skewed speed profiles during operational routes. Among the input variables, battery pack voltage and current variables also span wide ranges, reflecting the diverse operating conditions encountered during real-world bus operation. Thermal features display relatively steady behavior with lower variances. These statistics demonstrate that the dataset contains substantial variability in both battery and vehicle operating characteristics, providing a suitable foundation for evaluating the robustness of the proposed MIL-LGBM method.

Table 5. Descriptive statistics of the normalized bus dataset, with most feature means centered around the mid-range of the normalized scale (mean $\approx$ 0.5) and moderate standard deviations indicating variability under real-world operating conditions.

Feature Name	Min	Max	Mean	Std
speed	0.0	1.0	0.237	0.236
acceleration	0.0	1.0	0.535	0.069
hv_voltage	0.0	1.0	0.503	0.285
hv_current	0.0	1.0	0.525	0.237
bcell_maxVoltage	0.0	1.0	0.377	0.068
bcell_minVoltage	0.0	1.0	0.507	0.088

bcell_maxTemp	0.0	1.0	0.509	0.164
bcell_minTemp	0.0	1.0	0.577	0.182
dtemp	0.0	1.0	0.053	0.026
power	0.0	1.0	0.505	0.231
State of Charge (SoC)	16.0	100.0	75.173	15.932

Table 6 presents the summary statistics and distributional properties characterizing the passenger vehicle dataset. Similar to the bus dataset, all predictor variables were normalized prior to model training. The overall operational mean SoC is 65.447 with a standard deviation of 21.041, indicating a wider distribution of charge levels compared with the bus dataset. The driving dynamics reflect distinct passenger-car habits, yielding a standard deviation of 0.199 for the speed variable. The current measurements show relatively lower dispersion, while temperature-related variables exhibit substantial variation across operating conditions, reflecting changing environmental and driving conditions experienced by passenger vehicles. Overall, the statistics reveal considerable diversity in the underlying data distribution and confirm that the dataset encompasses a broad range of battery operating states, which is essential for developing and validating a generalized SoC estimation model.

Table 6. Descriptive statistics of the normalized passenger vehicle dataset, showing feature means distributed across the lower and middle regions of the normalized scale (mean < 0.6) and varying levels of standard deviations, reflecting diverse battery and driving behaviors.

Feature Name	Min	Max	Mean	Std
speed	0.0	1.0	0.176	0.199
acceleration	0.0	1.0	0.477	0.052
hv_voltage	0.0	1.0	0.573	0.209
hv_current	0.0	1.0	0.499	0.093
bcell_maxVoltage	0.0	1.0	0.137	0.061
bcell_minVoltage	0.0	1.0	0.561	0.207
bcell_maxTemp	0.0	1.0	0.377	0.130
bcell_minTemp	0.0	1.0	0.366	0.133
dtemp	0.0	1.0	0.382	0.143
power	0.0	1.0	0.490	0.088
State of Charge (SoC)	1.0	100.0	65.447	21.041

4.2. Data Preprocessing

A systematic data preprocessing pipeline, comprising data reduction, data filtering, data cleaning, data imputation, deduplication, and normalization, was executed to prepare the datasets for the SoC estimation models. In terms of “Data Reduction”, both the timestamp and cumulative traveled distance variables were excluded from the final feature set. Time was eliminated late in the preprocessing pipeline because it was necessary to preserve record order for secondary tasks, such as temporal interpolation and acceleration calculation, which required consecutive time records. For “Data Filtering”, outliers were identified using the interquartile range (IQR) method, where observations outside a strict threshold of seven times the IQR were removed. Furthermore, the dataset was filtered to include only driving mode (charging_signal = 3) records. For “Data Cleaning”, we handled several physically impossible and erroneous readings, such as 0 and 65,535 V for cell voltage and 1310.7 V for battery voltage, as well as cell temperatures of −40 °C and 25°C. Regarding “Data Imputation”, invalid or missing measurements were replaced using linear interpolation. This approach preserves the continuity of sensor signals while minimizing information loss caused by corrupted records. In the “Deduplication” step,

we removed several consecutive records with all features (except time) the same, as well as several rows that were empty.

In the ‘Normalization’ step, the Passenger Vehicle, Bus, NCM, and LFP datasets were normalized independently, preventing differences in feature distributions from introducing bias across datasets. In each dataset, all continuous numerical inputs, except the target SoC values, were transformed using min–max normalization as a preprocessing standardization procedure. Table 7 evaluates the influence of normalization on the performance of the proposed MIL-LGBM method. The results indicate that the MAE values obtained with normalized and non-normalized data are nearly identical across all datasets. This behavior can be attributed to the fact that tree-based learning algorithms perform feature partitioning through threshold-based splits and are therefore largely insensitive to feature scaling. Nevertheless, normalization was retained in the preprocessing pipeline to ensure a consistent feature representation and to facilitate comparisons across heterogeneous vehicle and battery datasets.

Table 7. Performance comparison (MAE) of the proposed MIL-LGBM method under normalized and non-normalized data configurations, showing nearly identical prediction performance.

	Normalized Data	Non-Normalized Data
Passenger Vehicle	0.799	0.799
Bus	3.983	3.988
NCM	0.713	0.712
LFP	2.918	2.926

4.3. Feature Engineering

In this study, the feature space for SoC estimation was expanded by constructing three distinct derived variables: power, acceleration, and temperature difference. The incorporation of these specific variables aims to enhance estimation accuracy by capturing factors that directly influence battery degradation and charge state.

Power is a critical feature in SoC estimation because it is physically coupled with the battery’s internal states and dictates the safe operational boundaries of the propulsion system [1,2]. Incorporating power-related variables allows estimation algorithms to capture the complex, non-linear electrochemical behavior of the battery under dynamic loads such as rapid acceleration or regenerative braking [1–3]. Furthermore, modern Battery Management Systems (BMS) increasingly employ joint estimation frameworks, where power is a vital feature to ensure consistency across multiple state variables; this approach improves accuracy throughout the battery’s lifecycle and enables essential functions such as range prediction and torque control [1–3]. By leveraging power as a feature, machine learning models can better account for internal resistance variations and thermal fluctuations, which are otherwise difficult to infer from direct measurements alone [2–4]. Power is given by;

$$Power = Voltage \times Current \quad (17)$$

Acceleration is also a critical feature for SoC estimation because it can subject the battery to extreme dynamic current stresses that significantly impact the battery’s usable capacity and internal temperature. These rapid changes generate substantial voltage drops and thermal gradients, requiring the BMS to accurately estimate internal resistance in real time to maintain stable SoC tracking [1,2]. Moreover, acceleration is a primary input for determining the State of Function (SOF) and State of Power (SOP), as it helps the BMS judge whether the current battery state can meet immediate torque demands without risking lithium plating or triggering safety cut-offs [1,3]. Persistent aggressive acceleration also contributes to the thickening of the solid electrolyte interphase (SEI) layer, which in-

creases internal resistance and alters the voltage-SoC relationship over the battery's lifecycle [1,4]. Finally, accounting for acceleration-related factors, such as mechanical vibrations, is necessary because they can degrade physical capacity, introducing further nonlinearities that must be handled by robust estimation models to ensure accuracy under natural driving conditions [1,5].

$$Acceleration_i = Speed_i - Speed_{i-1}, \quad i = 2, 3, \dots, N \quad (18)$$

The temperature difference between the maximum and minimum cell temperatures, which should ideally be maintained within approximately 3 °C, is a vital feature for SoC estimation because it directly dictates the consistency of the battery pack's internal states [1]. Operating a battery pack across uneven temperatures leads to uneven cell aging, causing their nominal capacities and internal resistances to diverge over time [1,2]. Because State-of-Charge is defined as the ratio of remaining usable capacity to total capacity, these thermal gradients ensure that cells in the same string will reach different charge levels even under identical current loads [1,3]. Furthermore, temperature differences alter electrochemical reaction kinetics and polarization effects, making it difficult for standard "large cell" models to maintain accuracy without accounting for cell-to-cell inconsistencies [3,4]. Modern BMSs utilize these thermal deltas to inform joint estimation frameworks and inconsistency compensation models, which adjust SoC calculations to account for local hotspots that could otherwise lead to model divergence or safety hazards like thermal runaway [1,2,4,5]. By monitoring temperature differences, the BMS can better manage cell balancing and provide more reliable range predictions throughout the battery's lifecycle [1–3].

$$DTemp = BcellMaxTemp - BcellMinTemp \quad (19)$$

4.4. Feature Selection

Mutual Information (MI) analysis was employed to determine the features that contributed most to predicting SoC. It successfully quantifies both linear and nonlinear dependencies between the candidate variables and the target variable. In other words, MI can capture complex, higher-order statistical variations without making strong assumptions about the underlying data distribution. Features with higher MI scores provide greater information about SoC and are expected to contribute more significantly to the predictive performance of the estimation model. Formally, mutual Information is given as:

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (20)$$

where $MI(X, Y)$ quantifies the mutual dependence between the feature X and the target Y . Specifically, the term $p(x, y)$ signifies the joint probability of these variables, while the symbols $p(x)$ and $p(y)$ refer to their marginal probabilities, respectively. According to the formula: If X and Y are totally independent, then $p(x, y) = p(x)p(y)$; therefore, $\log(1) = 0$ and $MI = 0$. If X provides information about Y , the MI value exceeds 0.

Figure 4 presents the computed MI scores of the input features for the bus dataset. The results indicate that voltage-related variables are the most informative predictors of SoC. Power also exhibits substantial predictive relevance, highlighting the importance of energy consumption rates in electric buses. Conversely, vehicle dynamics variables such as speed and acceleration yield comparatively lower scores, suggesting a more limited direct contribution to SoC estimation. Overall, the findings demonstrate that battery electrical dynamics represent the primary sources of information for robust SoC prediction in the electric bus dataset.

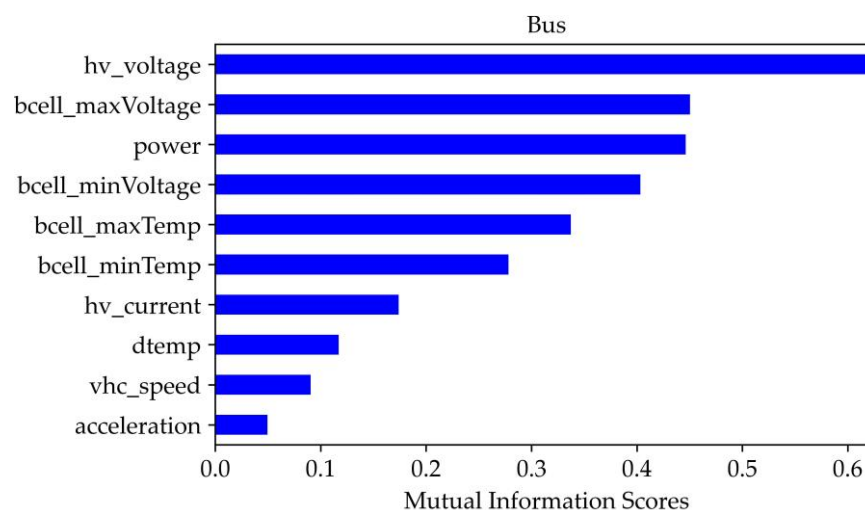


Figure 4. Feature importance scores for the bus dataset prior to model training, ranking the input variables according to their statistical relevance to the target SoC in descending order.

Figure 5 illustrates the feature dependency profile evaluated for the passenger vehicle dataset. Similar to the bus dataset, voltage-based measurements emerge as the most informative features. Power also provides considerable predictive information, revealing the comparative statistical influence of it on the SoC estimation process. Temperature, current, and driving-related variables exhibit relatively low contributions. All these results suggest that voltage measurements play an even more dominant role in passenger vehicle SoC estimation compared with the bus dataset.

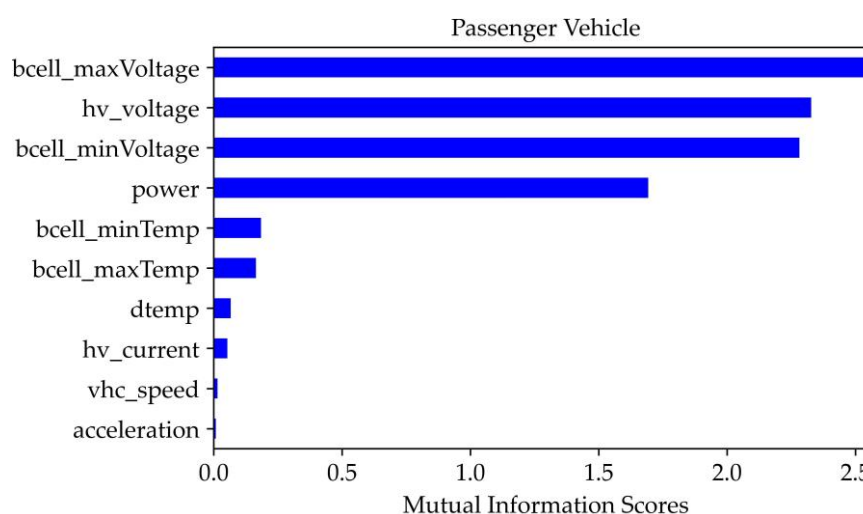


Figure 5. Feature importance scores for the passenger vehicle dataset prior to model training, ranking the input variables according to their statistical relevance to the target SoC in descending order.

After MI analysis, sequential feature pruning was executed by training the machine learning models on feature subsets, decreasing in cardinality from 10 to 4 based on their MI rankings. For both datasets, the best results were obtained by the nine-feature subset. Consequently, the lowest-ranked variable, namely acceleration, was systematically excluded from the input space, and the remaining nine features were utilized within the proposed MIL-LGBM framework.

4.5. Evaluation Metrics

The performance of the proposed SoC estimation framework was assessed using three universally accepted statistical metrics. Specifically, Mean Absolute Error (MAE),

Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2) were chosen to collectively provide a comprehensive assessment of prediction accuracy, error magnitude, and model fitting characteristics.

MAE acts as a statistical measure that quantifies the mean magnitude of absolute errors between estimated SoC values and their corresponding actual battery states. In this study, the MAE is highly useful as its final value is expressed in the same physical units as the original SoC measurements.

$$\text{MAE} = \frac{1}{N} \sum_{k=1}^N |\text{SoC}_{\text{pre}} - \text{SoC}_{\text{act}}| \quad (21)$$

where N is the number of samples, while SoC_{pre} and SoC_{act} represent the estimated and reference state-of-charge values within the testing dataset, respectively.

RMSE quantifies the precision of SoC estimates by determining the quadratic mean of the deviations between model predictions and reference battery states. By applying a squaring operation to the residuals, this metric inherently imposes a heavier penalty on substantial estimation inaccuracies.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^N (\text{SoC}_{\text{pre}} - \text{SoC}_{\text{act}})^2} \quad (22)$$

R^2 serves as an indicator of how well the predictions of a model match the real data. A value of 1 indicates that the predicted values perfectly match the actual data. A value of 0 corresponds to a baseline model, which can be achieved by just predicting the mean of all samples. Values less than 0 indicate that the model does not fit the data well, suggesting inadequate predictive capability.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (23)$$

4.6. Results

Table 8 presents a comparative performance evaluation of machine learning models on both the passenger vehicle and bus datasets in terms of MAE, RMSE, and R^2 . As investigated, all models were first implemented within the conventional Single-Instance Learning (SIL) framework. The results demonstrated that traditional linear architectures—namely Linear Regression, Ridge, Lasso, and Elastic Net—exhibited a limited capability to resolve SoC dynamics. Although the Long Short-Term Memory (LSTM) network obtained better results than the conventional linear models, it did not provide the best predictive performance. Non-linear machine learning models, including K-Nearest Neighbors and Neural Networks, produced competitive prediction performance but still exhibited room for improvement in predictive accuracy. Conversely, the tree-based gradient boosting (LGBM) substantially outperformed all these linear, non-linear, and neural-network-based methods across all evaluation metrics. In particular, SIL-LGBM achieved the lowest MAE and RMSE values while simultaneously producing the highest R^2 scores for both datasets. This superior performance of SIL-LGBM can be attributed to its ability to capture non-linear complex relationships among battery voltage, temperature, current, and other vehicle operating variables. Furthermore, unlike LSTM, which generally requires high computational cost, black-box, and complex model architectures, LGBM offers an efficient and interpretable learning framework through its tree-based structure. Based on these findings, the LGBM algorithm was selected as the base learner for the proposed MIL-based framework.

Table 8. Performance comparison of machine learning models under the Single-Instance Learning (SIL) framework for the passenger vehicle and bus datasets, demonstrating the superior performance of LGBM across all evaluation metrics (MAE, RMSE, and R^2).

Methods	Passenger Vehicle			Bus		
	MAE	RMSE	R^2	MAE	RMSE	R^2
SIL-Linear Regression	3.194	4.288	95.855	7.197	9.193	66.403
SIL-Ridge	3.195	4.288	95.855	7.195	9.193	66.400
SIL-Lasso	3.254	4.519	95.398	7.549	9.530	63.890
SIL-Elastic Net	6.671	8.785	82.607	10.858	13.314	29.523
SIL-K-Nearest Neighbors	1.042	1.918	99.171	4.509	7.527	77.472
SIL-LSTM	0.980	1.775	99.290	5.249	6.975	80.850
SIL-Neural Network	0.977	1.315	99.610	5.257	6.848	81.353
SIL-LGBM	0.887	1.203	99.674	4.410	5.924	86.048

Table 9 illustrates the performance enhancements achieved by transitioning from the Single-Instance Learning (SIL-LGBM) baseline to the proposed Multi-Instance Learning (MIL-LGBM) architecture after hyperparameter optimization. The empirical benchmarks consistently showed that MIL-LGBM achieved lower MAE and RMSE and higher R^2 scores than its SIL counterpart across both datasets. Specifically, for the passenger vehicle dataset, the MAE was decreased from 0.887 to 0.799. In particular, the proposed MIL representation reduced the MAE by 9.92% for the passenger vehicle dataset and 9.68% for the bus dataset while simultaneously decreasing RMSE. This systematic improvement stems from the inherent capability of the MIL paradigm to process time-series observations as aggregated data packets (“bags”) with shared labels. Unlike SIL, which treats each observation independently, MIL allows the model to learn collective battery operating behavior and reduces the influence of measurement noise and short-term sensor fluctuations while preserving meaningful battery state information. Consequently, the proposed MIL-LGBM method produced more robust and accurate state-of-charge estimations across different vehicle platforms.

Table 9. Performance comparison between the conventional SIL-LGBM and the proposed MIL-LGBM framework for SoC estimation on the passenger vehicle and bus datasets, including percentage improvements obtained by Multi-Instance Learning.

	Passenger Vehicle			Bus		
	SIL-LGBM	MIL-LGBM	Improvement	SIL-LGBM	MIL-LGBM	Improvement
MAE	0.887	0.799	9.92%	4.410	3.983	9.68%
RMSE	1.203	1.167	2.99%	5.924	5.765	2.68%
R^2	99.674	99.685		86.048	87.044	

Table 10 reports the performance of the proposed MIL-LGBM method on vehicles equipped with distinct lithium-ion battery chemistries, specifically contrasting Nickel Cobalt Manganese (NCM) and Lithium Iron Phosphate (LFP) systems. The results indicated that MIL-LGBM achieved high predictive accuracy for both battery types, demonstrating its capability to generalize across different electrochemical systems. Nevertheless, lower prediction errors are observed for NCM batteries compared with LFP batteries. This performance discrepancy is probably rooted in the underlying electrochemical behavior of the cells: NCM batteries generally exhibit a stronger and more continuous relationship between terminal voltage and SoC, whereas LFP batteries are characterized by extended voltage plateaus, making accurate charge-state estimation inherently more challenging. Despite these difficulties, the proposed MIL-LGBM model maintained high robustness on both chemistries, suggesting practical applicability across different battery technologies.

Table 10. Performance of the proposed MIL-LGBM framework on vehicles equipped with NCM and LFP battery chemistries, demonstrating its applicability across different battery technologies.

	NCM	LFP
MAE	0.713	2.918
RMSE	1.007	4.593
R ²	99.775	92.601

Figure 6 shows the variation in MAE with respect to the number of estimator parameters for the passenger vehicle and bus datasets. In both vehicle domains, prediction accuracy improved as more trees were incorporated into the ensemble. However, the rate of improvement gradually decreases as the number of estimators increases. For the passenger vehicle profile (Figure 6a), the error curve exhibited rapid exponential decay, whereas the bus profile (Figure 6b) demonstrated a more gradual convergence. For both vehicle types, improvements beyond approximately 50 estimators are minimal.

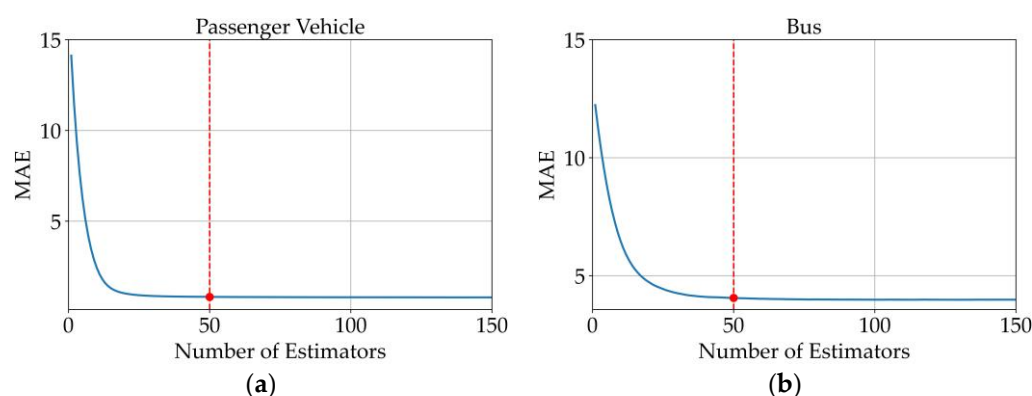


Figure 6. Convergence profiles of MAE with increasing number of estimators for (a) passenger vehicles and (b) buses, demonstrating rapid error stabilization with the optimal value marked at 50 decision trees by the red dashed line.

Figure 7 shows the variation in MAE with the number of estimators across datasets grouped by battery chemistry. Similar convergence behavior was observed for both NCM and LFP batteries. Prediction errors decrease rapidly during the initial boosting iterations and subsequently approach a stable region as the number of estimators increases. For both battery chemistries, improvements beyond approximately 50 estimators are minimal.

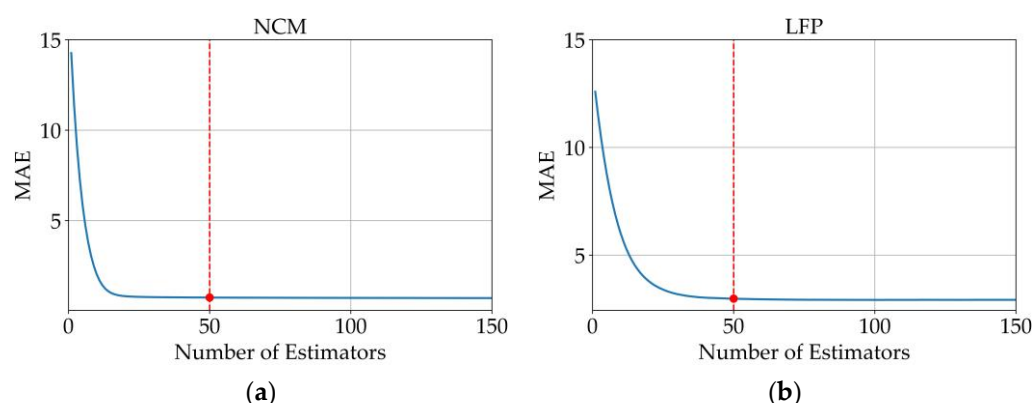


Figure 7. Error minimization trajectories with an increasing number of estimators evaluated for specific cell configurations: (a) NCM battery vehicles, and (b) LFP battery vehicles, demonstrating no meaningful improvement beyond the optimal value of 50 decision trees, indicated by a red dashed line.

Figure 8 presents the SoC prediction errors produced by the MIL-LGBM framework for 50 representative samples from the passenger vehicle and bus datasets. As illustrated in Figures 8a,b, the prediction errors are predominantly distributed around the zero-error line for the passenger vehicle model, whereas the bus model exhibits several larger deviations. Overall, the error distributions provide a visual indication that MIL-LGBM tends to achieve reasonable prediction performance.

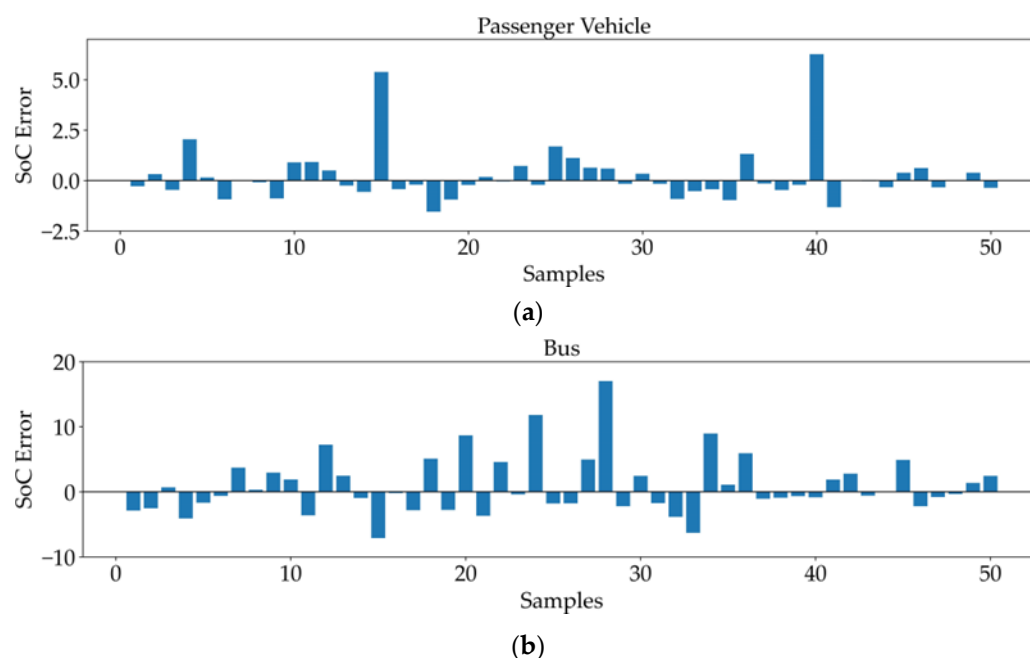
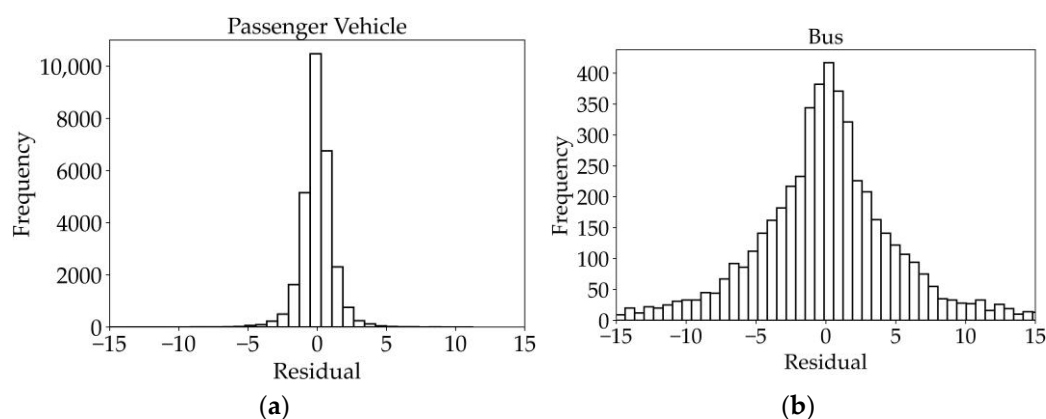


Figure 8. SoC prediction errors for 50 representative samples from (a) passenger vehicles and (b) buses, illustrating their magnitude and direction.

Figure 9 presents the error residual frequency distributions obtained by the MIL-LGBM method across all examined vehicle domains and battery chemistries. All residual histograms exhibited approximately bell-shaped distributions centered close to zero, validating that the learning model is free from substantial systematic prediction bias. Most prediction errors were concentrated within a relatively narrow range. The symmetry of the histograms suggests that overestimation and underestimation occurred with comparable frequency. Overall, the residual analysis confirmed the statistical reliability of the proposed framework for SoC predictions across different vehicle types and battery chemistries.



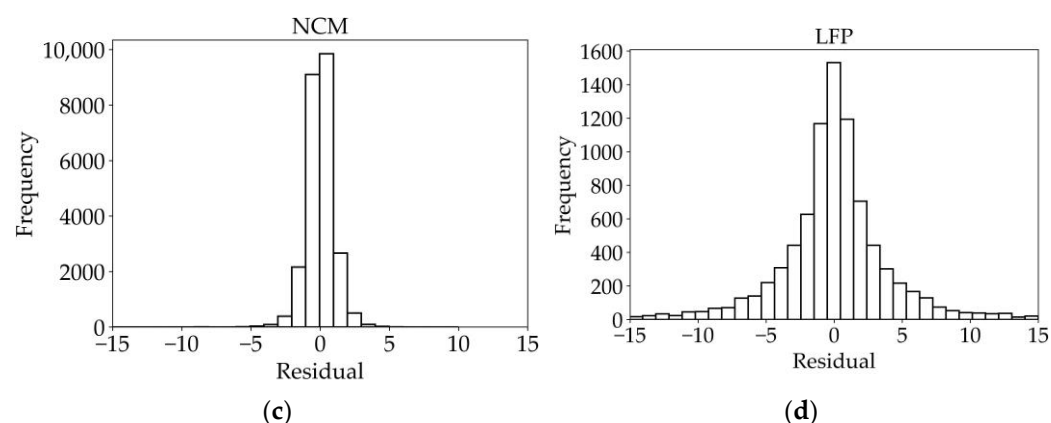


Figure 9. Residual distributions of the proposed MIL-LGBM framework for different vehicle and battery types: (a) passenger vehicles, (b) buses, (c) NCM battery vehicles, and (d) LFP battery vehicles, illustrating the prediction error characteristics across the evaluated datasets.

Figure 10 illustrates the relationship between the empirical SoC values and the model predictions across all experimental conditions. As shown in the figures, the majority of observations were concentrated near the diagonal reference line, indicating strong evidence of the accuracy of the models. Furthermore, no obvious systematic deviation was observed across the SoC range, suggesting that the model exhibits stable performance under both low and high charge conditions. In particular, the passenger vehicle and NCM scatter distributions showed minimal dispersion, demonstrating near-perfect correlation across the entire state-of-charge spectrum from deep discharge to full capacity (1–100%). Overall, the graphs confirmed that the MIL-LGBM framework successfully resolved structural-state tracking across diverse operational life cycles.

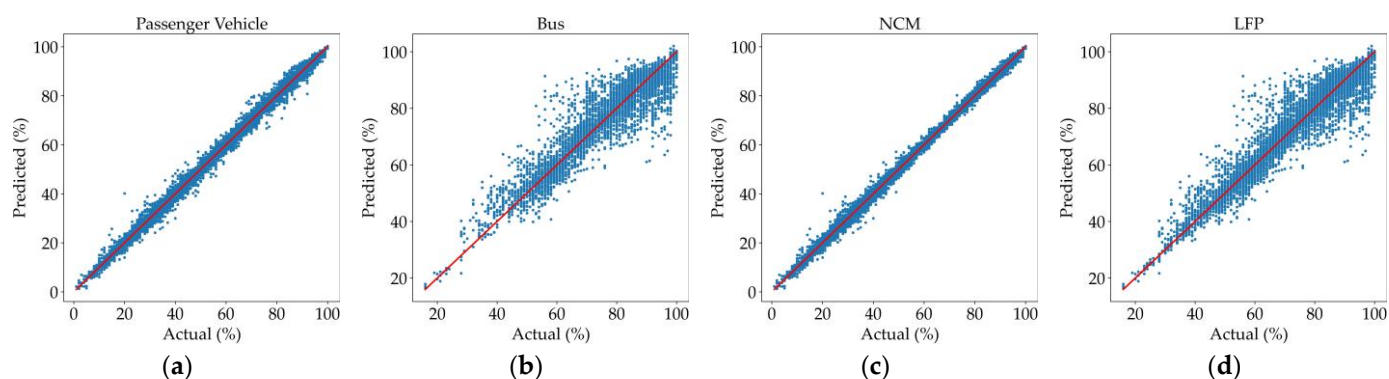


Figure 10. Scatter plots illustrating the agreement between actual and predicted SoC values with the red diagonal line representing the ideal prediction for different vehicle and battery types: (a) passenger vehicles, (b) buses, (c) NCM battery vehicles, and (d) LFP battery vehicles.

To investigate the benefit of the MIL paradigm over the SIL approach across different scales, a comprehensive sensitivity analysis was conducted by varying the bag size (w) from 5 to 15 in increments of 2. Table 11 presents the comparative MAE results and the corresponding relative improvement percentages of the proposed MIL-LGBM framework against the conventional SIL-LGBM baseline. The empirical evidence showed that MIL-LGBM consistently outperformed its SIL counterpart across all evaluated bag sizes for both vehicle categories. Depending on the bag size configuration, the MIL paradigm demonstrated a robust scalability profile, where the accuracy improvement fluctuated between 4.62% and 10.26% for the passenger vehicle segment, and between 2.74% and 9.68% for buses compared to the SIL baseline. The results confirm that the selected bag size provides an appropriate balance between preserving representative battery behavior and lim-

iting unnecessary information aggregation. Overall, the findings demonstrate that the performance gain achieved by MIL is systematic and consistent rather than incidental.

Table 11. Effect of different MIL-LGBM bag sizes on SoC estimation accuracy compared against the conventional SIL-LGBM baseline in terms of MAE, including improvement rates.

	Bag Size (w)	Passenger Vehicle		Bus	
		MAE	Improvement	MAE	Improvement
SIL-LGBM	-	0.887	-	4.410	-
MIL-LGBM	5	0.799	9.92%	3.983	9.68%
MIL-LGBM	7	0.796	10.26%	4.005	9.18%
MIL-LGBM	9	0.822	7.33%	4.032	8.57%
MIL-LGBM	11	0.831	6.31%	4.003	9.23%
MIL-LGBM	13	0.835	5.86%	4.106	6.89%
MIL-LGBM	15	0.846	4.62%	4.289	2.74%

Table 12 presents the computational efficiency of the proposed MIL-LGBM framework over the conventional SIL-LGBM approach under different bag sizes. The empirical profiles reveal a highly beneficial property of the MIL paradigm: it drastically accelerates the model training speed. For instance, under the $w = 5$ configuration, the training time decreased from 4.457 s to 1.800 s (59.61% reduction) for the passenger vehicle dataset and from 1.432 s to 0.765 s (46.58% reduction) for the bus dataset. Across all investigated window sizes, the reduction in training time ranges from 59.61% to 69.58% for the passenger vehicle dataset and from 46.58% to 55.17% for the bus dataset. This operational speedup is achieved because MIL operates on aggregated bags rather than individual observations, thereby reducing the effective number of records processed by the learning algorithm. Consequently, the proposed MIL framework not only improves SoC estimation accuracy but also substantially enhances computational efficiency, making it well-suited for large-scale real-world fleet applications.

Table 12. Training time comparison between the conventional SIL-LGBM and the proposed MIL-LGBM framework for different bag sizes, demonstrating the computational benefits of MIL.

	Bag Size (w)	Passenger Vehicle		Bus	
		Time (s)	Time Reduction	Time (s)	Time Reduction
SIL-LGBM	-	4.457	-	1.432	-
MIL-LGBM	5	1.800	59.61%	0.765	46.58%
MIL-LGBM	7	1.664	62.66%	0.740	48.32%
MIL-LGBM	9	1.500	66.35%	0.733	48.81%
MIL-LGBM	11	1.387	68.88%	0.678	52.65%
MIL-LGBM	13	1.356	69.58%	0.645	54.96%
MIL-LGBM	15	1.381	69.01%	0.642	55.17%

The benefits of the proposed MIL framework can be summarized as follows. First, MIL enables the model to learn collective battery behavior from groups of sequential observations instead of isolated samples, resulting in a more representative characterization of battery dynamics. Second, by aggregating temporally related instances into bags, the proposed framework becomes less sensitive to measurement noise and short-term fluctuations, thereby improving robustness and generalization. Third, this bag-level representation is particularly suitable for battery SoC estimation because SoC evolves more gradually than rapidly varying voltage, current, and temperature signals. Finally, the integration of MIL with LGBM provides a computationally efficient framework by learning from aggregated bags rather than individual observations, thereby reducing the effective train-

ing workload while simultaneously improving prediction accuracy over the conventional SIL framework.

4.7. Explainable AI Analysis for the Proposed Model

One of the principal strengths of the proposed MIL-LGBM framework is its inherent explainability and interpretable architecture. While high prediction accuracy is essential for reliable battery management, understanding why the model produces a particular SoC estimate is equally important for enhancing user confidence, validating model decisions, and helping engineers better understand battery behavior.

The XAI analysis provides several practical benefits over conventional feature ranking. First, it validates that the proposed MIL-LGBM model bases its predictions on physically meaningful battery characteristics rather than spurious correlations. Second, it enables battery engineers to identify the operating variables that most strongly influence SoC estimation under different vehicle categories. Finally, the decision tree structure and the SHAPs provide additional evidence regarding the robustness, transparency, and reliability of the proposed explainable MIL-LGBM framework. By revealing the vehicle-specific sensitivities, the XAI layer bridges the gap between data-driven black-box modeling and physical battery mechanisms, establishing the transparency required for trusted deployment in fleet monitoring systems.

To provide both global and local model interpretability, the proposed framework was analyzed using SHapley Additive exPlanations (SHAP). SHAP quantifies the contribution of each feature to the predicted SoC by assigning an additive importance value that reflects the marginal impact of that feature. Unlike conventional feature importance measures that only rank variables according to their overall contribution, SHAP additionally reveals whether high or low feature values drive the estimated SoC upward or downward, thereby providing a more comprehensive explanation of the trained model.

Figure 11a,b present the SHAP summary plots for the passenger vehicle and bus models, respectively. Each point corresponds to one observation, colored by its relative feature value ranging from low (blue) to high (red). The horizontal axis represents the SHAP value (impact on model output), where points to the right of the zero-line increase the predicted SoC, whereas points to the left decrease it. As observed in both vehicle models, battery cell maximum voltage emerges as the most dominant variable shaping the prediction process. Moreover, higher voltage values generally contribute positively to the estimated SoC, whereas lower voltage values shift the prediction toward lower SoC values. For the Bus model, temperature-related variables also exhibit noticeable contributions, reflecting the more diverse operational conditions experienced by these vehicles. High maximum cell temperatures (red dots) are generally associated with lower predicted SoC values, reflecting the elevated internal resistance and thermal stresses associated with heavy-duty stop-and-go bus operations.

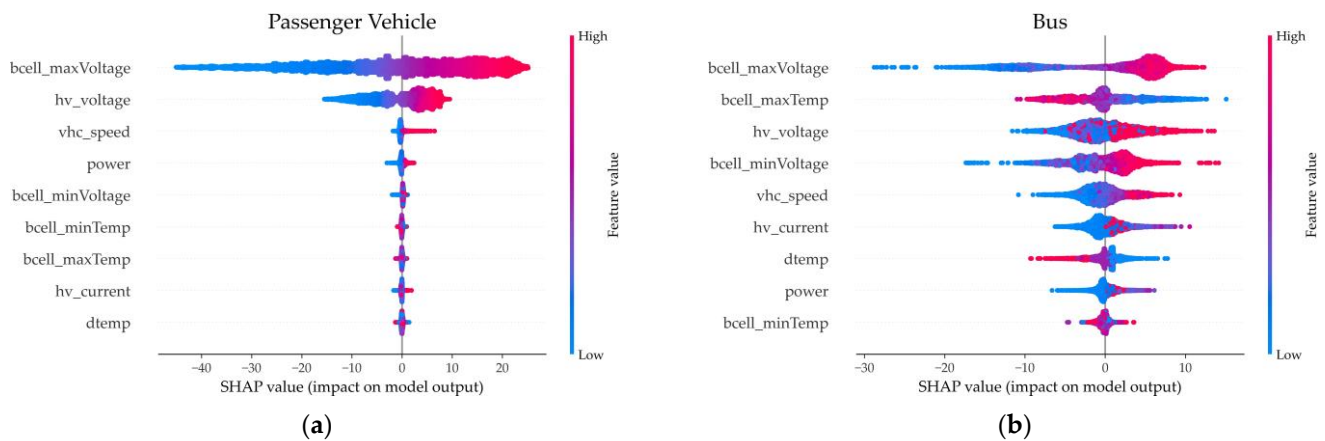


Figure 11. XAI-based SHAP summary plots illustrating global feature importance and directional dependencies on SoC estimation in the proposed MIL-LGBM framework for (a) passenger vehicles and (b) buses.

Figure 12 presents the decision tree derived from the proposed framework, providing an interpretable view of the learned decision-making process. One of the major advantages of the model is its human-readable structure, which enables users to trace how individual predictions are generated. In other words, the trained model has inherent explainability and transparent white-box properties. In general, features appearing near the upper levels of the tree have a greater influence on the final prediction. Consistent with the Mutual Information analysis presented in Figures 4 and 5, voltage-related variables occupy the most prominent decision nodes, highlighting their importance in battery SoC estimation. The tree also shows important decision points. For example, an initial branch decision is determined based on whether the pack voltage satisfies a threshold of 0.87. If true, the model follows the left branch to evaluate deeper boundaries; if false, it routes to the right to process other bounds. Such interpretable decision paths provide valuable insight into the physical factors influencing battery charge estimation and increase confidence in the model's predictions.

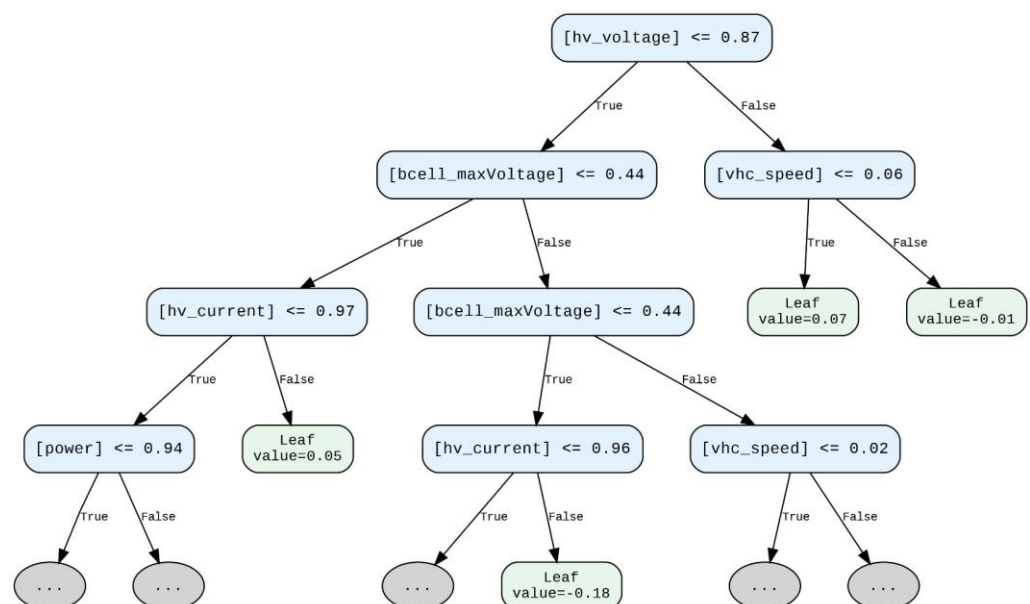


Figure 12. XAI-enabled decision tree illustrating the interpretable decision-making process, hierarchical logic, and feature-based prediction rules learned by the proposed MIL-LGBM framework.

Table 13 summarizes the feature importance statistics extracted from the decision tree. The results show the most influential variables, as indicated by the highest split count, split frequency, and depth-weighted importance score. Overall, the feature importance analysis reinforces the interpretability of the proposed framework and provides additional information.

Table 13. Feature importance statistics extracted from the XAI-enabled decision tree, indicating the dominant contribution of voltage-related variables to battery SoC estimation based on split frequency and weighted tree-depth scores.

Feature	Split Count	Split Frequency (%)	Depth Weighted Importance
bcell_minVoltage	52	26.130653	3.494765
vhc_speed	39	19.597990	3.071506
bcell_maxVoltage	35	17.587940	2.865682
hv_voltage	24	12.060302	2.391260
hv_current	26	13.065327	2.021943
power	11	5.527638	1.184526
dtemp	4	2.010050	0.345489
bcell_maxTemp	5	2.512563	0.323266
bcell_minTemp	3	1.507538	0.308824

5. Conclusions and Future Work

This study introduces MIL-LGBM, a specialized method for battery state-of-charge estimation. Unlike existing studies that rely on single-instance learning, focus on a single vehicle type or battery chemistry, and are evaluated under controlled laboratory conditions, this study was designed to investigate various vehicle platforms and battery technologies under realistic on-road environments. The proposed framework combines a multi-instance learning paradigm and a gradient boosting technique to capture collective battery operating behavior while maintaining model interpretability.

The proposed methodology was evaluated using datasets collected from both electric buses and passenger vehicles equipped with NCM and LFP batteries. Feature engineering and mutual information-based feature selection were employed to enhance data quality, while hyperparameter optimization further improved model performance. Experimental results demonstrated that the MIL representation consistently improved prediction performance compared with the conventional SIL approach. In both bus and passenger datasets, the proposed model achieved meaningful reductions in both MAE and RMSE, while an increase was observed in R^2 . For passenger vehicles, the proposed MIL-LGBM method achieved an MAE of 0.799. The findings of the study confirm that the integration of MIL enables the learning algorithm to exploit collective battery behavior patterns more effectively than conventional instance-based approaches. The results indicate that diverse on-road datasets coming from real vehicles under real road conditions can be effectively utilized by the proposed MIL-LGBM method. Moreover, explainability analyses were provided based on feature importance rankings and a representative decision tree.

Future research directions will focus on developing a user-friendly on-board software interface to integrate the proposed model into practical EV ecosystems. This interface will aim to provide end-users with diagnostic insights to enhance energy management, enabling predictive trip and charging planning. Such systems can provide drivers with accurate battery-state information and help extend battery lifespan.

Author Contributions: Conceptualization, I.A.K.; methodology, I.A.K.; software, I.A.K.; validation, D.B.; formal analysis, I.A.K.; investigation, D.B.; resources, I.A.K.; data curation, I.A.K.; writing—original draft preparation, I.A.K. and D.B.; writing—review and editing, I.A.K. and D.B.; visualization, I.A.K. and D.B.; supervision, D.B.; project administration, D.B.; funding acquisition, I.A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset utilized in this study is publicly available [30] on the GitHub repository (<https://github.com/Translab-SCUT/Electric-vehicle-operation-data>, accessed on 8 March 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AEKF	Adaptive extended Kalman filter
AM	Attention mechanism
AUKF	Adaptive unscented Kalman filter
BiLSTM	Bidirectional long short-term memory
CNN	Convolutional neural network
DAKF	Dual adaptive Kalman filter
DL	Deep learning
DT	Decision tree
ECM	Equivalent circuit model
EKF	Extended Kalman filter
ET	Extra Trees
EV	Electric vehicles
FNN	Feedforward neural network
FOA	Fractional-order adaptive
GMMCC	Generalized mixture maximum correlation-entropy criterion
GRU	Gated recurrent unit
KF	Kalman Filter
KNN	K-nearest neighbors
LFP	Lithium iron phosphate
LGBM	Light gradient boosting machine
LMO	Lithium manganese oxide
LR	Linear regression
LSTM	Long short-term memory
MAE	Mean absolute error
MaxE	Maximum error
MIL	Multi-instance learning
ML	Machine learning
MSE	Mean squared error
NARX	Nonlinear autoregressive exogenous
NCM	Nickel cobalt manganese
NN	Neural network
QPSO	Quantum-behaved particle swarm optimization
RF	Random forest
RMSE	Root mean squared error
SIL	Single instance learning
SMO	Sliding mode observers
SoC	State of charge
SoH	State of health
SVM	Support vector machine
TFF	Triple forgetting factor
TCN	Temporal convolutional network

UKF	Unscented Kalman filter
XAI	Explainable artificial intelligence
XGBoost	Extreme gradient boosting

References

1. Albakri, M.; Darwish, A. A Review of Modelling, State of Charge Estimation and Management Methods of EV Lithium-Ion Batteries. *Batteries* **2026**, *12*, 92. <https://doi.org/10.3390/batteries12030092>.
2. Yang, Y.; Liang, J. Hybrid state-of-charge estimation for lithium-ion batteries: A structured survey. *Syst. Sci. Control Eng.* **2026**, *14*, 2640255. <https://doi.org/10.1080/21642583.2026.2640255>.
3. Zhang, Q.; Rong, H.; Zhao, D.; Pei, M.; Dong, X. A Critical Review of the State Estimation Methods of Power Batteries for Electric Vehicles. *Energies* **2025**, *18*, 3834. <https://doi.org/10.3390/en18143834>.
4. Marques, T.M.B.; dos Santos, J.L.F.; Castanho, D.S.; Ferreira, M.B.; Stevan, S.L., Jr.; Illa Font, C.H.; Antonini Alves, T.; Piekarski, C.M.; Siqueira, H.V.; Corrêa, F.C. An Overview of Methods and Technologies for Estimating Battery State of Charge in Electric Vehicles. *Energies* **2023**, *16*, 5050. <https://doi.org/10.3390/en16135050>.
5. Sarda, J.; Patel, H.; Papat, Y.; Hui, K.L.; Sain, M. Review of Management System and State-of-Charge Estimation Methods for Electric Vehicles. *World Electr. Veh. J.* **2023**, *14*, 325. <https://doi.org/10.3390/wevj14120325>.
6. Zhao, D.; Li, H.; Zhou, F.; Zhong, Y.; Zhang, G.; Liu, Z.; Hou, J. Research Progress on Data-Driven Methods for Battery States Estimation of Electric Buses. *World Electr. Veh. J.* **2023**, *14*, 145. <https://doi.org/10.3390/wevj14060145>.
7. Zhang, D.; Zhong, C.; Xu, P.; Tian, Y. Deep Learning in the State of Charge Estimation for Li-Ion Batteries of Electric Vehicles: A Review. *Machines* **2022**, *10*, 912. <https://doi.org/10.3390/machines10100912>.
8. Kang, S.; Lee, Y.; Jang, G.; Lee, S. Lightweight 1D-CNN-Based Battery State-of-Charge Estimation and Hardware Development. *Electronics* **2026**, *15*, 704. <https://doi.org/10.3390/electronics15030704>.
9. Qiu, J.; Zhang, Z.; Zhu, Z.; Luo, C. Accurate State of Charge Estimation for Lithium-Ion Batteries Using a Temporal Convolutional Network and Bidirectional Long Short-Term Memory Hybrid Model. *Batteries* **2026**, *12*, 50. <https://doi.org/10.3390/batteries12020050>.
10. Li, R.; Hao, Y.; Zhang, M.; Lv, Y. State-of-Charge Estimation of Lithium-Ion Batteries Based on the CNN-Bi-LSTM-AM Model Under Low-Temperature Environments. *Sensors* **2026**, *26*, 264. <https://doi.org/10.3390/s26010264>.
11. Yang, Y.; Zhou, Y.; Liu, Z.; Jiang, S.; Xue, H.; Guo, P. A temporal convolutional network-Transformer-CrossAttention model for state of charge estimation of lithium-ion batteries. *J. Energy Storage* **2026**, *141*, 119417. <https://doi.org/10.1016/j.est.2025.119417>.
12. Nohra, C.; Ghandour, R.; Nehme, B.; Khaled, M.; Outbib, R. A Novel μ -Analysis-Based Estimator for State of Charge and State of Health Estimation in Lithium-Ion Batteries for Electric Vehicles. *World Electr. Veh. J.* **2026**, *17*, 86. <https://doi.org/10.3390/wevj17020086>.
13. Yang, L.; He, M.; Ren, Y.; Gao, B.; Qi, H. State of Charge Estimation for Lithium-Ion Batteries via a Diffusion-Enhanced Equivalent Circuit Model and an Improved Extended Kalman Filter. *Energy* **2026**, *345*, 140181. <https://doi.org/10.1016/j.energy.2026.140181>.
14. Guo, H.; Li, Z.; Xue, H.; Luo, J. State of Charge Estimation Method for Lithium-Ion Batteries Based on Online Parameter Identification and QPSO-AUKF. *Batteries* **2026**, *12*, 84. <https://doi.org/10.3390/batteries12030084>.
15. Li, F.; Wang, H.; Chen, H.; Geng, L.; Wu, C. State-of-Charge Estimation of Lithium-Ion Batteries Based on GMMCC-AEKF in Non-Gaussian Noise Environment. *Batteries* **2026**, *12*, 29. <https://doi.org/10.3390/batteries12010029>.
16. Li, S.; Yang, S.; Jiang, Z.; Jiang, W.; Zhu, Z.; Ma, Y. An online state of charge estimation method: Support vector machine battery model fusing a filtering algorithm. *Electr. Power Syst. Res.* **2026**, *253*, 112530. <https://doi.org/10.1016/j.epsr.2025.112530>.
17. Dharani, A.; Patange, S.S.; Krishna, M.; Vidya, R.; Bindu, R.; Subramanya, K.N.; Kiran, P.V.R.S.; Raja, S. A Digital Twin Framework for Dual-Path Estimation of Battery State of Charge and State of Health Using Hybrid Physics-Informed Machine Learning. *IEEE Access* **2026**, *14*, 6911–6926. <https://doi.org/10.1109/ACCESS.2026.3652279>.
18. Souza, K.M.d.; Galvão, J.R.; Mondadori, J.A.P.; França, M.B.d.M.; Junior, P.B.; Corrêa, F.C. Online Estimation of Lithium-Ion Battery State of Charge Using Multilayer Perceptron Applied to an Instrumented Robot. *Batteries* **2026**, *12*, 25. <https://doi.org/10.3390/batteries12010025>.
19. Du, L.; Shen, X.; Li, X.; Li, X.; Wei, Y. Battery state of charge estimation from electrochemical impedance spectroscopy data based on machine learning with fractional order model parameters and distribution relaxation time features. *J. Energy Storage* **2026**, *144*, 119754. <https://doi.org/10.1016/j.est.2025.119754>.

20. Joh, W.; Park, K.; Shin, D.; Kim, J. GPS-Assisted State-of-Charge Prediction for Electric Vehicles in Shuttle Service Applications. *Electronics* **2026**, *15*, 1014. <https://doi.org/10.3390/electronics15051014>.
21. Long, C.; Li, B.; Wen, J. A state of charge feature selection method for electric buses via fuzzy information measures. *Appl. Soft Comput.* **2026**, *186*, 114190. <https://doi.org/10.1016/j.asoc.2025.114190>.
22. Volturino, S.; Di Martino, A.; Longo, M. Machine Learning-Based State-of-Charge Prediction for Electric Bus Fleet: A Critical Analysis. *Electronics* **2025**, *14*, 4147. <https://doi.org/10.3390/electronics14214147>.
23. ElMenshawhy, M.S.; Massoud, A.M.; Guglielmi, P. State-of-Charge Estimation Using Triple Forgetting Factor Adaptive Extended Kalman Filter for Battery Energy Storage Systems in Electric Bus Applications. *IEEE Trans. Transp. Electrification* **2025**, *11*, 2. <https://doi.org/10.1109/TTE.2024.3514704>.
24. Vieira, R.; Kollmeyer, P.; Pitault, L.; Emadi, A. Comprehensive comparison of machine learning and Kalman filter battery state of charge estimators. *IEEE Access* **2025**, *13*, 36321–36338. <https://doi.org/10.1109/ACCESS.2025.3545380>.
25. El-Sayed, E.I.; ElSayed, S.K.; Alsharef, M. Data-Driven Approaches for State-of-Charge Estimation in Battery Electric Vehicles Using Machine and Deep Learning Techniques. *Sustainability* **2024**, *16*, 9301. <https://doi.org/10.3390/su16219301>.
26. Zhao, X.; Zhang, M.; Xue, G. Data-Driven Algorithm Based on Energy Consumption Estimation for Electric Bus. *World Electr. Veh. J.* **2023**, *14*, 329. <https://doi.org/10.3390/wevj14120329>.
27. Zeng, J.; Wang, S.; Cao, W.; Zhang, M.; Fernandez, C.; Guerrero, J.M. Improved fractional-order hysteresis-equivalent circuit modeling for the online adaptive high-precision state of charge prediction of urban-electric-bus lithium-ion batteries. *Int. J. Circuit Theory Appl.* **2024**, *52*, 420–438. <https://doi.org/10.1002/cta.3746>.
28. Reshma, P.; Manohar, V.J. Collaborative evaluation of SoC, SoP and SoH of lithium-ion battery in an electric bus through improved remora optimization algorithm and dual adaptive Kalman filtering algorithm. *J. Energy Storage* **2023**, *68*, 107573. <https://doi.org/10.1016/j.est.2023.107573>.
29. Srirattanawichaikul, W.; Wirasanti, P. Evaluation of lightweight battery management system with field test of electric bus in campus transit system. *Int. J. Electr. Comput. Eng.* **2020**, *10*, 6202–6213. <https://doi.org/10.11591/ijece.v10i6.pp6202-6213>.
30. Jing, H.; Hu, J.; Ou, S.; Lv, Z.; Lyu, R.; Zhao, J. Scalable and Generalizable Deep Learning for Battery State of Health Estimation in On-Road Electric Vehicles. *J. Energy Chem.* **2025**, *110*, 823–841. <https://doi.org/10.1016/j.jechem.2025.07.020>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.