International Journal of Science and Technology Research Archive

ISSN: 0799-6632 (Online)

Journal homepage: https://sciresjournals.com/ijstra/

(REVIEW ARTICLE)

Check for updates

Explainable Artificial Intelligence (XAI) in Battery Management Systems: A Comprehensive Review

Jong Myoung Kim *

Department of Artificial Intelligence and Big Data, Sehan University.

International Journal of Science and Technology Research Archive, 2025, 08(02), 014-026

Publication history: Received on 26 February 2025; revised on 07 April 2025; accepted on 09 April 2025

Article DOI: https://doi.org/10.53771/ijstra.2025.8.2.0034

Abstract

Battery Management Systems (BMS) are crucial for the safe and efficient operation of lithium-ion batteries in applications ranging from electric vehicles to grid storage. While Artificial Intelligence (AI) and Machine Learning (ML) have significantly advanced BMS capabilities, particularly in state estimation and fault diagnosis, the inherent 'black-box' nature of many complex models raises concerns about reliability, trustworthiness, and safety. Explainable Artificial Intelligence (XAI) offers methods to render these AI/ML models transparent and interpretable. This paper provides a comprehensive review of the application of XAI techniques within various BMS tasks. We survey the literature on XAI applied to state-of-charge (SOC), state-of-health (SOH), and remaining useful life (RUL) estimation, as well as fault detection and diagnosis, and charging management. Key XAI methodologies employed in BMS research, such as SHAP, LIME, attention mechanisms, and inherently interpretable models, are discussed. We analyze current trends, identify significant challenges including real-time implementation, evaluation of explanations, and data limitations, and suggest promising future research directions. This review aims to serve as a valuable resource for researchers and practitioners seeking to develop more transparent, reliable, and trustworthy intelligent BMS solutions.

Keywords: Explainable AI (XAI); Battery Management System (BMS); Lithium-Ion Battery; State Estimation; Fault Diagnosis; Machine Learning

1 Introduction

Lithium-ion batteries (LIBs) have become ubiquitous energy storage solutions, powering electric vehicles (EVs), portable electronics, and grid-scale storage systems [1]. Ensuring their safe, reliable, and efficient operation hinges on sophisticated Battery Management Systems (BMS) [2, 3]. A BMS performs critical functions, including monitoring battery states, ensuring operation within safe limits, optimizing performance, and prolonging lifespan [2].

Driven by the increasing complexity of battery systems and the availability of large datasets, Artificial Intelligence (AI) and Machine Learning (ML) techniques have been increasingly integrated into BMS functionalities [4]. ML models have demonstrated remarkable success in tasks such as estimating State of Charge (SOC), State of Health (SOH), Remaining Useful Life (RUL) [5, 6], and detecting and diagnosing various fault conditions [7, 8], often surpassing traditional methods in accuracy.

However, many high-performing AI/ML models, particularly deep learning architectures, operate as "black boxes," providing predictions without clear explanations of their reasoning [9]. This lack of transparency poses significant challenges in safety-critical applications like BMS. Operators and engineers may hesitate to trust or act upon predictions without understanding their basis [10], potentially hindering the adoption of advanced AI solutions. Furthermore,

* Corresponding author: Jong Myoung Kim

understanding *why* a model makes a certain prediction is crucial for debugging, validation, and identifying potential biases or reliance on spurious correlations.

Explainable Artificial Intelligence (XAI) has emerged as a vital field to address this opacity [11]. XAI encompasses a range of techniques designed to make AI/ML model decisions understandable to humans. By providing insights into model behavior—such as identifying influential input features or visualizing decision boundaries—XAI aims to foster trust, enhance safety, and facilitate human-AI collaboration [11, 12].

While previous reviews have focused on ML in BMS [4] or XAI for specific tasks like fault diagnosis [13], a comprehensive overview of XAI applications across the *breadth* of BMS functions is still needed. This review aims to fill that gap by systematically surveying the state-of-the-art literature on XAI techniques applied to various core BMS tasks. We will discuss the methodologies employed, summarize key findings, identify common challenges, and outline future research trajectories. This paper intends to provide researchers and practitioners with a holistic understanding of how XAI is being used and can be further leveraged to create more intelligent, reliable, and trustworthy BMS.

The remainder of this paper is organized as follows: Section 2 briefly outlines key BMS tasks. Section 3 discusses the general application of AI/ML in BMS. Section 4 introduces relevant XAI techniques. Section 5 reviews XAI applications in specific BMS tasks. Section 6 discusses challenges and open issues. Section 7 suggests future research directions, and Section 8 concludes the paper.

2 Overview of Battery Management System Tasks

An effective BMS performs several critical functions to ensure battery safety, performance, and longevity [2, 3]. **Table 1** provides a summary of these core tasks, which are further elaborated below.

Table 1 Core Functions of a Battery Management System (BMS)

Task Name	**Primary Purpose/Goal**
State Estimation	Estimate internal states (SOC, SOH, RUL) that cannot be directly measured.
State of Charge (SOC)	Determine remaining battery capacity (fuel gauge).
State of Health (SOH)	Assess current condition relative to new (degradation level).
Remaining Useful Life (RUL)	Predict time/cycles until end-of-life.
Fault Detection & Diagnosis (FDD)	Identify occurrence, type, and location of faults to ensure safety and prevent hazardous events.
Charging Management	Optimize charging process for speed, safety, and minimal degradation.
Cell Balancing	Equalize charge among cells in a pack to maximize usable capacity and prevent cell stress.
Thermal Management	Monitor and control battery temperature to maintain optimal operating range for performance and safety.

- **State Estimation:** This involves accurately estimating internal battery states that cannot be directly measured. Key states include:
    - *State of Charge (SOC):* Represents the available capacity relative to the maximum, akin to a fuel gauge, crucial for range prediction and operational planning.
    - *State of Health (SOH):* Assesses the battery's current condition compared to its fresh state, typically reflecting capacity fade and impedance increase, which informs maintenance and replacement decisions.
    - *Remaining Useful Life (RUL):* Predicts the time or number of cycles remaining until the battery reaches its end-of-life criteria, essential for warranty and operational planning. Accurate estimation of these states is challenging due to complex electrochemical processes and varying operating conditions.
- **Fault Detection and Diagnosis (FDD):** This function is paramount for safety. It involves identifying the occurrence, type (e.g., internal short circuit, overcharge, sensor malfunction), and sometimes location of faults

within the battery pack to enable timely intervention and prevent potentially hazardous events like thermal runaway.

- **Charging Management:** This task focuses on optimizing the charging process. It aims to balance rapid charging with minimizing battery degradation and ensuring safety by controlling charging current, voltage, and temperature according to the battery's state and limits.
- **Cell Balancing:** In multi-cell packs, individual cell capacities and resistances inevitably diverge. Cell balancing aims to equalize the SOC across cells, either passively (dissipating energy from higher-charged cells) or actively (transferring energy between cells), thereby maximizing the usable pack capacity and preventing premature aging of individual cells.
- **Thermal Management:** Maintaining the battery within its optimal temperature range is critical for both performance and safety. This involves monitoring cell temperatures and controlling heating or cooling systems (e.g., fans, liquid cooling) to mitigate overheating during high power operation or underheating in cold conditions.

3 AI/ML Applications in BMS

While traditional BMS often rely on model-based approaches (e.g., equivalent circuit models, Kalman filters) which provide physical interpretability but can struggle with accuracy under diverse conditions [14], AI/ML techniques have emerged as powerful data-driven alternatives [4]. These methods learn complex patterns and non-linear relationships directly from sensor data (voltage, current, temperature, etc.), often leading to improved performance in various BMS tasks, as summarized in **Table 2**.

Table 2 Examples of AI/ML Techniques Applied to BMS Tasks

BMS Task	Common AI/ML Techniques Used	Example Application/Goal
State Estimation (SOC/SOH/RUL)	RNNs (LSTM, GRU), SVM, GPR, Ensemble Methods	Estimate internal states accurately under dynamic conditions.
Fault Detection & Diagnosis (FDD)	Classifiers (SVM, RF, CNN), Anomaly Detection (Autoencoders, OC-SVM)	Identify known fault types or detect unexpected deviations.
Charging Management	Reinforcement Learning (RL), Optimization Algos.	Learn optimal charging policies balancing speed vs. health.
Cell Balancing	ML Classifiers/Regressors	Predict optimal balancing currents or control strategies.
Thermal Management	Regression Models (NNs, GPR), Control Algorithms	Predict temperature distribution, optimize cooling/heating.

- **State Estimation (SOC, SOH, RUL):** AI/ML models excel at mapping the complex, non-linear relationships between measurable signals (like voltage curves during operation) and internal battery states. Techniques adept at handling time-series data, such as Recurrent Neural Networks (RNNs, including LSTMs and GRUs), are widely employed. Support Vector Machines (SVM), Gaussian Process Regression (GPR), and various ensemble methods are also frequently used to estimate SOC, SOH, and RUL, often achieving higher accuracy compared to traditional methods, especially under dynamic operating profiles and when accounting for complex aging effects [5, 6]. Their strength lies in capturing subtle degradation indicators from operational data that are challenging to represent with purely physics-based models.
- **Fault Detection and Diagnosis (FDD):** The ability of AI/ML to recognize intricate patterns in high-dimensional data makes it particularly well-suited for FDD. For identifying known fault types, supervised learning models like SVM, Random Forests (RFs), and Convolutional Neural Networks (CNNs) – which can effectively process spatial or temporal patterns in sensor readings – are trained to classify faults based on their unique signatures [7, 8, 15]. Furthermore, unsupervised learning methods, especially anomaly detection algorithms like One-Class SVM or Autoencoders [29], offer a valuable approach for detecting unexpected deviations from normal operational behavior, potentially identifying novel or incipient fault conditions without requiring pre-labeled fault data.
- **Charging Management:** Optimizing the charging process involves a complex trade-off between charging speed, efficiency, safety, and minimizing long-term battery degradation. Reinforcement Learning (RL) stands out as a promising approach, allowing agents to learn optimal, state-dependent charging policies through

interaction with the battery (or a model of it), potentially outperforming predefined charging protocols, especially in adapting to varying conditions or battery health. Other ML techniques, such as regression models, might also assist in predicting optimal charging parameters or end-of-charge points.

- **Cell Balancing and Thermal Management:** While perhaps less extensively covered in broad AI/BMS reviews compared to state estimation and FDD, ML techniques offer potential improvements in these areas as well. ML models can be applied to optimize cell balancing strategies, for instance, by predicting the optimal balancing currents based on cell states [22 - *Verify XAI relevance or note as future work area*]. Similarly, for thermal management, ML can enhance performance by enabling more accurate prediction of internal temperature distributions based on limited sensor data or by learning optimal control strategies for cooling/heating systems directly from operational data [30].

Despite these performance advantages and diverse applications, the inherent complexity of many high-performing AI/ML models often results in a lack of transparency ('black-box' behavior), making it difficult to trust their outputs in safety-critical scenarios. This crucial limitation directly motivates the integration and study of Explainable AI (XAI) techniques within BMS, which will be discussed further in subsequent sections.

4 Explainable AI (XAI) Techniques Relevant to BMS

A diverse array of XAI techniques exists, aiming to provide insights into complex AI/ML models used in BMS [11, 12]. These methods vary in their scope (local vs. global), applicability (model-agnostic vs. model-specific), and the type of explanation they provide. Selecting the appropriate technique depends on the specific BMS task, the underlying AI/ML model, the type of data, and the target audience for the explanation. **Table 3** summarizes key XAI techniques relevant to BMS applications, which are discussed in more detail below.

Table 3 Overview of XAI Techniques Applicable to BMS

Category	Specific Method	Scope	Output Type	Key Strength	Key Weakness / BMS Consideration	Refs
Model-Agnostic Local	LIME	Local	Feature importance scores (for one prediction)	Intuitive, easy to apply to any model	Explanation instability, local fidelity only	[16]
	SHAP	Local/ Global	Feature attributions (Shapley values)	Theoretically grounded, consistent, provides global summary	Computationally expensive, feature independence assumption	[17]
Model-Agnostic Global	Permutation Feature Importance	Global	Feature importance scores (model-wide)	Simple concept, widely applicable	Can be misleading with correlated features, costly	
	PDP / ALE	Global	Plots showing feature effect on prediction	Visualizes feature impact	Max 1-2 features, assumes independence (PDP)	
Model-Specific	Attention Mechanisms	Local/ Global	Attention weights/maps	Built-in interpretability for sequence models	Correlation, not causation; interpretation can be tricky	[18]
	Integrated Gradients / LRP	Local	Feature/Input attribution scores	Applicable to deep networks	Requires model internals, gradient issues possible	
Inherently Interpretable	Linear/Logistic Regression	Global	Coefficients, p-values	Simple, easy to understand	May underfit complex battery dynamics	[9]
	Decision Trees / Rule Lists	Global	Rules, Tree structure	Explicit logic, human-readable	Can become complex, prone to overfitting	[9]

4.1 Model-Agnostic Methods

These methods treat the AI/ML model as a black box and can be applied to any model type, which is advantageous given the variety of models used in BMS.

- **LIME (Local Interpretable Model-agnostic Explanations):** LIME explains an individual prediction by learning a simple, interpretable model (e.g., linear regression) in the local neighborhood of the instance being explained [16]. It works by generating perturbations (small variations) of the input instance, obtaining the black-box model's predictions for these perturbations, and then fitting a weighted, interpretable model to this local data. The coefficients or rules of this local model serve as the explanation, indicating the importance of each feature for that specific prediction. *Strength:* Its intuition and applicability to virtually any ML model make it a popular starting point for interpretability. *Weakness/BMS Consideration:* Explanations primarily offer local fidelity and might not represent the model's global behavior accurately. The definition of the 'neighborhood' and the perturbation strategy can significantly affect the explanation's stability. Applying LIME effectively to time-series data, common in BMS, requires careful consideration of how to generate meaningful temporal perturbations.
- **SHAP (SHapley Additive exPlanations):** Based on cooperative game theory concepts, SHAP assigns a unique contribution value (Shapley value) to each feature for a specific prediction, ensuring properties like local accuracy and consistency [17]. It calculates the marginal contribution of a feature by considering its effect across all possible combinations of other features. SHAP provides rich local explanations (e.g., force plots visualizing contributions) and powerful global explanations by aggregating Shapley values across many instances (e.g., summary plots showing overall feature importance and impact direction). *Strength:* Offers a strong theoretical foundation based on Shapley values, providing consistent and reliable local and global explanations. *Weakness/BMS Consideration:* The main drawback is its computational expense, which can be significant for models with many input features (common in BMS with multiple sensors and time steps) or complex internal structures. While model-agnostic in principle, efficient implementations often exist for specific model types (e.g., TreeSHAP). Additionally, interpreting SHAP values requires understanding the concept of feature contributions relative to a baseline expectation. Some SHAP implementations might assume feature independence, which needs careful consideration for potentially correlated battery sensor data.
- **Global Methods (Feature Importance, PDP/ALE):** Beyond SHAP's global summary, other methods provide global insights. Permutation Feature Importance measures a feature's overall importance by quantifying the drop in model performance when that feature's values are randomly shuffled across the dataset. Partial Dependence Plots (PDP) and Accumulated Local Effects (ALE) plots aim to visualize the average marginal effect of one or two features on the model's predictions. *Strength:* These methods offer a high-level overview of which features matter most globally or how features generally influence predictions. *Weakness/BMS Consideration:* Permutation importance can be misleading for highly correlated features. PDP assumes feature independence, while ALE attempts to address this but is more complex. Both PDP and ALE are typically limited to visualizing the effects of only one or two features at a time.

4.2 Model-Specific Methods

These methods are designed for particular classes of models, often leveraging their internal architecture.

- **Attention Mechanisms:** Widely used in sequence models like Transformers and some RNN variants (e.g., LSTM with attention), attention layers learn to dynamically weight different parts of the input sequence (e.g., specific time steps or sensor channels) when generating an output [18]. These learned attention weights can be visualized (e.g., as heatmaps) to infer which input segments the model considered most important for its prediction. *Strength:* Provides interpretability directly integrated within the model architecture, naturally suited for time-series data prevalent in BMS state estimation and RUL prediction. *Weakness/BMS Consideration:* Attention weights highlight model focus or correlation, but do not necessarily equate to causal feature importance. Interpreting complex, multi-head attention patterns can still be non-trivial.
- **Gradient-based / Propagation-based Methods:** Techniques like Integrated Gradients, DeepLIFT, or Layer-Wise Relevance Propagation (LRP) are primarily used for deep neural networks. They work by back-propagating gradients or relevance scores from the output layer back to the input features, thereby attributing the prediction to specific input elements. *Strength:* Can provide fine-grained, pixel-level, or input-element-level attributions for deep learning models. *Weakness/BMS Consideration:* These methods require access to the model's internal structure and gradients. They can sometimes be sensitive to implementation choices, and their theoretical underpinnings or interpretation can be complex (e.g., handling gradient saturation).

4.3 Inherently Interpretable Models

This approach prioritizes transparency by using models whose decision-making process is directly understandable, rather than explaining a complex black box post-hoc [9].

- **Linear Models / GAMs:** Linear and logistic regression models offer straightforward interpretability through their coefficients, which directly indicate the weight and direction of each feature's influence. Generalized Additive Models (GAMs) extend this by allowing non-linear relationships for individual features while maintaining additivity, offering a balance between flexibility and interpretability. *Strength:* Simple, transparent, well-understood relationship between inputs and outputs. *Weakness/BMS Consideration:* Their inherent linearity or additivity might limit their ability to capture the highly complex, non-linear, and interactive dynamics often governing battery behavior and degradation, potentially leading to lower predictive accuracy compared to more complex models.
- **Decision Trees / Rule Lists:** These models make predictions using an explicit set of hierarchical rules (trees) or a list of rules. The path leading to any prediction can be easily followed and understood. *Strength:* Provide highly transparent, human-readable logic. *Weakness/BMS Consideration:* Single decision trees can be unstable (small data changes can lead to different trees) and prone to overfitting complex datasets. While ensemble methods like Random Forests or Gradient Boosted Trees significantly improve predictive performance, they sacrifice the inherent interpretability of single trees, often requiring model-agnostic XAI techniques like SHAP for explanation.

The selection of an XAI technique for a specific BMS application should carefully consider the trade-offs between the desired level and type of explanation, the complexity of the underlying AI/ML model, the nature of the battery data, computational constraints (especially for on-board BMS), and the needs of the end-user interpreting the explanation. Often, employing a combination of different XAI techniques can provide a more robust and multifaceted understanding of the AI/ML model's behavior.

5 XAI Applications in Specific BMS Tasks

Having outlined the key BMS tasks and relevant XAI techniques, this section reviews how XAI has been specifically applied in the literature to interpret AI/ML models across different BMS functions. The goal is to understand the types of insights gained and the common practices in the field. ***Table 4*** *provides a high-level summary of reported XAI applications and findings for core BMS tasks, based on the reviewed literature. The subsequent subsections elaborate on these applications.*

Table 4 Summary of XAI Applications and Findings in BMS Tasks (Illustrative based on Literature)

BMS Task	Common AI/ML Models Used (Examples)	Applied XAI Techniques (Examples)	Key Insights/Findings from XAI (Examples from Literature)
State Estimation (SOC)	NN, LSTM, GPR	LIME, SHAP, Feature Importance	Identification of key input features (voltage, current, temp.) under different conditions; Understanding model reliance shifts.
State Estimation (SOH)	LSTM, GPR, Ensemble	SHAP, Feature Importance, LIME	Highlighting influential features (e.g., voltage curve shapes, ICA peaks, impedance); Understanding feature interactions for degradation.
State Estimation (RUL)	LSTM, Transformer, CNN	Attention Mechanisms, SHAP	Identifying critical historical cycles or stress events influencing prediction; Temporal focus of the model.
Fault Detection & Diagnosis (FDD)	CNN, RF, SVM, Autoencoders	SHAP, LIME, Decision Trees	Pinpointing specific sensor readings (voltage drops, temp spikes) indicative of faults; Revealing model decision rules; Anomaly explanation.
Charging Management	RL, Optimization Algos.	SHAP (on policy/value nets), Feature Importance	Understanding trade-offs learned by RL agents (speed vs. health); Identifying factors influencing optimal charging decisions. (Less common)

Cell Balancing	ML Classifiers/Regressors	Feature Importance	Determining factors influencing optimal balancing current prediction. (Less common)

5.1 XAI for State Estimation (SOC, SOH, RUL)

Explainable AI techniques are increasingly applied to demystify the complex AI/ML models used for estimating critical battery states, thereby enhancing trust and providing deeper insights into battery behavior. The literature reveals several common approaches and findings:

- **State of Health (SOH) Estimation:** Understanding SOH is vital for assessing battery degradation. Studies applying XAI often focus on identifying which input features, derived from voltage, current, and temperature profiles during operation or specific tests (like charging), are most indicative of capacity fade or impedance increase according to the ML model. For instance, Li et al. [26] used SHAP with an LSTM network, demonstrating how features from the charging voltage plateau and their interactions significantly influenced SOH predictions. Such analyses help validate if the model learns physically meaningful degradation indicators and understand the relative importance of different operational phases or derived health features.
- **State of Charge (SOC) Estimation:** For SOC estimation, especially under dynamic conditions where traditional methods struggle, XAI helps understand how ML models adapt. Techniques like LIME can provide instance-specific explanations, revealing how a model might dynamically shift its reliance between voltage-based estimation and current integration depending on factors like temperature or recent load history, as exemplified by Wang et al. [27] using LIME with a neural network. This offers transparency into the model's adaptive behavior in specific, potentially challenging, operational moments, building confidence in its robustness.
- **Remaining Useful Life (RUL) Prediction:** Predicting RUL inherently involves long-term dependencies. XAI methods, particularly attention mechanisms integrated within sequence models like LSTMs or Transformers, are valuable here. They allow researchers to visualize which parts of the battery's historical usage data the model focused on most when making its RUL prediction. Ren et al. [28] utilized attention in a Transformer network to effectively highlight specific cycles or stress events (e.g., high C-rate periods) that heavily influenced the end-of-life forecast, linking past usage to predicted lifespan and potentially informing usage recommendations for extending life.

Overall, applying XAI to state estimation models allows researchers to move beyond accuracy metrics, validate model reasoning against domain expertise, and potentially discover new indicators of battery state and health learned implicitly by the models. This fosters greater understanding and trust in these critical estimations.

5.2 XAI for Fault Detection and Diagnosis (FDD)

Given the safety-critical nature of fault detection, XAI plays a crucial role in validating and trusting FDD models. Key applications synthesized from the literature include:

- **Identifying Fault Indicators:** Post-hoc methods like SHAP are frequently used to determine which specific sensor readings or derived features contribute most to a fault classification made by models like CNNs or Random Forests. Studies often confirm that XAI highlights expected physical indicators, such as significant voltage drops, rapid temperature increases, or deviations in cell-to-cell consistency for faults like internal short circuits [19]. This helps confirm the model is learning relevant physics.
- **Understanding Model Logic:** Inherently interpretable models, such as Decision Trees or rule-based systems, provide explicit diagnostic rules that can be directly examined and compared with engineering knowledge or established diagnostic procedures [20]. This offers a high degree of transparency, although potentially at the cost of some predictive performance compared to complex models.
- **Explaining Anomaly Detection:** For unsupervised anomaly detection models used to flag unexpected behavior or sensor faults [29], XAI techniques like LIME [16] or SHAP [17] can be applied (sometimes with modifications) to attempt explaining why a particular data point was flagged as anomalous, potentially pointing towards the specific deviating sensor or unusual pattern, aiding in root cause analysis. Comparing explanations from different methods [Discussion needed based on literature] can also help assess the robustness of the findings for unexpected events.

Synthesizing findings from various studies [7, 8, 15, 19, 20, and others] reveals that XAI significantly aids in verifying that FDD models are learning correct fault signatures and provides valuable diagnostic insights beyond a simple fault flag, increasing confidence in automated diagnostic systems.

5.3 XAI for Charging Management & Other Tasks

The application of XAI to other BMS tasks like charging optimization and cell balancing is currently less mature but holds significant potential for improving transparency and user acceptance:

- **Charging Management:** As RL agents are developed to find optimal charging strategies that balance speed and battery health [21 - Verify XAI relevance...], XAI techniques (e.g., applying SHAP to the RL policy or value network) could be used to understand the complex trade-offs learned by the agent (e.g., why it reduces current at a certain SOC or temperature) and the factors driving its charging decisions under different states. This could lead to more trustworthy adaptive charging systems that users understand and accept.
- **Cell Balancing:** For ML models predicting optimal balancing currents or control actions [22 - Verify XAI relevance...], feature importance analysis could reveal the key cell parameters (voltage, temperature, estimated resistance) influencing the balancing decisions, helping to validate and potentially refine the balancing strategy.

While dedicated XAI studies in these specific areas are still emerging, the principles applied in state estimation and FDD are transferable. Future work will likely see increased use of XAI to interpret complex control strategies learned by AI for charging, balancing, and thermal management, making these optimizations more transparent and reliable.

6 Challenges and Open Issues for XAI in BMS

Despite the growing interest and potential benefits, the practical deployment and widespread adoption of XAI techniques in real-world BMS face several significant challenges and open research questions [11, 13]. These hurdles need to be addressed to fully realize trustworthy and effective explainable battery management. **Table 5** summarizes the key challenges discussed in this section.

Table 5 Major Challenges and Open Issues for XAI in BMS Applications

Challenge Area	Description	Key Implications for BMS	Potential Mitigation Strategies
Real-time Constraints	High computational cost of many XAI methods (e.g., SHAP sampling).	Difficulty in generating on-the-fly explanations on resource-limited BMS hardware.	Lightweight XAI, approximate methods, hardware acceleration, offline analysis.
Data Scarcity and Quality	Lack of diverse, labelled, high-quality fault and degradation data.	Explanations may be unreliable, overfit to limited data, or fail on unseen scenarios.	Data augmentation (Generative AI), transfer learning, physics-informed methods.
Evaluation of Explanations	Difficulty in objectively measuring explanation quality (fidelity, robustness, utility).	Hard to validate explanation correctness and ensure reliability for critical decisions.	Domain-specific metrics, user studies, counterfactual evaluation, benchmarking.
Human Interpretation & Usability	Explanations may be complex or not tailored to the end-user's needs/expertise.	Risk of misinterpretation, over-trust, or under-trust; lack of actionable insights.	Human-centered XAI design, adaptive explanations, user training, clear visualization.
Fidelity vs. Performance Trade-off	Post-hoc explanations might not perfectly reflect complex model reasoning; interpretable models might lack accuracy.	Dilemma between using high-performing black boxes vs. transparent but potentially simpler models.	Hybrid approaches, research into high-fidelity explanations, risk assessment.
Standardization & Regulation	Lack of standard formats, protocols, and regulatory guidelines for XAI in BMS.	Difficulty in comparing methods, ensuring consistency, and certifying XAI for safety use.	Development of BMS-specific XAI standards, collaboration with regulatory bodies.

6.1 Real-time Constraints and Computational Cost

Many powerful XAI methods, particularly model-agnostic techniques like SHAP that rely on extensive sampling or perturbation, entail significant computational overhead [23]. Generating explanations for complex models operating on high-frequency sensor data can be time-consuming, often taking seconds or even minutes per explanation depending on the method, model complexity, and data dimensionality. This poses a major challenge for on-board BMS applications, which typically run on resource-constrained microcontrollers with limited processing power (MHz range) and memory (KB or MB range). Generating real-time, on-the-fly explanations needed for immediate diagnostic alerts or adaptive control adjustments is often infeasible with current sophisticated XAI methods. Therefore, significant research is needed into developing lightweight XAI algorithms specifically designed for embedded systems, computationally cheaper approximation techniques (e.g., optimized or selective SHAP variants), strategies for offline explanation generation targeting common scenarios, or exploring hardware-software co-design involving dedicated accelerators for XAI computations within the BMS architecture [See Section 7].

6.2 Data Scarcity and Quality

The performance and reliability of both the underlying AI/ML model and the XAI technique applied to it heavily depend on the quality, quantity, and diversity of the training data [10, 13]. As previously noted, obtaining comprehensive, accurately labeled datasets covering various battery chemistries, form factors, operating conditions (temperature, load profiles), fault types (including incipient and combined faults), and degradation stages is extremely challenging in the battery domain. Models trained on limited or biased data may learn spurious correlations or fail to generalize to unseen conditions. Consequently, XAI methods applied to such models might highlight irrelevant features, provide misleading explanations, or generate explanations that are not robust when applied to slightly different operational scenarios. Explanations derived from models trained primarily on lab data might not hold true in real-world applications with different noise levels and unmodeled dynamics. Strategies like generative AI for data augmentation, transfer learning from simulation to reality or across different battery types, and incorporating physics-based constraints into both the ML model and the XAI process are potential avenues to mitigate these data limitations, but require further research and validation.

6.3 Evaluation of Explanations

Evaluating the 'goodness' or 'quality' of an explanation generated by an XAI method is notoriously difficult and remains a significant open research area [24]. Unlike standard ML model evaluation based on predictive accuracy metrics (e.g., accuracy, F1-score), there is often no objective "ground truth" against which to compare an explanation. Key facets of explanation quality need consideration:

- *Fidelity:* How accurately does the explanation reflect the model's actual internal reasoning or decision boundary? A persuasive explanation might not be faithful to the model.
- *Robustness:* How stable is the explanation? Small, insignificant perturbations to the input should ideally not lead to drastically different explanations [23].
- *Understandability/Interpretability:* Is the explanation clear, concise, and easily comprehensible to the intended human user?
- *Usefulness/Actionability:* Does the explanation help the user achieve a specific goal, such as debugging the model, making a decision, or learning about the domain? Developing quantitative metrics and standardized benchmarks specifically for evaluating XAI methods in the context of BMS tasks is crucial. This could involve assessing alignment with known battery physics, comparing explanations to expert knowledge, conducting user studies to measure understandability and usefulness, or using counterfactual analysis to test explanation validity.

6.4 Human Interpretation and Usability

Ultimately, the value of an explanation lies in its correct interpretation and effective use by a human [25]. The target users of XAI in BMS can range from battery researchers and design engineers to field technicians, vehicle operators, or even fleet managers, each with different levels of expertise and different information needs. Explanations generated by XAI algorithms, such as high-dimensional SHAP value plots or complex decision tree structures, may not be inherently understandable to all users. Presenting explanations in a format tailored to the user's background and the specific task context is critical for effective communication. Furthermore, human cognitive biases can influence how explanations are perceived, potentially leading to over-trust in plausible but incorrect explanations or under-trust due to information overload or lack of clarity. Research in human-centered XAI, focusing on user studies within the BMS domain, developing adaptive and context-aware explanation interfaces, and designing effective visualizations, is essential to ensure that XAI outputs translate into genuinely improved understanding and decision-making.

6.5 Fidelity vs. Performance Trade-off

An ongoing debate in the XAI community revolves around the choice between explaining complex, high-performance black-box models post-hoc versus using inherently interpretable models from the outset [9]. Post-hoc explanations (like LIME, SHAP) aim to approximate the reasoning of potentially very accurate but opaque models. However, there is a risk that the explanation itself might lack perfect fidelity – it might be a plausible simplification that doesn't fully capture the true, complex internal logic of the black box, especially regarding feature interactions. Misleading explanations, even for accurate models, can be detrimental in safety-critical systems. Conversely, inherently interpretable models (like linear models, GAMs, simple decision trees) offer complete transparency, but their simpler structure may limit their predictive power on complex, non-linear BMS tasks, potentially sacrificing accuracy. Navigating this trade-off requires careful consideration of the specific application's requirements for both performance and trustworthiness. Hybrid approaches combining model-based insights with interpretable ML, or research into verifying the fidelity of post-hoc explanations, are important directions.

6.6 Standardization and Regulation

The field of XAI, particularly its application in specific domains like BMS, currently lacks widely accepted standards. There is a need for standardization in several areas: terminology, methodologies for generating explanations, formats for presenting explanations, and protocols for evaluating explanation quality. Such standards would facilitate better comparison between different XAI techniques and promote consistency and reproducibility in research and development. Furthermore, as AI/ML systems become more integrated into safety-critical automotive and energy storage applications, regulatory bodies may start requiring evidence of system transparency and reliability, potentially including specific requirements for the explainability of AI components within certified BMS. Establishing clear industry guidelines and potentially certification processes for XAI in battery management will be an important factor for building public trust and enabling widespread, responsible adoption.

Addressing these multifaceted challenges is paramount for transitioning XAI in BMS from a promising research area to a set of reliable, practical tools deployed in real-world battery systems.

7 Future Research Directions

Addressing the challenges outlined in the previous section requires concerted research efforts across multiple fronts. Based on the current state-of-the-art and identified gaps, several promising future research directions emerge for advancing the field of XAI in BMS:

- **Lightweight and Real-time XAI Algorithms:** Given the strict computational constraints of on-board BMS hardware, a critical need exists for developing XAI techniques that are significantly more efficient than current methods like full SHAP. Research should focus on lightweight algorithms specifically designed for embedded systems, computationally cheaper approximation strategies (e.g., efficient sampling for SHAP, simplified LIME variants), model distillation techniques applied to explanations (training a simpler model to mimic the complex model's explanations), and methods tailored for edge computing environments. The goal is to enable near real-time explanation generation without compromising the BMS's primary control functions, potentially allowing for adaptive control based on explained predictions.
- **Physics-Informed Explainable AI (PIXAI):** Integrating domain knowledge from battery electrochemistry and physics can significantly enhance the reliability and meaningfulness of XAI explanations. Future work should explore PIXAI approaches where physical laws (e.g., conservation laws, thermodynamic principles), known constraints (e.g., voltage bounds, monotonic degradation trends), or simplified physical models (e.g., equivalent circuit parameters) are incorporated into the AI/ML model training process (as regularizers or hybrid architectures) or directly into the XAI method itself (e.g., constraining explanations to be physically plausible). This could lead to explanations that are not only data-driven but also physically consistent and robust, preventing spurious correlations and improving trust, especially when data is scarce.
- **XAI with Limited and Imbalanced Data:** The persistent challenge of data scarcity, particularly for rare faults or specific operating corners, necessitates research into XAI methods that perform reliably in low-data regimes. This includes exploring how generative models can be used not just for data augmentation to improve model accuracy, but specifically for generating diverse scenarios that enhance explanation robustness. Furthermore, applying techniques like transfer learning (leveraging knowledge from simulations, different battery chemistries, or related domains), few-shot learning, and meta-learning in the context of *explainable* battery models warrants investigation. The aim is to generate reliable explanations even for rare events or newly deployed battery technologies where extensive historical data is unavailable.

- **Quantitative and Domain-Specific Evaluation Metrics:** Moving beyond qualitative assessments requires the development of objective, quantitative metrics tailored for evaluating XAI methods within the specific context of BMS. Research is needed to define and validate metrics that measure different facets of explanation quality, such as fidelity (how well the explanation reflects the model's behavior), robustness (stability against minor input changes), understandability (measured via user studies), and task-based utility (how much the explanation improves human performance on a relevant task, like fault diagnosis). These metrics could involve comparing XAI outputs against physics-based simulations, checking consistency with known degradation mechanisms, or assessing alignment with expert judgments encoded in knowledge graphs. Establishing benchmark datasets and standardized evaluation protocols is also crucial for objective comparison of different XAI techniques.
- **Human-Centered XAI for BMS Stakeholders:** Explanations are ultimately consumed by humans, whose background and needs vary widely. Future research must adopt a human-centered design approach. This involves identifying the specific explanatory needs of different BMS stakeholders (e.g., design engineers requiring deep model insights vs. field technicians needing quick diagnostic pointers vs. EV drivers needing understandable safety alerts). User studies are needed to evaluate the effectiveness of different explanation formats (e.g., feature attributions, rules, counterfactuals, visualizations) for different users and tasks. Developing adaptive interfaces that tailor the complexity and format of explanations based on the user and context, and studying how to effectively calibrate user trust through explanations, are key areas for investigation [25, 31].
- **Hardware-Software Co-design for Efficient XAI:** Enabling real-time XAI on resource-constrained BMS platforms may necessitate synergistic hardware and software optimization. Research into hardware-software co-design could explore developing specialized hardware accelerators (e.g., using FPGAs, ASICs, or neuromorphic chips) optimized for common XAI computations (like perturbation-based methods or gradient calculations). Simultaneously, software optimizations, including applying model compression techniques (like quantization and pruning) not just to the primary ML model but also to the explanation generation process itself, could significantly reduce computational demands.
- **Standardization and Regulatory Frameworks:** As AI/XAI becomes more integral to safety-critical systems like automotive BMS or grid energy storage, establishing industry standards and clear regulatory guidelines will be essential for ensuring safe and responsible deployment. Future efforts should involve collaboration between researchers, battery manufacturers, automotive OEMs, energy companies, standards organizations (like ISO, IEC, SAE), and regulatory bodies. This collaboration is needed to develop standardized terminology, common formats for reporting XAI results, protocols for validating the performance and reliability of explainable systems, and potentially certification requirements for AI/XAI components used in safety-critical BMS applications.

These research directions are often interconnected and highlight the need for interdisciplinary collaboration between AI/ML experts, battery scientists and engineers, control systems engineers, hardware designers, human-computer interaction researchers, and policymakers to achieve truly trustworthy, effective, and widely adopted explainable battery management systems.

8 Conclusion

The integration of Artificial Intelligence and Machine Learning has undeniably advanced the capabilities of Battery Management Systems, offering enhanced performance in critical tasks such as state estimation and fault diagnosis. However, the prevalent use of complex, opaque models introduces significant concerns regarding trustworthiness, reliability, and safety, particularly given the safety-critical nature of battery applications in electric vehicles and large-scale energy storage. Explainable AI (XAI) emerges as a crucial enabler to bridge this gap, providing the necessary tools and methodologies to foster transparency and understanding in intelligent BMS.

This paper has provided a comprehensive review of the current landscape of XAI applications within the domain of BMS. We surveyed the state-of-the-art literature, examining how various XAI techniques – including model-agnostic methods like LIME and SHAP, model-specific approaches like attention mechanisms, and inherently interpretable models – are being employed across core BMS functions. Our review highlighted that XAI is increasingly utilized not only for fault diagnosis, where it helps identify key fault indicators and validate model reasoning, but also for state estimation (SOC, SOH, RUL), offering deeper insights into the factors driving battery degradation and state changes as learned by AI/ML models. Early applications in charging management and other areas also show promise.

Despite these advancements, significant challenges impede the widespread practical deployment of XAI in BMS. As discussed, these include the computational demands of many XAI methods conflicting with real-time BMS constraints, the scarcity of diverse and high-quality battery data impacting explanation reliability, the inherent difficulty in objectively evaluating explanation quality, ensuring human interpretability for various stakeholders, navigating the trade-off between model performance and interpretability, and the lack of standardization and regulatory frameworks.

Addressing these challenges through dedicated research, as outlined in the future directions – including developing lightweight and physics-informed XAI, establishing robust evaluation metrics, adopting human-centered design principles, pursuing hardware-software co-design, and fostering standardization – is paramount. Continued interdisciplinary collaboration will be key to advancing this field.

In conclusion, Explainable AI is poised to play an indispensable role in the future of battery management. By rendering complex AI/ML models transparent and interpretable, XAI not only enhances trust and facilitates debugging but also unlocks deeper understanding of battery behavior and failure mechanisms. Moving forward, the systematic integration of robust, efficient, and user-centric XAI solutions will be essential for developing the next generation of intelligent BMS that are not only high-performing but also demonstrably safe, reliable, and trustworthy. The pursuit of explainability is fundamental to ensuring the responsible and effective deployment of AI in critical energy storage technologies.

References

[1] Dunn B, Kamath H, Tarascon JM. Electrical energy storage for the grid: a battery of choices. Science. 2011;334(6058):928-35.

[2] Lu L, Han X, Li J, Hua J, Ouyang M. A review on the key issues for lithium-ion battery management in electric vehicles. J Power Sources. 2013;226:272-88.

[3] Andrea D. Battery Management Systems for Large Lithium-Ion Battery Packs. Artech House; 2010.

[4] Vidal C, Malysz P, Kollmeyer P, Emadi A. Machine Learning Applied to Electrified Vehicle Battery State of Charge and State of Health Estimation: State-of-the-Art. IEEE Access. 2020;8:52796-814.

[5] Chemali E, Kollmeyer PJ, Preindl M, Emadi A. Long Short-Term Memory Networks for Accurate State-of-Charge Estimation of Li-ion Batteries. IEEE Trans Ind Electron. 2018;65(8):6730-9.

[6] Berecibar M, Gandiaga I, Villarreal I, Omar N, Van Mierlo J, Van den Bossche P. Critical review of state of health estimation methods of Li-ion batteries for real applications. Renewable Sustainable Energy Rev. 2016;56:572-87.

[7] Samanta A, Chowdhuri S, Williamson SS. Machine learning-based data-driven fault detection/diagnosis of lithium-ion battery: A critical review. Electronics. 2021;10(11):1309.

[8] Zhao R, Liu J, Gu J. Fault diagnosis of lithium-ion battery packs based on voltage anomaly detection: A data-driven approach. Renewable Sustainable Energy Rev. 2017;75:1309-19.

[9] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-215.

[10] Li W, Sengupta N, Dechent D, Howey D, Annaswamy A, Sauer DU. Data-driven battery health monitoring: Challenges and opportunities. Joule. 2020;4(12):2587-92.

[11] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115.

[12] Adadi A, Berrada M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access. 2018;6:52138-60.

[13] Faraji Niri M, Aslansefat K, Haghi S, Hashemian M, Daub R, Marco J. A review of the applications of explainable machine learning for lithium–ion batteries: from production to state and performance estimation. Energies. 2023;16(17):6360.

[14] He H, Xiong R, Fan J. Model-based methods for battery states estimation in electric vehicles: Current status and future trends. J Power Sources. 2012;198:296-308.

[15] Jia Y, Li J, Yao W, Li Y, Xu J. Precise and fast safety risk classification of lithium-ion batteries based on machine learning methodology. J Power Sources. 2022;548:232064.

[16] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD'16). 2016. p. 1135-1144.

[17] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. p. 4765-4774.

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017.

[19] Yang D, Wang Y, Pan R, Chen R, Chen Z. Early fault diagnosis of lithium-ion batteries using convolutional neural network and Shapley additive explanations. Energy. 2022;244:123190.

[20] Xu C, Li L, Xu Y, Han X, Zheng Y. A vehicle-cloud collaborative method for multi-type fault diagnosis of lithium-ion batteries. eTransportation. 2022;12:100172.

[21] Liu T, Tan X, Rui X, Li H, Xu S. Optimal charging of lithium-ion batteries using deep reinforcement learning with consideration of battery degradation. Energy. 2022;239:122099. (Verify XAI relevance or note as future work area)

[22] Ma K, Wang C, Zhang C, Xu G. Machine learning-based active balancing method for lithium-ion battery packs. Journal of Energy Storage. 2021;42:103065. (Verify XAI relevance or note as future work area)

[23] Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. ICML Workshop on Human Interpretability in Machine Learning (WHI). 2018.

[24] Mohseni S, Zarei N, Ragan ED. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems (TiiS). 2021;11(3-4):1-45.

[25] Liao QV, Gruen D, Miller T. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. ACM Computing Surveys (CSUR). 2021;54(9):1-37.

[26] Li Y, Liu K, Foley AM, Zülke A, Berecibar M, Nanini-Maury E, Van Mierlo J, Hoster HE. SHAP-based interpretable state-of-health estimation for lithium-ion batteries using LSTM network. Energy and AI. 2023; 11:100209.

[27] Wang S, Takyi-Aninakwa P, Jin S, Yu C, Fernandez C. Interpretable Lithium-Ion Battery State of Charge Estimation Using LIME Tuned Neural Networks. IEEE Transactions on Transportation Electrification. 2022; 8(4):4787-4796.

[28] Ren L, Zhao L, Hong S, Zhao S, Wang H, Zhang L. Remaining useful life prediction for lithium-ion batteries based on a transformer network with attention mechanism. Reliability Engineering & System Safety. 2021; 212:107622.

[29] Zhang Y, Xiong R, He H, Pecht MG. Lithium-ion battery fault diagnosis based on autoencoder and restricted Boltzmann machine. Journal of Energy Storage. 2020; 31:101697.

[30] Panchal S, Khasow R, Dincer I, Agelin-Chaab M, Fraser R, Fowler M. Machine learning based thermal model for lithium-ion battery packs. Journal of Power Sources. 2021; 485:229312.