IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

XAI-Based Fault Diagnosis In Multi-Cell Battery Packs

[1]JongMyoung Kim

[1]Professor

[1]Department of Artificial Intelligence and Big Data, Sehan University, South Korea

Abstract: Battery packs in electric vehicles and energy storage systems can suffer various faults that pose safety hazards, including cell internal short circuits (ISC), thermal runaway (TR), sensor failures, and battery management system (BMS) malfunctions. Data-driven machine learning methods have shown promise in detecting such faults early, but their black-box nature raises trust and safety concerns. Explainable Artificial Intelligence (XAI) techniques have emerged to bridge this gap by providing insight into model decisions. This review surveys recent research on XAI-based fault diagnosis in multi-cell lithium-ion battery packs. We discuss traditional model-based approaches (using first-principles battery models and state observers) versus data-driven approaches (machine learning classifiers and anomaly detectors), highlighting how XAI methods can augment the latter. Key XAI techniques including feature attribution methods like SHAP and LIME, attention-based deep learning models, and surrogate modeling are explained and compared. Recent studies have applied XAI to identify the sensor signals and features most indicative of faults (e.g. sudden drops in cell voltage or rises in internal resistance). We compile these findings, noting each method's strengths and limitations. Challenges such as the need for large labeled fault datasets, real-time explainability, and validation of XAI outputs are identified as open issues. Overall, XAI can enhance the transparency and reliability of data-driven fault diagnosis in multi-cell battery packs, enabling safer battery management through human-interpretable insights.

Index Terms **- Battery Fault Diagnosis, Explainable AI, Lithium-Ion Battery, Multi-Cell Battery Pack**

I. INTRODUCTION

Lithium-ion battery packs are widely used in electric vehicles (EVs) and stationary storage, where reliability and safety are paramount [1]. As these battery systems age or encounter abuse, they can develop faults that lead to performance degradation or catastrophic failures. Prompt detection and diagnosis of faults – such as BMS malfunctions, cell imbalance or degradation, internal short circuits (ISC), overcharging, over-discharging, and thermal runaway (TR) – is essential to mitigate risks and prevent accidents [1]. In multi-cell packs, faults may originate at the cell level (e.g. an ISC in a single cell) or at the pack level (e.g. a faulty sensor or cooling failure), with potentially severe consequences if left unaddressed.

Conventional fault diagnosis approaches fall into two broad categories: model-based methods and data-driven methods [2, 3]. Model-based methods use physics-based or equivalent-circuit battery models combined with observers or analytical redundancy to detect inconsistencies caused by faults [3]. For example, state observers can estimate each cell's state-of-charge or resistance, and deviations from the expected model behavior can indicate a fault. Such approaches benefit from physical interpretability and can sometimes detect certain faults early [3]. However, model-based methods require accurate models and can be sensitive to modeling errors, parameter uncertainties, and varying operating conditions [3]. They may also be computationally intensive and require extensive tuning for complex multi-cell systems [1, 3].

In contrast, data-driven approaches rely on patterns learned directly from data (e.g. voltage, current, temperature measurements) using statistical or machine learning techniques. These methods, which include signal processing algorithms, machine learning classifiers, and neural networks, can handle large amounts of data and automatically adapt to complex fault signatures [4, 5]. Machine learning has become increasingly popular for battery fault diagnosis due to its ability to capture nonlinear relationships and multivariate interactions beyond what first-principles models can easily provide [4]. Indeed, numerous studies have applied classifiers (SVM, random forest, neural networks, etc.) or anomaly detection algorithms to identify faulty batteries or predict failures [2, 6, 7]. However, a major drawback of purely data-driven models is their lack of transparency: high accuracy models often act as "black boxes," making it difficult for engineers to trust their predictions or understand the reasons behind a detected fault [4, 5]. This lack of explanation is both a practical and an ethical issue in safety-critical applications [4]. It is in this context that Explainable Artificial Intelligence (XAI) techniques have a critical role to play.

XAI refers to methods that make the decision-making process of AI models understandable to humans [8]. Instead of simply outputting a fault alarm, an XAI-enabled model can provide additional information such as which sensor readings or features most contributed to the alarm, or how different input conditions would affect the model's output. Such explanations can increase user trust, aid in root-cause analysis, and ensure the model is aligning with domain knowledge (rather than exploiting spurious correlations). In the field of battery management, XAI is especially important because of the high stakes: an incorrect fault diagnosis can lead to unsafe operation or unnecessary shutdowns, so operators need confidence and insight into automated decisions. Moreover, regulatory frameworks and safety standards increasingly demand transparency from AI systems in vehicles and energy grids.

Recent research on battery fault diagnosis has begun incorporating XAI techniques to interpret data-driven models. A 2021 critical review by Samanta et al. noted that while many machine learning approaches for battery fault detection show high performance, they often do not explain their results, making it hard to assess their reliability in real-world conditions [2]. Similarly, Faraji Niri et al. (2023) observed that despite the proliferation of data-driven battery management algorithms, only a small number of studies have focused on explainability, indicating a significant gap in the literature [9]. This review addresses that gap by surveying the state-of-the-art in XAI applications for battery fault diagnosis. We first outline the main approaches to battery fault detection and diagnosis, then discuss the XAI techniques applied in this domain – from feature attribution methods like SHAP and LIME to interpretable model architectures and surrogate modeling. We highlight representative case studies, compare the strengths and weaknesses of different methods, and identify open challenges and future research directions for XAI in battery fault diagnosis.

II. Approaches to Battery Fault Diagnosis in Multi-Cell Packs

Battery fault diagnosis methods can be broadly categorized as model-based (or physics-based) and non-model-based (data-driven) approaches [1, 3]. In practice, effective battery management often uses a combination of both. Here we outline these approaches and their relevance to explainability.

Model-Based Methods:

Model-based fault diagnosis relies on mathematical models of battery behavior. Two common model types are electrochemical models (describing the internal electrochemistry, e.g. using coupled differential equations) and electrical equivalent circuit models (ECMs), which approximate battery dynamics with RC circuits [3]. For fault diagnosis, models are typically augmented with estimation algorithms or observers (e.g. Kalman filters, Luenberger observers) to track internal states and parameters [3]. Faults are detected by monitoring residuals – deviations between measured outputs and model-predicted outputs – or by observing parameter changes beyond normal ranges. For instance, an internal short circuit can be detected by a sudden drop in estimated cell resistance or capacity in a model-based observer [6, 3]. Model-based approaches have the advantage of built-in interpretability: when a fault is declared, it is usually because a physically meaningful residual exceeded a threshold, directly indicating the type or location of anomaly (e.g. an increasing difference between cell open-circuit voltage and model estimate may point to cell capacity loss or an ISC). These methods leverage expert knowledge and are often preferred in high-trust settings. However, developing accurate battery models that capture cell-to-cell variations, temperature effects, and aging is challenging [3]. Model-based methods can also be computationally heavier and slower to respond [3]. They may fail if the actual fault deviates from assumptions in the model. In summary, model-based diagnostics are interpretable by design (the model's parameters and residuals have physical meaning) but can be limited in adaptability and scope of detectable faults.

Data-Driven Methods:

Data-driven fault diagnosis uses patterns in data to detect anomalies or classify fault types without requiring an explicit battery model. Early data-driven techniques include signal processing and statistical analysis – for example, using the correlation between cell voltages in a series pack to identify a faulty cell (a significantly lower correlation can indicate an out-of-line cell) [10]. More recently, machine learning techniques have been applied, such as supervised classifiers to identify fault types from sensor features, or unsupervised methods to flag abnormal behavior. Data-driven methods can utilize a wide range of measured variables: cell voltages, pack current, temperatures, pressure, etc., often combined into features that capture trends (voltage drops, dV/dt, entropy changes, etc.). These approaches are powerful in handling complex or subtle fault signatures and do not require detailed battery knowledge upfront. For instance, a random forest or neural network can learn to differentiate normal vs. various faulty conditions (like ISC vs. connection fault vs. sensor drift) from training data [2, 6, 7]. A key limitation, however, is that complex ML models are typically opaque – it's not immediately clear why a certain decision was made. Purely data-driven diagnostics risk overfitting to particular datasets and might pick up spurious correlations that don't generalize. Without explanations, it's hard to trust these models in safety-critical applications. This lack of transparency has been identified as a major barrier to deployment of advanced ML in battery management [2, 9]. XAI techniques aim to address this by making data-driven methods more interpretable, as we explore in the next section.

It's worth noting that the line between model-based and data-driven is increasingly blurred. Hybrid approaches exist, where data-driven models are constrained by or informed by physical models (for example, a neural network might estimate model parameters, or a diagnostic system might use a qualitative physics model to validate ML outputs) [9]. Additionally, some simpler data-driven techniques (like decision trees or linear regression models) are themselves interpretable and can be considered inherently explainable. Regardless of approach, the goal is to accurately and promptly detect faults with understanding of the reasoning – which is where XAI comes in for the more opaque methods.

III. XAI Techniques for Battery Fault Diagnosis

Explainable AI techniques provide tools to interpret and visualize the decisions of complex models. In the context of battery fault diagnosis, XAI can reveal which sensor readings or derived features most influenced a model's prediction of a fault, or how a model differentiates between fault types. This section reviews the main XAI methods applied or applicable to multi-cell battery fault diagnosis, grouped by their approach: post-hoc feature attribution, interpretable model architecture (attention mechanisms), and surrogate modeling and rule extraction. We also provide examples from recent studies for each category (summarized in Table 1).

3.1 Feature Attribution Methods (SHAP, LIME, and Feature Importance)

Feature attribution is a post-hoc XAI approach that assigns an importance value to each input feature of a model for a given prediction. In battery fault diagnosis, features could be sensor measurements (voltages, temperatures) or engineered features (e.g. voltage drop during a load pulse). By quantifying each feature's influence on the model's output, engineers can verify if the model is attending to meaningful indicators of faults.

- **SHAP (SHapley Additive exPlanations):** SHAP is a popular model-agnostic method based on cooperative game theory [11]. It calculates feature importance by considering the contribution of each feature to the prediction across many feature value combinations, using Shapley values for fair attribution. A key advantage of SHAP is that it provides consistent global and local explanations – the SHAP values for a single instance explain that particular prediction, and if we average SHAP values over many instances we get each feature's overall importance [11]. SHAP is also grounded in solid theoretical guarantees (local accuracy and consistency) [11]. In battery applications, SHAP can be applied to any trained fault detector model. For example, in a recent study on battery state prediction, SHAP was used to identify the most influential inputs affecting the model's estimation of battery health and impending failure [12]. The output might show, for instance, that an abnormal drop in one cell's voltage contributed heavily (with a large negative SHAP value) to the model's decision that an internal short had occurred. This aligns with human expectations and thus builds trust. However, SHAP can be computationally expensive for models with many features or complex correlations, as it requires evaluating numerous feature subsets. Despite this, its use in battery diagnostics is growing; Faraji Niri et al. (2023) found SHAP to be the second most frequently employed XAI method in battery research, often used alongside tree-based models [9].

- **LIME (Local Interpretable Model-Agnostic Explanations):** LIME is another widely-used technique that explains a prediction by training a simple interpretable model (such as a linear model) locally around the instance of interest [13]. In practice, LIME perturbs the input features of the instance in many ways, observes the complex model's outputs, and then fits a weighted linear regression that approximates the complex model in that local vicinity [13]. The coefficients of this local surrogate model serve as feature importance scores. In a battery fault context, LIME could be used to explain why a specific driving cycle was flagged as "thermal runaway risk" by an AI: it would generate slight variations of the sensor readings (voltage, temperature profiles) and see how the black-box model responds, then deduce which readings (perhaps a rapidly rising temperature in one module) drive the prediction. LIME's strength is its simplicity and model-agnostic nature – it can explain any classifier or regressor. It tends to be faster than SHAP but can suffer from instability (different perturbations may yield slightly different explanations). As of this review, LIME has been less commonly reported in battery fault studies compared to SHAP, but it remains a valuable tool. It provides quick, human-intelligible insights (e.g. "voltage sensor 5's reading had the highest positive weight in the local linear model, indicating that sensor was a main driver of the fault prediction"), which complements more rigorous methods.
- **Permutation and Gini Importance:** In addition to SHAP and LIME, many researchers use simpler feature importance measures, especially with tree-based models. Permutation importance involves shuffling one feature's values among instances and observing how much the model's error increases; a large increase indicates the model was relying strongly on that feature [14]. Gini importance (used in random forests) measures how much a feature reduces uncertainty or impurity across the decision splits in the tree ensemble [14]. These methods were used, for example, by Jia et al. (2022) in a fault classification model for lithium-ion cells [6]. Their machine learning model (comparing SVM, decision tree (DT), random forest (RF), etc.) could classify whether a cell was experiencing an internal short or thermal runaway with up to 95% accuracy [6]. To explain these predictions, they examined feature importances via permutation and the RF's Gini importance [6]. The results showed that features like the initial and final voltage derivative, the end-of-discharge voltage, and the voltage integral were the most significant indicators of an impending internal short circuit [6]. Such information is very useful – it tells battery engineers that the model is focusing on sensible precursors (voltage drop patterns) rather than any meaningless artifact. Similarly, Xu et al. (2022) trained a decision tree to classify multiple fault types in a battery pack (normal vs. short circuit vs. connection fault vs. capacity degradation) [7]. By examining the tree's structure, they found that it prioritized features related to the mean open-circuit voltage difference between cells and the mean internal resistance difference as top decision nodes, which contributed about 64% and 13% respectively to the model's fault decisions [7]. This aligns with intuition: a cell with an internal short will exhibit a notably lower open-circuit voltage and altered resistance compared to its peers [7]. These feature-attribution approaches (permutation, Gini importance) are less computationally intensive and easily applied to tree models, but they provide only a global sense of importance and can sometimes be misleading if features are correlated. Nonetheless, they have proven effective in validating that ML models for battery fault diagnosis are 'paying attention' to the right signals.

Overall, feature attribution XAI methods like SHAP and LIME (and simpler importance metrics) are a cornerstone of explainability in battery diagnostics. They answer the critical question: "What sensor readings or features caused the model to think the battery is faulty?" – which is invaluable for engineers performing troubleshooting or deciding on control actions (such as bypassing a suspect cell).

3.2 Interpretable Model Architectures and Attention Mechanisms

Another route to explainability is to use or design models that are more transparent by their very structure. In battery fault diagnosis, this can mean choosing inherently interpretable models (like decision trees or rule-based systems), or incorporating architectural features like attention mechanisms in neural networks to highlight important inputs.

- **Decision Trees and Rule-Based Models:** Decision tree classifiers (and their ensembles, if combined with careful interpretation) are often used because they produce human-readable rules. A decision tree might, for example, first check if any cell's voltage drops below a threshold under load (yes/no), then check the difference between the highest and lowest cell voltage, and so on – forming a logical path to a fault diagnosis. Such rules are relatively easy to interpret and verify against engineering knowledge. In multi-cell battery packs, diagnostic decision trees can be used to isolate a faulty cell or determine the fault type based on a hierarchy of symptom checks. The downside is that unconstrained decision trees

can become very large and complex when fitted to data, but techniques like limiting tree depth or using tree ensembles with global explanation tools (as described in 3.1) can mitigate this. In one case, decision trees were used to classify battery faults and then the resulting tree was examined directly to understand the model's logic [7]. The tree's splits confirmed known fault indicators (voltage and resistance disparities), giving confidence in both the model and the diagnostic result.

- **Attention Mechanisms in Neural Networks:** Attention mechanisms, widely used in sequence models, allow a neural network to weight the importance of different parts of the input when making a prediction. In battery systems, attention can be applied in time-series models (to highlight which time steps in a voltage or temperature trajectory are most informative) or across different sensors/cells (to highlight which cell in a battery pack is contributing most to an anomaly). For example, an LSTM-based model for early fault prediction could include an attention layer that produces weights for each time step, indicating where the model "attended" – perhaps strongly weighting a moment when a cell's voltage abruptly dropped. If that time aligns with a known fault initiation event, it provides an explanation for the model's prediction. Likewise, a neural network taking multiple cell measurements could employ an attention mechanism over cells, effectively identifying the cell that is likely faulty by giving it a higher attention weight. Although attention weights are not a perfect explanation (they highlight correlation, not necessarily causation), they have been successfully used to interpret deep learning models in related domains (e.g. machinery fault diagnosis) [14] and are beginning to find use in battery applications. One recent work on battery state-of-health prediction developed a self-attention neural network that not only improved accuracy but also indicated which segments of the cycle data were most predictive of battery aging [9]. Such information can indirectly point to fault-related stress periods in usage (for instance, attention might focus on high-temperature charging intervals which accelerate degradation). The advantage of attention is that it provides an internal explanatory signal without separate post-processing. However, interpreting attention requires caution: it tells us where the model looked, which is helpful, but not exactly why the model made its decision in a causal sense.
- **Inherently Interpretable Models:** Apart from trees, other inherently interpretable models include linear models, logistic regression, and case-based reasoning (k-Nearest Neighbors with explanation by similar cases). In battery diagnostics, linear models are sometimes too simplistic given nonlinear behaviors. However, a notable example is the elastic net model used by Chen et al. (2017) for on-board failure identification [15]. The elastic net is essentially a linear model with regularization that selects a subset of features. Chen et al. trained this model on partial charging curves of cells to classify whether a cell is healthy or failing due to capacity loss [15]. The model inherently performed feature selection, effectively choosing the two most relevant voltage/capacity features and ignoring the rest [15]. This yielded a simple linear decision boundary in that feature space – something that can be easily interpreted by engineers (the selected features corresponded to peaks in the incremental capacity curve, which are known indicators of cell aging) [15]. The interpretable nature of the model meant no separate XAI tool was needed; the model's coefficients directly indicated how much each feature contributed to the diagnosis. The trade-off was a slight drop in predictive performance compared to black-box models, but many would argue this is worthwhile in safety-critical diagnosis. This philosophy echoes the viewpoint of Rudin (2019), who argues that whenever possible, one should use interpretable models instead of explaining black-boxes – to avoid the risk of misleading explanations [4]. In battery fault diagnosis, if a simple rule-based algorithm or linear model achieves near-human accuracy, it is often preferable for operational use due to its transparency and ease of validation.

IV. Surrogate Modeling and Hybrid Explanations

Surrogate modeling involves creating a secondary, interpretable model that approximates the original complex model's behavior. This is similar in spirit to LIME (which uses local surrogates), but here we discuss global surrogates and other hybrid strategies combining model-based and data-driven insights.

- **Global Surrogate Models:** A global surrogate is an interpretable model (like a decision tree or a set of if-then rules) trained to mimic the predictions of a complex model over the entire input space. For instance, suppose we have a trained neural network that flags cells as faulty or healthy based on a plethora of features. We could generate a large number of examples, get the neural network's predictions, and then train a decision tree on this synthesized dataset (inputs vs. the network's outputs) to serve as a stand-in. If the tree achieves high fidelity in reproducing the neural network's decisions, it effectively becomes a concise explanation of the network. Researchers have not extensively reported

global surrogates specifically for battery fault models yet, but this approach is promising. It could yield a set of human-understandable rules like: "IF cell_5 voltage drop > X and temperature > Y THEN Fault = True" which approximate the neural net's complex boundary. The risk is that the surrogate may not perfectly capture the original model, especially if the model is highly nonlinear. Still, even a partial fidelity surrogate can provide insight. Rule-based surrogate extraction has been noted as a viable path for explaining ensemble models in related power systems, which by analogy applies to battery systems.

- **Hybrid Physics-Data Explanations:** In multi-cell battery packs, certain fault types can be more easily explained by physics, while others emerge from patterns in data. An emerging idea is to use domain knowledge to guide explanations of data-driven models. For example, if an ML model predicts an "internal short" fault, one could automatically correlate that with known symptoms such as heat generation and voltage divergence, and check if those were present. If the model's prediction is indeed based on those symptoms, the explanation can be phrased in those terms (e.g. "Cell 7 is predicted to have an internal short because its voltage dropped 0.2 V more than the pack average and its surface temperature spiked"). This kind of explanation uses a surrogate description rooted in battery physics to rationalize the black-box output. Another hybrid approach is supplementing data-driven models with estimated parameters from a battery model. For instance, one might feed features like "estimated cell resistance" or "capacity fade %" (obtained via a model-based estimation) into an ML classifier alongside raw sensor data. The classifier's use of those features can then be interpreted directly (if the classifier heavily weights a high resistance estimate, it is aligning with the physical notion that increased resistance indicates a fault). This merges interpretability with predictive power, and some studies have found it improves both performance and explainability [9, 3].
- **Case-Based Reasoning:** Although not yet prominent in battery fault literature, case-based reasoning can be intuitive: the system can retrieve similar fault cases from history to explain a new diagnosis. For example, "Cell 12 is flagged faulty because it behaved similarly to Cell 8 in Pack 3 last year, which had an internal short." This approach requires a library of past fault data and a way to measure similarity, but it provides a very human-friendly explanation by example. As more battery operational data become available, one can envision XAI systems that provide such analogical explanations.
- **Visualization Tools:** Lastly, a mention should be made of visualization-driven XAI. In battery packs, one can visualize the spatial or temporal patterns that led to a fault. Heatmaps showing cell-wise anomaly scores or time plots highlighting segments where the model found anomalies can be effective explanatory tools for engineers. For instance, a plot of all cell voltages where the faulty cell's voltage curve is highlighted at the points the model deemed problematic can immediately show when and which cell diverged. These visual explanations often accompany formal XAI methods: the SHAP values for each cell can be depicted on a pack diagram, etc. Such techniques are more supplemental but are important for practical acceptance of XAI outputs.

Table 1. Key XAI methods for battery fault diagnosis – their approach, advantages, limitations, and example applications.

Technique	Explanation Approach	Pros	Cons	Example Usage
SHAP (SHapley Additive exPlanations)	Model-agnostic feature attribution based on Shapley values (game theory). Quantifies each feature's contribution to a specific prediction [11].	Consistent, theoretically fair; provides local and global explanations; works with any model.	Computationally intensive for large models; assumes feature independence (kernel SHAP) which may not hold, requiring careful interpretation.	Identifying which sensor features indicate an internal short or thermal runaway in ML classifiers [6, 7]. Widely used to explain battery health predictions [12].
LIME (Local Interpretable Model-agnostic Explanations)	Post-hoc local surrogate modeling. Fits a simple interpretable	Fast and flexible; easily explains individual predictions in human terms (e.g.	Local fidelity only (may not reflect global behavior); explanations can vary with sampling; less	Explaining a particular fault decision by approximating the battery model's

	model (e.g. linear) around the neighborhood of an instance to explain a complex model's prediction [13].	linear weights); no need for model internals.	effective if features interact in complex ways.	logic near that data point. Not yet widely reported in battery domain, but applicable for explaining e.g. "why this cycle was flagged as faulty." [13]
Feature importance (Permutation/Gini)	Measures influence of each feature on model performance (permutation testing) or impurity reduction (for tree models) [14].	Simple to compute; gives quick global importance ranking; works out-of-the-box for tree ensembles.	Global only (no instance-specific insight); permutation can be misleading if features are correlated.	Used in Random Forest models to show voltage drop and dV/dt are top fault indicators [6, 7]. Helps validate model focus aligns with known fault symptoms.
Attention Mechanisms	Neural network layers that learn weightings for input components (time steps, sensor channels) for each output prediction.	Offers built-in indication of "where the model is looking"; can improve model performance and interpretability simultaneously.	Attention weights are not a complete explanation; can sometimes be diffuse or hard to interpret if multiple fault indicators exist.	In a battery pack sequence model, attention highlights the time segment of an anomaly (e.g. a sudden voltage drop) as the reason for fault prediction. Emerging use in deep learning for battery state forecasting [9].
Surrogate Models / Rules	Train an interpretable model (tree, rule set) to mimic the behavior of the black-box model, either globally or for specific regions.	Provides a human-readable approximation of complex model; leverages rich forms like if-then rules which align with engineering reasoning.	May sacrifice accuracy/fidelity; difficult to capture highly nonlinear behavior with simple surrogates; risk of oversimplification.	Decision tree surrogates to explain a neural net that detects faulty cells, yielding rules like "IF cell#5 voltage drop > X AND temp > Y THEN fault." Conceptually demonstrated in power system diagnostics.

As shown in **Table 1**, each XAI technique has its strengths and trade-offs. In practice, multiple methods are often used in combination to build a comprehensive picture. For example, a team diagnosing an EV battery pack might use a neural network for fault detection (for high accuracy), apply SHAP to identify the top contributing sensor signals for each alert, and also maintain a simple rule-based checklist as a sanity check (e.g. any cell with >0.1 V deviation is flagged). The SHAP explanation might reveal that a particular cell's voltage anomaly is driving the network's prediction, which matches the rule-based trigger – reinforcing confidence in the diagnosis. If SHAP instead highlighted an unexpected feature (say, an auxiliary temperature reading), engineers would know to investigate further before taking action, as it could indicate either a subtle fault mechanism or a modeling issue.

V. Discussion: Benefits, Challenges, and Future Directions

Benefits of XAI in Battery Fault Diagnosis: Incorporating XAI offers several clear advantages. Firstly, it increases operator trust in automated battery management systems. Technicians are more likely to act on a model's recommendation (such as disconnecting a module) if the model can explain its reasoning in familiar terms (e.g. pointing out a specific cell's abnormal behavior) [2, 9]. XAI also aids in root cause analysis: by knowing which features were important, engineers can trace back to physical causes (for instance, a low-voltage reading might be due to an open circuit or a failed fuse). In research and development, XAI helps model validation – it can reveal when a model is relying on spurious correlations or noise. For example, if an explanation shows a model focused on a temperature sensor that is known to be unreliable or far from the fault location, that model can be scrutinized or retrained. Overall, XAI contributes to safer battery operation by adding a layer of transparency and error-checking to the decision process. It also facilitates knowledge transfer: insights gained from XAI (like which sensor patterns herald failure) can inform improved diagnostic rules and sensor designs in next-generation BMS.

Despite these benefits, there are significant challenges and open questions in applying XAI to battery fault diagnosis:

- **Limited Fault Data and Ground Truth:** Data-driven models, and by extension XAI analyses of them, are only as good as the data they are trained on. Real battery failures (thermal runaways, internal shorts) are rare and often not reproducible in volume due to safety. This means models might be trained on simulated faults or accelerated aging data, which may not capture all real-world nuances. XAI might faithfully explain a model's reasoning on the training distribution, but if the model never saw a certain fault scenario, its "explanation" might not correspond to the true cause. A related issue is the lack of ground truth for some faults – we might know a model is detecting an anomaly, but verifying the exact fault type or cause can be difficult without destructive analysis. As a result, one future need is the creation of richer open datasets of battery fault events (including sensor logs from failed batteries) to train and evaluate explainable models [9]. Researchers are beginning to share data from battery lifecycles and abuse tests, which could be used to improve XAI methods' reliability.
- **Real-Time Constraints:** Implementing XAI in an online BMS environment has computational and usability constraints. Methods like SHAP can be too slow to run on-the-fly in an embedded processor for every prediction. Simplified or approximate XAI methods might be required, or one might pre-compute explanation templates offline. There is ongoing work on speeding up XAI algorithms and on developing lightweight surrogate models that can run in real time. Moreover, the BMS has limited user interface capabilities – any explanation must be distilled to key information (e.g. which cell is faulty) rather than overwhelming a technician with dozens of SHAP values. Balancing detail and clarity in real-time explanations is an open challenge.
- **Evaluation of Explanations**: How do we know if an explanation is correct or useful? This is a general challenge in XAI. An explanation could be persuasive yet wrong (the model might be right for the wrong reasons). For high-stakes systems like batteries, it is crucial to validate explanations. One approach is to perform controlled experiments: if the model says "cell 7's voltage triggered this alarm," one could manipulate cell 7's data and see if the alarm changes accordingly (counterfactual testing). Another is expert review – have battery engineers assess whether the explanations align with known physics. Developing quantitative metrics for explanation quality in this domain (e.g. how well they align with a first-principles simulator's diagnosis) would advance the field. The work by Guidotti et al. (2018) suggests a need to explicitly consider the fidelity and persuasiveness of explanations in evaluations [16].
- **Integration with Decision-Making:** Ultimately, an explanation is only as useful as the actions it enables. Future BMS may act autonomously (e.g. triggering pack reconfiguration or emergency cooling). How should XAI inform automated actions? One idea is to set up human-in-the-loop systems where the BMS provides an explanation and waits for a human override or confirmation for certain critical decisions. This requires careful interface design and may involve training operators to interpret XAI outputs correctly. Additionally, explanations could be logged for post-event analysis even if no human is in the loop at runtime, creating a trace of "why" decisions were made to satisfy regulatory or forensic requirements (like investigating an EV fire).
- **Generalizability and Robustness:** Battery systems vary widely (chemistry, format, use profile). An XAI technique or model explanation that works for one scenario might not directly transfer to another.

For example, the importance of a feature like "voltage drop under 10C load" might be valid for one cell type but irrelevant for another that fails in a different way. Ensuring that XAI methods remain robust across different datasets and even detecting when a model is extrapolating beyond its trained conditions (which might manifest as unusual or low-confidence explanations) is an open problem. Some research is looking at uncertainty estimates alongside explanations to convey when a model is less sure about its reasoning.

- **Future Directions:** We foresee several avenues for advancing XAI-based fault diagnosis in battery packs. One is the development of standardized diagnostic explanation frameworks in BMS software – analogous to OBD-II codes in gasoline cars, there could be standardized "explanation codes" for battery faults (e.g. a code for "voltage discrepancy" fault cause) that an XAI system maps to. Another is leveraging transfer learning and multi-modal data: combining mechanical, electrical, and thermal models such that explanations draw from multiple domains (e.g. correlating an internal short's electrical signature with the thermal signature for a more complete explanation). As machine learning models become part of digital twin frameworks for batteries, their explainability will be key to bridging the virtual model with real-world understanding [12]. Additionally, regulatory bodies may start requiring explainability for AI in safety applications; this will push XAI from a research topic to a required feature in industry. Finally, interdisciplinary efforts between battery scientists and AI experts will be crucial – to ensure explanations are grounded in electrochemical reality and to identify new fault indicators that ML discovers. The ultimate goal is a transparent BMS where every alert or action is accompanied by a clear rationale, thus enabling safer and more effective management of multi-cell battery packs.

VI. CONCLUSION

XAI-based fault diagnosis in multi-cell battery packs is an emerging and promising area that marries the predictive power of advanced machine learning with the transparency needed for safety-critical systems. This review has surveyed the landscape of techniques and studies at this intersection. Data-driven models can detect subtle and complex fault patterns that traditional methods might miss, but explaining their decisions is vital for trust and adoption. Techniques such as SHAP and LIME provide post-hoc insights into black-box models, identifying which sensor measurements or derived features most strongly influence fault predictions [13, 11]. Interpretable model designs, like decision trees and attention-equipped networks, offer an alternative path by incorporating explainability directly into the model structure [7, 15]. Early applications of these approaches to battery packs have yielded intuitive and actionable explanations – for instance, pinpointing a single aberrant cell out of hundreds as the root of a fault [7], or linking a sudden voltage dip to an internal short diagnosis [6]. These successes demonstrate that XAI can turn battery fault diagnosis from a mysterious verdict ("pack failure") into a transparent process with supporting evidence.

At the same time, our review highlights that XAI in this domain is still in its infancy. Only a handful of works (spanning roughly the last five years) explicitly incorporate explainability into battery fault diagnostics [9]. There is ample room for deeper exploration – from developing fast, real-time explainable diagnostics to creating large-scale benchmarks for evaluating explanation methods in battery management. Challenges such as ensuring explanation fidelity, handling data scarcity, and integrating XAI with human decision-making processes will need to be addressed. Encouragingly, the trajectory is set by analogous fields (like healthcare and finance) where XAI has moved from theoretical to practical, and battery research is following suit.

In conclusion, explainable AI has the potential to significantly enhance the safety and effectiveness of battery fault diagnosis. It provides the means not only to detect faults but also to understand them. By opening the black box of machine learning models, XAI allows engineers to validate diagnostic decisions against domain expertise and thus bridges the gap between algorithm and operator. The synergy of model-based and data-driven methods, under the unifying lens of explainability, offers a powerful toolkit for managing the complexity of modern battery packs. As this field progresses, we anticipate that future battery management systems will routinely employ XAI – delivering transparent diagnostics that can accelerate troubleshooting, inform preventive maintenance, and ultimately contribute to the development of more resilient battery systems. An explainable future is a safer future for energy storage technology.

REFERENCES

[1] Rao KD, Pujitha NNL, Ranga MR, Manaswi C, Dawn S, Ustun TS, et al. Fault mitigation and diagnosis for lithium-ion batteries: a review. Front Energy Res. 2025;13:1529608
[2] 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
[3] Xu Y, Ge X, Guo R, Shen W. Recent advances in model-based fault diagnosis for lithium-ion batteries: a comprehensive review. Renew Sustain Energy Rev. (Submitted, 2024)
[4] 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
[5] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51(5):93
[6] 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
[7] 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
[8] 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.
[9] 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
[10] Yu Q, Li J, Chen Z, Pecht M. Multi-fault diagnosis of lithium-ion battery systems based on correlation coefficient and similarity approaches. Front Energy Res. 2022;10:891637
[11] 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.
[12] Njoku JN, Nwakanma CI, Kim DS, An DH. Explainable data-driven digital twins for predicting battery states in electric vehicles. IEEE Access. 2024;12:83480-83501
[13] 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.
[14] Brusa E, Cibrario L, Delprete S, Di Maggio LG. Explainable AI for machine fault diagnosis: understanding features' contribution in ML models for industrial condition monitoring. Appl Sci. 2023;13(4):2038
[15] Chen K, Zheng F, Jiang J, Zhang W, Jiang Y, Chen K. Practical failure recognition model of lithium-ion batteries based on partial charging process. Energy. 2017;138:1199-1208