batteries

Article

Early and Uncertainty-Aware Detection of Impending Voltage Outliers in Battery Packs via a Probabilistic Hierarchical Adaptive Framework

Teng Liu *, Wei Li, Zhiqiang Li and Shangbo Wu

Chongqing Seres Phoenix Intelligent Innovation Technology Co., Ltd., Chongqing 400041, China
* Correspondence: liut_bistu@163.com

Abstract

The global adoption of electric vehicles (EVs) highlights the critical role of lithium-ion battery packs in ensuring safety and performance, while voltage outliers as precursors to thermal runaway pose significant risks. Existing fault detection methods suffer from limited adaptability, poor uncertainty quantification, and inadequate handling of long-term temporal dynamics. To address these gaps, this study proposes a Probabilistic Hierarchical Adaptive Framework (PHAF) for early, uncertainty-aware detection of impending voltage outliers. PHAF integrates three core innovations: (1) the Weighted Outlier Depth (WOD) metric, which fuses Boltzmann-weighted voltage deviations and gradient-based thermal penalties to sensitively capture electro-thermal anomalies, especially under thermal stress (>45 °C); (2) the Learnable Spectral Convolution Network (LSCN), a novel architecture that combines adaptive spectral modulation and dual-path convolutions to model long-range frequency patterns and local temporal dependencies in voltage sequences; and (3) a hierarchical multi-model system that dynamically selects specialized models (LSCN, GRU, and LSTM) across four prediction horizons (160–40 min), leveraging quantile regression for uncertainty quantification and an early-termination mechanism to optimize computational efficiency. Evaluated on real-world data from 60 AITO EVs, PHAF achieves 95.4% classification accuracy for Level 1 (early-stage) faults at the 160 min horizon, >90% accuracy for critical Level 3 faults within 80 min, and a maximum AUC of 0.943 for long-term anomaly detection. This framework enables a transition from passive remediation to active prevention of battery thermal runaway, providing reliable, confidence-aware monitoring for safety-critical EV applications.

Keywords: electric vehicles; weighted outlier depth; uncertainty-aware fault prediction; learnable spectral convolution network

Academic Editor: Odne S. Burheim

Received: 25 May 2026
Revised: 30 June 2026
Accepted: 3 July 2026
Published: 6 July 2026

1. Introduction

1.1. Background

The global push toward electric vehicles (EVs) is widely seen as a necessary step to cut down on fossil fuel use and move closer to a low-carbon future [1,2]. As EV adoption grows, lithium-ion batteries have become the heart of this transformation not just because they store the energy needed to drive, but because their performance, safety, and cost shape the entire user experience [3,4]. Over the past decade, the number of EVs sold each year has jumped from just over 100,000 in 2012 to more than 10 million in 2022, showing how quickly this technology is entering mainstream use [5]. While increasing battery energy density helps reduce range anxiety and makes EVs more practical, it also introduces safety trade-offs [6]. For example, batteries with a high nickel content offer better capacity but are less stable under heat, making them more likely to overheat or fail if not carefully managed [7]. Around the world, major automakers have recalled hundreds of thousands of EVs due to battery safety concerns, which not only costs billions of dollars but also shakes consumer confidence [8]. Although full-scale battery failures remain rare thanks to better engineering, when they do happen the results can be dramatic: rapid temperature spikes, smoke, fire, or even explosions [9]. Often, these events are not sudden. They begin with small, hidden problems that grow quietly over months or years before turning into full thermal runaway [10]. These slow-developing faults rarely announce themselves with loud alarms; instead, they whisper through tiny changes in voltage, temperature, or internal resistance, signals easily drowned out by normal operational noise or dismissed as harmless fluctuations [11]. That is what makes them so dangerous: by the time the problem becomes obvious, it may already be too late to stop it [12].

In real-world EVs, one of the earliest and most telling signs of a developing battery problem is a voltage outlier, a small but unusual shift in the voltage of a single cell compared to others in the pack [13,14]. These shifts may look minor at first and be easily dismissed as normal variation or sensor noise, but they often signal deeper issues like internal shorts, lithium plating, or separator wear [15]. Left unchecked, these early electrical anomalies can trigger chemical reactions that generate heat, which then spreads to neighboring cells and can eventually lead to catastrophic failure [16]. Conditions like overcharging, extreme temperatures, or physical damage can accelerate this process, but even under normal use, tiny manufacturing flaws or material inconsistencies can slowly grow into serious faults [17]. Spotting these signals early is extremely difficult [18]. In a typical EV battery pack, close to a hundred cells work together, each with a slightly different aging rate and behavior, all while being exposed to constantly changing driving and weather conditions [19]. A cell that behaves oddly on a cold morning might look perfectly normal on a warm afternoon [20]. A voltage dip during aggressive acceleration could be mistaken for a sensor glitch rather than the start of a fault [21]. And because battery packs are designed to balance themselves; early problems in one cell can be masked by the behavior of healthy neighbors, making detection even harder [22]. Although researchers have mapped out many of the chemical and physical processes behind battery failure, including oxygen release from cathodes, reactions between lithium and the electrolyte, and breakdown of protective surface layers, turning this knowledge into practical, real-time warning systems for everyday vehicles is still a major hurdle [23–25]. The ability to catch these early signs before they become dangerous is critical, not just for safety, but for maintaining trust in EV technology as it scales globally [26]. Without reliable early warnings, we are forced to wait until something goes wrong—and in the world of high-energy batteries, waiting can be costly, or even deadly [27].

As lithium-ion batteries become more common in homes, cities, and transport systems, their safety and reliability matter more than ever [28]. A single failure in one vehicle can have ripple effects, triggering recalls, damaging brand reputation, or even slowing down the broader adoption of clean energy technologies [29]. While lab tests can simulate failure under controlled conditions, real-world operation is far messier: batteries face unpredictable loads, extreme temperatures, partial charging cycles, and long idle periods, all of which affect how faults develop and how easy they are to detect [30,31]. The path from a tiny defect to a full thermal event is rarely linear or fast; it often involves slow, hidden changes that only become visible when it is almost too late [32]. Without early detection, we remain in a reactive mode, fixing problems only after they happen [33]. But with better

https://doi.org/10.3390/batteries12070245

tools to spot the earliest warning signs, like small voltage drifts under thermal stress, we can shift toward prevention, giving drivers, fleet operators, and battery management systems the time they need to act before things go wrong [34]. This is especially urgent as EVs move from early adopters to mass-market users, where safety expectations are higher and the consequences of failure are greater. Building a future where EVs are truly safe at scale means going beyond just making better batteries; it means building smarter ways to listen to what those batteries are trying to tell us, long before they fail [9]. Furthermore, the transition of battery technologies from electric vehicles to broader energy storage applications highlights the critical need for targeted lithium-ion battery diagnostics to ensure safety and reliability across diverse operational scenarios [35].

1.2. Current Methods for Early Battery Fault Warning

Predicting early battery faults in real-world electric vehicles using traditional rule-based [16] or physics-driven [36] models remains a difficult task, often held back by many practical limitations [37]. Especially in large-scale EV fleets, challenges come from unpredictable cell variations, inconsistent manufacturing quality, and the high cost of collecting and labeling high-resolution field data [38]. Even when lab experiments simulate faults such as forced overcharge or mechanical crush, these controlled scenarios rarely reflect how failures actually emerge during daily driving, where stress is slow, irregular, and hidden [24]. The gap between clean lab data and messy, real-world battery behavior makes it hard to translate academic models into reliable tools for automakers or battery management systems [39]. In recent years, however, the rapid growth of data science, cloud computing, and sensor networks has created new ways to tackle these problems not by building perfect physical models but by learning directly from what batteries actually do in the field [40]. These new approaches rely on intelligent algorithms that find patterns in voltage, temperature, and current data without needing to understand every chemical reaction inside the cell [41]. They belong to a family of methods called machine learning, which can handle complex, noisy, and nonlinear relationships in time-series data, exactly the kind of signals that appear before a fault becomes dangerous [42]. In addition to algorithmic advancements, the comprehensive analysis of open lithium-ion battery testing datasets plays a pivotal role in advancing degradation modeling and life cycle management, providing the necessary foundation for training robust data-driven models [43].

In battery health monitoring, many current methods depend on supervised machine learning, meaning that they need large amounts of pre-labeled “good” and “bad” data to train on [24]. For instance, one study used random forest classifiers to detect mechanical damage by analyzing eight key features extracted from charge–discharge cycles, such as end-of-charge voltage, open-circuit voltage slope, and internal resistance trends [44]. Another approach applied a multiclass relevance vector machine to identify internal short circuits by tracking four main indicators: how much a cell’s SOC drifts from its neighbors, how its temperature behaves abnormally, how its terminal voltage deviates under load, and how its resistance reacts to current changes [45]. These methods work well in controlled tests, but they require engineers to manually choose which features matter most, a time-consuming process that does not always capture the full picture [46]. Other studies have combined machine learning with physics-based simulations, such as using finite element models to generate virtual battery stress data, then training classifiers to predict safety boundaries under different mechanical loads [34,36]. These hybrid methods help bridge the gap between simulation and reality, but they still rely heavily on assumptions that may not hold true in actual EV operation [47].

To improve accuracy and reduce reliance on hand-crafted features, recent efforts have focused on enhancing supervised learning itself, not by replacing it but by making it smarter and more adaptive [48,49]. Instead of using fixed feature sets, newer supervised models take raw or minimally processed time-series data like voltage, temperature, and current sequences and learn directly which patterns are most predictive of early faults [50]. This shift allows the model to discover subtle, nonlinear relationships that human engineers might overlook [33]. For example, instead of calculating "voltage deviation over 10 min" as a fixed rule, the model might learn that a specific combination of slow voltage drop followed by a sharp temperature rise under moderate load is a stronger predictor of impending failure. Supervised learning also benefits from richer, higher-resolution datasets collected from real-world EV fleets, data that captures the true complexity of battery behavior across seasons, driving styles, and aging stages [51]. When trained on diverse, high-quality labeled examples, supervised models can generalize better and detect faults earlier than ever before, especially when combined with techniques like data augmentation, ensemble learning, and multi-horizon prediction [52].

The most powerful supervised approaches today use deep neural networks, especially Long Short-Term Memory (LSTM) networks and their variants, to model how battery signals evolve over time [53]. LSTMs are particularly good at this because they can remember past states and use them to interpret current behavior, making them ideal for spotting slow-developing faults that unfold over hours or days [54]. For example, an LSTM can be trained to predict the next 30 min of a cell's voltage curve, and if the real voltage starts to drift away from the prediction, that is a warning sign. Some models go further, predicting multiple parameters at once, voltage, temperature, and SOC, to get a more complete picture of cell health. Hybrid models that combine LSTMs with equivalent circuit models (ECMs) have also shown promise, using physics-based constraints to guide the learning process and improve interpretability [55]. While training deep networks can be challenging, with issues like vanishing gradients and long training times, recent advances in optimization, regularization, and model architecture have made them more stable and easier to deploy [56]. And unlike unsupervised or semi-supervised methods, supervised deep learning gives clear, quantifiable performance metrics, accuracy, precision, and recall, making it easier to validate, compare, and trust in safety-critical applications like EV battery monitoring [16].

In battery applications, Transformer-based models have already shown strong results in estimating SOC, predicting SOH, and forecasting remaining life under real driving cycles [5]. Some recent studies have even used Transformers to build thermal warning systems, combining voltage data with simulated or real thermal images to detect cells that are heating up abnormally, often the last step before thermal runaway [53]. These models do not just predict one value; they learn the full relationship between multiple signals: how voltage dips relate to temperature spikes, how charge speed affects internal resistance, and how small imbalances grow over time. And because they work directly on raw sequences, they adapt better to new battery types, new driving styles, and new environmental conditions without needing engineers to redesign them from scratch. While still evolving, these deep learning approaches represent the most promising path toward truly early, reliable, and automatic fault warnings, the kind that can catch a problem not when it is obvious but when it is just beginning. Furthermore, transfer learning has also made significant advances in battery state estimation and diagnosis, such as enabling battery health prognosis using impedance spectrum data [57]. Although not the primary focus of this study, integrating such techniques represents a promising future direction to overcome data scarcity in real-world applications.

While physics-based models and equivalent circuit models provide interpretable insights into battery degradation, they heavily rely on precise parameter identification and struggle to adapt to the complex nonlinear aging processes and individual cell variations in large-scale fleets [58]. Traditional machine learning approaches, such as random forests and support vector machines, require extensive manual feature engineering, which often

https://doi.org/10.3390/batteries12070245

fails to capture the implicit high-dimensional temporal dependencies hidden in raw voltage sequences [59]. Although standard deep learning architectures like Long Short-Term Memory networks have demonstrated superior performance in sequence modeling, they predominantly focus on local temporal steps and suffer from vanishing gradients over very long horizons [60]. Moreover, most of these existing models output deterministic point predictions without providing any measure of prediction confidence, severely limiting their reliability in safety-critical applications where false alarms must be strictly minimized.

1.3. Challenges in Predicting Real-Life Battery Failure Problems

Despite encouraging progress in data-driven battery fault detection, reliably predicting early voltage outliers under real-world conditions remains a major unsolved problem, especially when working within the practical limits of onboard battery management systems (BMSs). While existing methods show promise in controlled or idealized settings, they often fail to adapt to the messy, uncertain, and constantly changing nature of real EV operation. Many models are too rigid, using fixed time windows, single architectures, or ignoring the thermal context of voltage changes, which makes them miss subtle early warnings or raise too many false alarms. Others lack the ability to say how confident they are in their predictions, leaving operators blind to whether a warning is urgent or uncertain. To build a system that works reliably across different vehicles, climates, and driving styles, we must overcome four critical challenges:

(a) Current fault detection methods often analyze voltage or temperature in isolation, even though real battery failures are driven by their combined behavior. A small voltage dip might be harmless on a cool day, but dangerous under high heat. Yet most models do not account for this coupling. They use static thresholds or simple statistical rules that treat all deviations the same, regardless of thermal context. This leads to missed detections when thermal stress amplifies risk, or unnecessary alarms when deviations are benign. What is needed is a scoring system that weights voltage anomalies based on their thermal environment, giving more attention to cells that drift under high temperature, where failure risk is greatest.

(b) Voltage outliers do not appear suddenly, they evolve slowly, often over hours or days, hidden within complex frequency patterns that standard models like LSTMs or GRUs struggle to capture. These models focus on local, step-by-step dependencies but miss the bigger picture, like how a cell's voltage rhythm is changing over the long term or how its frequency signature is shifting as it degrades. Without seeing these global patterns, models cannot spot the earliest signs of trouble. What is needed is a new kind of architecture, one that looks at voltage sequences not just as strings of numbers, but as signals with frequency content, like audio or vibration data. By learning to modulate and analyze these spectral patterns, amplifying the right frequencies, suppressing noise, and combining them with local trends, we can detect the slow, hidden drifts that precede failure, long before they become obvious in the time domain.

(c) Most deep learning models give only a single "best guess" prediction, with no measure of how certain they are. But in safety-critical systems like EV batteries, knowing the confidence behind a warning is just as important as the warning itself. Without uncertainty bounds, operators cannot tell, leading to either ignored alarms or unnecessary shutdowns. Even when models try to estimate confidence, they often use crude approximations or add-on methods that do not integrate with the core prediction. What is needed is a built-in, probabilistic approach, like quantile regression, that learns to predict not just the most likely outcome, but also the full range of possible outcomes, from optimistic to worst-case. This lets the system trigger alarms only

https://doi.org/10.3390/batteries12070245

when the risk is both high and certain, reducing false alarms without sacrificing sensitivity.

(d) Almost all existing algorithms use a single fixed model with one input length and one prediction window, whether it is 10 min or 2 h. But in reality the best time to warn about a fault depends on how fast it is developing. A slow voltage drift might need a 160 min forecast to catch it early; a sudden drop might only need 40 min. Fixed models cannot adapt; they either warn too early or too late (and miss the window for action). Worse, they force the same architecture to handle all time scales, even if it is only good at one. A truly adaptive system should dynamically choose the right model and the right time horizon based on what the data is showing, switching from long-term spectral analysis to short-term sequence modeling as the fault progresses, and stopping early if no risk is found, to save computation.

1.4. Contributions and Organization

To clearly position the novelty of the proposed framework relative to existing methods, it is essential to highlight the specific advancements over prior works. Unlike rule-based thermal runaway warning systems that treat voltage and temperature independently [61], the proposed Weighted Outlier Depth metric explicitly fuses electrochemical and thermal signals to capture coupled anomalies under thermal stress. Compared to recent Transformer-based approaches that model global attention but ignore spectral characteristics [62], the proposed Learnable Spectral Convolution Network uniquely integrates adaptive spectral modulation with dual path convolutions, enabling the extraction of both global frequency signatures and local temporal dependencies. Furthermore, while current uncertainty quantification methods in battery prognostics often rely on Monte Carlo dropout or ensemble techniques that are computationally prohibitive for online use [63], the proposed framework embeds lightweight quantile regression directly into a hierarchical multi-horizon architecture, providing calibrated confidence bounds with minimal overhead. This comprehensive integration ensures that the framework not only detects early faults with higher sensitivity but also supports reliable decision making in real-time safety-critical scenarios.

The fundamental hypothesis of this research is that significant voltage deviations observed among battery cells during vehicle standstill periods, where current is negligible, are primarily driven by genuine internal degradation mechanisms such as micro short circuits rather than sensor noise or load transient effects. Building upon this physical and operational premise, it is further hypothesized that integrating electrothermal features with probabilistic deep learning can reliably capture these early fault signatures and quantify prediction uncertainty.

To address the critical need for early and uncertainty-aware detection of voltage outliers in real-world EV battery systems, a novel Probabilistic Hierarchical Adaptive Framework (PHAF) has been developed. This framework enables proactive fault warnings under dynamic and thermally stressed operating conditions without relying on fixed thresholds or monolithic model architectures. Validated using field data from 60 AITO electric vehicles, the system achieves 95.4 percent classification accuracy for early-stage voltage anomalies at the 160 min horizon and maintains over 90 percent accuracy for critical faults within 80 min. By integrating physics-informed anomaly scoring, spectral temporal deep learning, multi-horizon model specialization, and quantile-based uncertainty quantification, the proposed framework represents a significant advancement toward deployable and confidence-aware battery safety monitoring. The core innovations are structured around three key technical contributions:

(1) A novel Weighted Outlier Depth (WOD) metric is introduced to jointly evaluate voltage and temperature signals through a probabilistic scoring mechanism. Drawing on

https://doi.org/10.3390/batteries12070245

the Boltzmann distribution, the method assigns higher weights to cells exhibiting larger voltage deviations, especially under elevated temperatures above 45 °C, where electrochemical instability is more likely to escalate. A parallel thermal penalty component further amplifies scores for probes showing both statistical deviation and absolute overheating. This dual emphasis ensures that anomalies are flagged in the context of their thermal risk, enabling earlier and more meaningful warnings than voltage-only or rule-based systems.

(2) A specialized deep learning architecture named the Learnable Spectral Convolution Network (LSCN) is designed to capture long-term evolution patterns in battery voltage sequences. Unlike conventional recurrent networks that focus on local step-by-step dependencies, the LSCN processes sequences in both time and frequency domains. It employs adaptive spectral modulation to learn which frequency bands carry the most predictive power for fault development, while dual path convolutions extract complementary short-term temporal features. This hybrid approach allows the model to detect subtle and slow-drifting voltage abnormalities by recognizing changes in their underlying spectral signatures over hours of operation.

(3) A probabilistic hierarchical multi-model prognosis system is implemented to handle varying prediction horizons with specialized architectures and explicit uncertainty quantification. Four time windows of 160, 120, 80, and 40 min are assigned to models best suited for each scale, namely, the LSCN for spectral long-range forecasting, GRUs for intermediate dynamics, and LSTM for fine-grained sequential prediction. Quantile regression is embedded into every level to provide calibrated uncertainty estimates, outputting prediction intervals that reflect the range of plausible future voltages. An intelligent gating mechanism evaluates fault likelihood at each level and terminates further computation early if no risk is detected within the confidence interval. This structure ensures high sensitivity for early warnings while minimizing unnecessary processing for resource-constrained onboard deployment.

The remainder of this paper is organized as follows: Section 2 introduces the dataset and data preprocessing procedures. Section 3 details the proposed Probabilistic Hierarchical Adaptive Framework. Section 4 presents the experimental results and performance evaluations. Section 5 concludes the paper.

2. Data Description

This study utilizes real-world vehicle data collected from an AITO electric vehicle fleet during operational cycles. Regarding the data sources, the dataset is exclusively derived from the cloud monitoring platform of AITO pure electric vehicles, where raw signals are continuously uploaded by the Telematics Control Unit during daily driving. This ensures that the data reflects genuine and complex operating conditions rather than idealized laboratory environments. Concerning the key roles of this data, it serves three fundamental purposes in this research. First, it provides the massive time-series samples required for training and validating the deep learning models within the proposed framework. Second, it offers authentic voltage and temperature profiles to verify the effectiveness of the Weighted Outlier Depth metric in capturing electro-thermal anomalies. Third, the diverse fault patterns contained in the dataset act as the benchmark for evaluating the predictive accuracy and generalization ability of the multi-horizon prognosis system, thereby confirming the practical engineering value of the proposed method.

The data, transmitted in real time via the Telematics Control Unit to a cloud-based monitoring platform, establishes a comprehensive foundation for analyzing battery system state characteristics under actual driving conditions. The dataset comprises complete operational logs from 60 vehicles, with 50 exhibiting well-defined voltage outlier events and the remaining 10 operating normally throughout the collection period. Raw data was

https://doi.org/10.3390/batteries12070245

sampled at 1 Hz, capturing high-resolution dynamic information about the battery system. Each vehicle monitoring stream includes critical signals: (a) voltage readings from all 96 individual cells within the battery pack; (b) temperature measurements from 16 strategically placed sensors across key pack locations; (c) essential battery management system parameters such as state of charge, total current, and total voltage.

The dataset comprises complete operational logs from 60 vehicles, collected over a continuous period of 18 months. The vehicle usage diversity is substantial, as the fleet consists of private passenger cars subjected to highly diverse driving cycles, varying ambient temperatures ranging from minus 10 to 40 degrees Celsius, and different user charging habits. This ensures the data reflects genuine and complex operating conditions. Regarding the fault distribution, a total of 142 independent voltage outlier events were recorded across the 50 faulty vehicles, while the remaining 10 vehicles operated normally throughout the collection period. These independent fault events include various degradation stages, providing a comprehensive basis for evaluating the early-warning capability of the proposed framework.

To ensure a robust evaluation of the framework across varying temporal scales, the sample distribution was quantified for each prediction horizon (40, 80, 120, and 160 min). As the prediction window increases, the total number of available sequences naturally decreases due to the stricter requirements for continuous historical data and future ground truth labels. The specific breakdown of sample counts across the four fault severity levels (Normal, Level 1, Level 2, and Level 3) for each specialized model is summarized in Table 1. Notably, the dataset exhibits a higher prevalence of Level 1 anomalies compared to Level 3 faults, reflecting the real-world progression of battery degradation where early warnings are more frequent than terminal events.

Table 1. Sample size distribution across different prediction horizons and fault severity levels.

Prediction Horizon	Normal	Level 1	Level 2	Level 3	Total Samples
40 min	1321	5407	2654	2192	11,574
80 min	443	3461	1167	401	5472
120 min	178	1220	913	350	2661
160 min	59	670	191	205	1125

To prevent information leakage and ensure a fair evaluation under real-world deployment scenarios, the dataset was partitioned at the vehicle level based on unique vehicle identification numbers (VINs). Specifically, 70% of the vehicles were randomly selected as the training set, 15% were assigned to the validation set for hyperparameter tuning and early stopping, and the remaining 15% were reserved as an independent test set for final performance evaluation. All temporal sequences were constructed exclusively within each subset, ensuring that no data segments from the same vehicle appeared in more than one split. This VIN-based partitioning strategy preserves temporal integrity and reflects realistic generalization conditions in fleet-level battery fault diagnosis.

A rigorous preprocessing pipeline was implemented to ensure data integrity and consistency. This involved linear interpolation to address missing values and the elimination of spurious outliers arising from sensor communication anomalies. For the fault early-warning task, computational efficiency and preservation of fault-relevant features necessitated resampling the original 1 Hz data to 5 min intervals. This resampling strategy effectively reduces model inference overhead while retaining long-term temporal patterns indicative of emerging faults.

Precise annotation of voltage outlier events is critical for developing and validating supervised learning models. The annotation protocol adheres to operational rules from the cloud monitoring system, specifically targeting cells with abnormally low voltage during vehicle standstill, defined as periods where current approaches negligible levels. This standstill criterion mitigates interference from ohmic voltage drops induced by load variations, thereby isolating genuine anomalies potentially linked to internal micro shorts or similar degradation mechanisms. Consequently, the annotations reflect physically meaningful fault signatures rather than transient operational artifacts.

To identify voltage outliers within a battery pack consisting of $N = 96$ cells, a systematic labeling procedure is implemented based on inter-cell voltage variations. At each timestamp t, the cell voltages are denoted as $V_i(t)$, where $i = 1, 2, \ldots, N$. The procedure begins by evaluating the overall voltage dispersion across the pack through the voltage range:

$$\Delta V_{\mathrm{diff}}(t) = \max_i\left(V_i(t)\right) - \min_i\left(V_i(t)\right) \tag{1}$$

If $\Delta V_{\mathrm{diff}}(t)$ < volt_set, the pack is considered normal at this timestamp and subsequent steps are skipped; otherwise, proceed.

The average voltage of all cells is computed as

$$V_{\mathrm{avg}}(t) = \frac{1}{N}\sum_{i=1}^{N} V_i\,(t) \tag{2}$$

which serves as a reference for evaluating individual cell deviations. Subsequently, the cell with the minimum voltage, $V_{\mathrm{min}}(t) = min_i V_i(t)$, is identified, and its deviation from the average is calculated:

$$\Delta V_{\mathrm{dev}}(t) = V_{\mathrm{min}}(t) - V_{\mathrm{avg}}(t) \tag{3}$$

Given that abnormally low voltages are of primary concern in this context, the value of this deviation, $\Delta V_{\mathrm{dev}}(t)$, is compared against a set of predefined thresholds to determine the severity level of the anomaly. The corresponding alarm level is then assigned according to Table 2.

Table 2. Threshold configuration table.

Name	Description	Unit	Value	Alarm Level
volt_set	Voltage difference threshold	mV	20	/
rank_threshold	Grade threshold	mV	50	Level 1
		mV	100	Level 2
		mV	150	Level 3

The selection of these specific threshold values is grounded in practical engineering experience from battery management systems and relevant industry standards. The voltage threshold of 20 mV serves as a noise filter to prevent false alarms caused by minor sensor fluctuations, while the rank threshold values of 50, 100, and 150 mV correspond to progressive stages of cell imbalance and internal degradation risk, ensuring that the alarm levels align with actual safety intervention requirements.

Voltage outliers reflect the degree of consistency degradation inside the battery and the state of micro short circuits, which is one of the early manifestations of thermal runaway. Figure 1 shows a voltage characteristic curve of a lithium battery cell sampled per minute. Before 7500 min, most of the judgment values are in the first-level and second-level fault states. With the charging and discharging of the lithium battery, the maximum cell voltage, minimum cell voltage and average voltage show obvious fluctuations with ups and downs. After 7500 min, these three cell voltage characteristic curves basically overlap, indicating that the voltage outlier fault has disappeared temporarily. Taking the judgment value curve as the target, this paper constructs simplified thermal–electrical

https://doi.org/10.3390/batteries12070245

model features of lithium batteries and builds an efficient early-warning model for voltage outlier faults. By predicting the early manifestations of thermal runaway, it achieves a key transformation from passive remediation to active prevention.

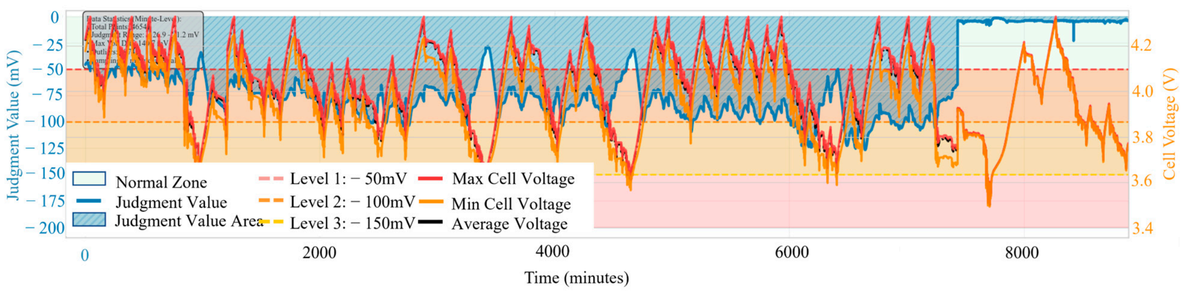

Figure 1. Voltage outlier fault curve.

3. Proposed Probabilistic Hierarchical Adaptive Framework

3.1. Background and Model Architecture

In recent years, the rapid advancement of EV technology has placed increasing demands on the safety, reliability, and efficiency of lithium-ion battery systems [64]. As battery packs grow in size and complexity, traditional fault detection methods based on fixed thresholds or simple statistical models have proven insufficient for capturing the intricate electro-thermal coupling behaviors that precede critical failures such as thermal runaway [11]. While conventional approaches often rely on voltage deviation or temperature monitoring in isolation, they frequently overlook the nuanced interplay between electrical and thermal anomalies, leading to delayed or inaccurate early warnings [65]. Moreover, existing data-driven techniques, particularly those rooted in classical machine learning, struggle to model long-term temporal dependencies and quantify prediction uncertainty—both of which are essential for reliable prognostics under dynamic operating conditions [66]. To address these limitations, deep learning-based frameworks have emerged as promising alternatives, offering enhanced representation learning capabilities and adaptability to complex nonlinear dynamics [67]. However, most current deep learning models focus primarily on point predictions without explicitly accounting for uncertainty, thereby limiting their applicability in safety-critical applications where confidence estimation is paramount.

To bridge this gap, we propose a novel PHAF that integrates physics-informed anomaly metrics with uncertainty-aware deep learning architectures for real-time fault early warning in large-scale battery systems. This framework leverages a multi-dimensional outlier detection strategy by combining Boltzmann-weighted voltage analysis and gradient-based thermal penalty modeling to compute cell-level anomaly scores, namely, WOD scores, which reflect both electrochemical inconsistency and thermal risks. These physical insights are then fed into a hierarchical prognosis architecture composed of four specialized deep learning models—each tailored to capture distinct temporal scales: Model A (LSCN) for long-term spectral–temporal patterns, Models B and C (GRU-based) for medium- and short-term dynamics, and Model D (LSTM) for immediate-term sequential behavior. The integration of quantile regression enables probabilistic forecasting across all horizons, providing not only accurate voltage trajectory predictions but also explicit uncertainty bounds. By sequentially evaluating fault likelihood at multiple time scales and employing an adaptive early-termination mechanism based on confidence intervals, the PHAF achieves a balance between computational efficiency and diagnostic accuracy, making it highly suitable for deployment in resource-constrained onboard battery management systems. The overall structure, illustrated in Figure 2, demonstrates how the proposed method unifies physical modeling, deep learning, and probabilistic reasoning into a cohesive, scalable framework for proactive battery health monitoring.

To explicitly clarify how the frame structure in Figure 2 achieves online battery fault warning, the operational workflow during the deployment phase is detailed as follows. During real-world operation, the Telematics Control Unit continuously streams raw sensor data to the cloud platform. The incoming data is dynamically partitioned into fixed-length sliding windows. Each newly formed window is immediately processed by the feature extraction module to compute the WOD metrics. Subsequently, the extracted features are sequentially fed into the hierarchical deep learning models. The system evaluates the prediction confidence at each hierarchical level. If the uncertainty bounds indicate a high probability of normal operation, the early termination mechanism is triggered to bypass subsequent shorter horizon models, thereby saving computational resources. Otherwise, the evaluation proceeds to the next level until a definitive fault warning is generated. This continuous streaming and sliding window pipeline ensures that proactive and reliable fault warnings are delivered in real time without requiring offline batch processing.

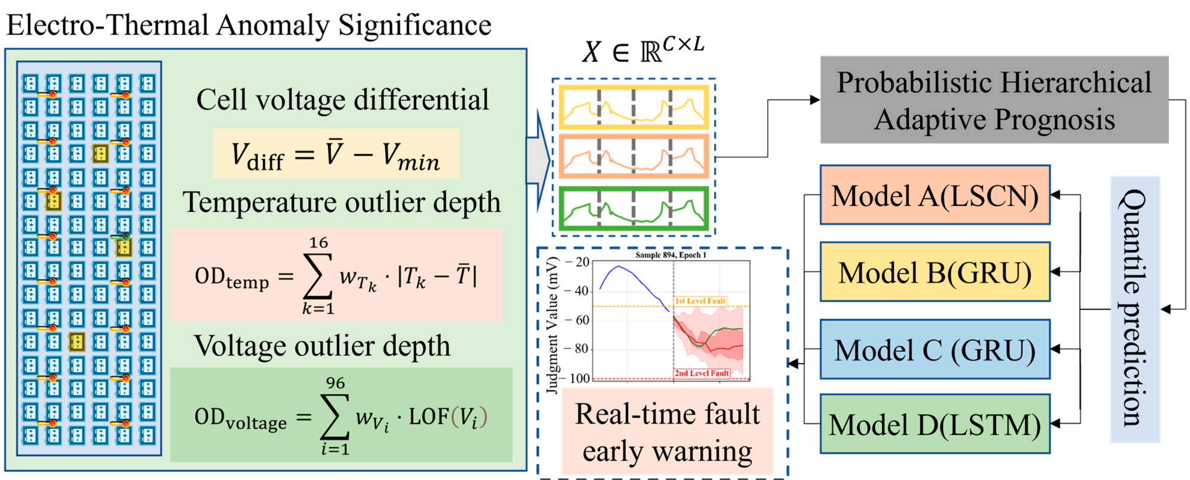

Figure 2. Proposed Probabilistic Hierarchical Adaptive Framework.

3.2. Modeling of the Boltzmann-Weighted Voltage Outlier Model and the Gradient-Based Thermal Penalty Model

This study proposes a novel WOD metric for detecting anomalies in lithium-ion battery systems through the joint analysis of voltage and temperature distributions. The theoretical foundation for voltage weighting is derived from the Boltzmann distribution [68], a core concept in statistical thermodynamics that describes the probability of a physical system occupying a specific energy state. This distribution enables the model to prioritize cells with significant voltage deviations. The proposed method integrates three key components: Boltzmann-based voltage weighting, a gradient-based thermal penalty, and Local Variance Ratio (LVR) computation. The Boltzmann-based voltage weighting leverages the distribution's inherent capacity to amplify differences between normal and abnormal cell states, while the entire integrated framework quantifies cell-level abnormalities under electro-thermal coupling constraints, which is an essential scenario for ensuring battery safety and optimizing performance management.

Step 1: Boltzmann-Weighted Voltage Distribution.

To provide a stronger theoretical foundation for the proposed metric, the physical rationale connecting voltage deviations to Boltzmann statistics is explicitly elaborated. In statistical thermodynamics, the Boltzmann distribution governs the probability of a physical system occupying a specific energy state at thermal equilibrium. Within a battery pack, the voltage deviation of an individual cell from the pack average can be physically mapped to its electrochemical potential energy state. Healthy cells naturally reside in a low-energy stable equilibrium, whereas cells experiencing internal degradation or micro short circuits manifest as high-energy excited states due to significant voltage drops. By adopting the Boltzmann factor, the proposed weighting mechanism exponentially amplifies the statistical weights of these high-energy anomalous states while effectively suppressing the background fluctuations of normal cells.

In this formulation, the parameter β_v physically represents the inverse of the effective system temperature, functioning as a sensitivity or focusing coefficient. A larger β_v sharpens the probability distribution, thereby increasing the model sensitivity to minute voltage deviations. This physical interpretation provides a rigorous justification for the adaptive adjustment of β_v under thermal stress conditions. When the ambient or internal temperature exceeds 45 degrees Celsius, the internal electrochemical reactions accelerate, and the risk of thermal runaway escalates. Consequently, increasing β_v under such conditions is physically equivalent to lowering the effective temperature in the statistical model, which forces the distribution to focus more intensely on the most critical anomalous cells, ensuring that early-warning signals are not masked by normal operational noise.

Voltage deviations are weighted using a Boltzmann distribution to emphasize abnormal cells, as this distribution inherently assigns higher probabilities to states that deviate from the system's equilibrium (represented by the average voltage here). The normalized voltage weight $w_v^{(i)}$ for cell i is computed as

$$w_v^{(i)} = \frac{\exp(-\beta_v \cdot |v_i - \bar{v}|)}{\sum_{j=1}^{96} \exp\left(-\beta_v \cdot \left|v_j - \bar{v}\right|\right)} \tag{4}$$

where $\bar{v} = \frac{1}{96}\sum v_i$ denotes the average voltage and β_v is a voltage focusing parameter dynamically adjusted based on thermal conditions. The adaptive β_v implementation

$$\beta_v = \begin{cases} 0.8\beta_v, \max(t) > 45°C \\ \beta_v \end{cases} \tag{5}$$

effectively increases sensitivity to voltage deviations during thermal stress. This adjustment is motivated by the fact that elevated temperatures (above 45 °C) can accelerate battery degradation mechanisms, making voltage anomalies more predictive of critical faults and thus necessitating enhanced detection sensitivity.

Step 2: Gradient-Based Thermal Penalty with Overheat Compensation.

Temperature weights incorporate both normalized deviation and explicit overheat penalties, as temperature anomalies in lithium-ion batteries not only indicate performance issues but also pose fire and explosion risks if unaddressed. The temperature weight $w_t^{(k)}$ for the *k*-th temperature probe is defined as

$$w_t^{(k)} = \frac{|t_k - \bar{t}|}{\max|t_m - \bar{t}|} + \lambda \cdot \mathbb{I}(t_k > 45°C) \tag{6}$$

where $\bar{t}$ is the average temperature across 16 probes, $\lambda = 0.1$ is the penalty factor (calibrated to balance statistical deviation and safety priorities), and $\mathbb{I}(\cdot)$ denotes the indicator function (which equals 1 if the overheat condition is met and 0 otherwise). This dual mechanism amplifies weights for both statistically deviant and critically overheated regions, ensuring that the model does not overlook either subtle thermal drifts or severe overheating events, both of which are critical for comprehensive battery health monitoring.

Step 3: Computation of the Local Variance Ratio.

To quantify the local isolation of anomalous cells while maintaining computational efficiency, a simplified metric named the Local Variance Ratio (LVR) is introduced. Unlike the conventional Local Outlier Factor which relies on complex local density estimation, the LVR is primarily based on neighborhood variance ratios. This approach targets the top 5% voltage outliers, as it effectively measures how isolated a data point is relative to its local variance distribution. Restricting the analysis to the top 5% of candidates reduces computational complexity while focusing on the most impactful anomalies. The simplified LVR for cell i is calculated as

$$\mathrm{LVR}(i) = \frac{\mathrm{Var}\big(\mathcal{N}_k(v_i)\big)}{\mathrm{Var}(\mathbf{v}) + \epsilon} \tag{7}$$

where $\mathcal{N}_k(v_i)$ represents the $k = 5$ nearest voltage neighbors of cell i, $\epsilon = 10^{-8}$ prevents division by zero, and computation is restricted to candidate set $\mathcal{C} = \{i : d_v^{(i)} \in \text{top } 5\%\}$.

Step 4: Computation of Voltage- and Temperature-Specific WOD Metrics.

To quantify anomaly signals from the voltage and temperature of lithium-ion battery systems separately, weighted integration is performed for each dimension. This approach preserves the distinct physical meanings of voltage and temperature data, as voltage primarily reflects the electrochemical state of individual cells, while temperature directly indicates thermal safety status:

$$WOD_{voltage} = \sum_{i=1}^{96} w_v^{(i)} \cdot \mathrm{LVR}(i) \tag{8}$$

$$WOD_{temp} = \sum_{k=1}^{16} w_t^{(k)} \cdot |t_k - \bar{t}| \tag{9}$$

In the above equations, the product of $w_v^{(i)}$ and $\mathrm{LVR}(i)$ ensures that cells exhibiting both large voltage deviations and strong isolation contribute more significantly to the $WOD_{voltage}$ score. This targeted weighting enables the metric to prioritize cells that are most likely to have electrochemical abnormalities. The thermal weight $w_t^{(k)}$, which integrates both the normalized temperature deviation and an explicit overheat penalty, is multiplied by the raw temperature deviation $|t_k - \bar{t}|$. This multiplication amplifies the contribution of thermally critical probes, enabling the WOD_{temp} to effectively capture both mild thermal drifts and severe overheating events.

The curve data of $WOD_{voltage}$ and WOD_{temp} is shown in Figure 3. Throughout the entire dataset, the value of WOD_{temp} fluctuates slowly within the range of 4–11, while the value of $WOD_{voltage}$ fluctuates more drastically, ranging between 0.005 and 0.015. These two features enable the evaluation of the inconsistency degree of the battery system in both electrical and thermal dimensions.

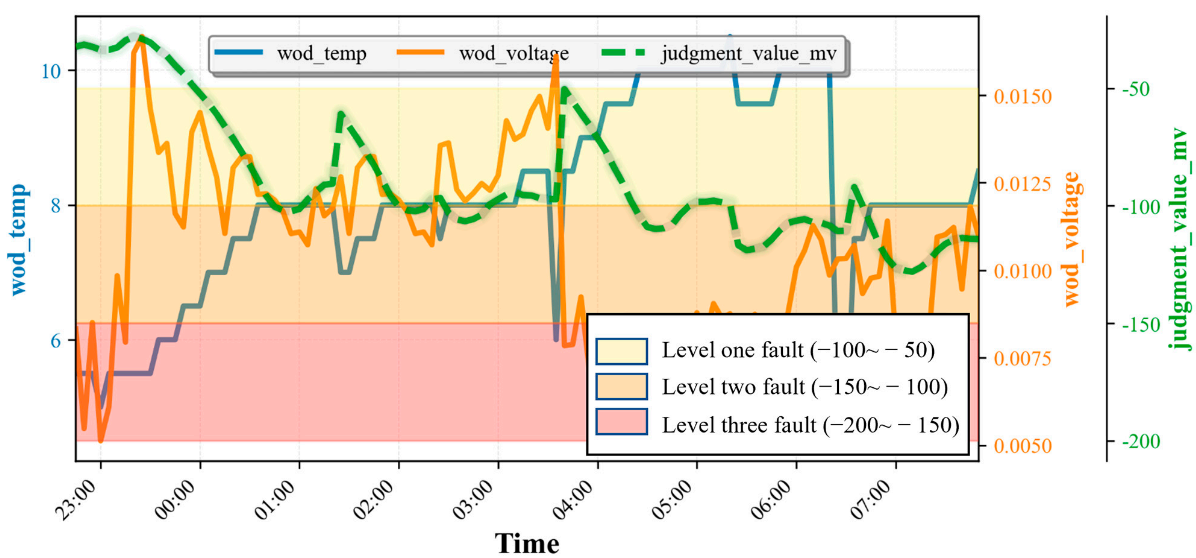

Figure 3. Voltage outlier, thermal penalty and judgement value curve.

3.3. Uncertainty-Aware Deep Learning Model Training Method

To effectively quantify prediction uncertainty in lithium-ion battery voltage outlier early warning, a quantile regression-based loss function is employed during model training. This approach enables simultaneous prediction of multiple quantiles, thereby providing a probabilistic forecasting framework that captures both central tendency and prediction intervals. The model is trained to estimate several key quantiles, specifically [0.1, 0.3, 0.5, 0.7, 0.9], which represent different probability levels in the conditional distribution of the target voltage signal.

The quantile loss function is formally defined for each quantile level $q \in [0,1]$. For a given quantile q, the loss between the predicted value $\hat{y}_q$ and the true observation y is computed as

$$L_q(y, \hat{y}_q) = \max\left((q-1)(y-\hat{y}_q), q(y-\hat{y}_q)\right) \tag{10}$$

This function is asymmetric, penalizing over-prediction and under-prediction differently depending on the chosen quantile. The overall training objective is to minimize the sum of quantile losses across all target quantiles. Let $Q = \{q_1, q_2, \dots, q_k\}$ denote the set of desired quantiles; the total loss over a batch of sequences is given by

$$\mathcal{L} = \frac{1}{k}\sum_{i=1}^{k} L_{q_i}(y, \hat{y}_{q_i}) \tag{11}$$

In this implementation, the model outputs a tensor of shape [B, T, 5], where the last dimension corresponds to the five pre-specified quantiles. The training procedure optimizes all quantile predictions jointly, enabling the model to not only provide accurate point forecasts (via the median quantile, i.e., q = 0.5) but also to quantify uncertainty through interval forecasts derived from extreme quantiles such as q = 0.1 and q = 0.9. This uncertainty-aware training method significantly enhances the reliability of early-warning signals for voltage anomalies in electric vehicle batteries by communicating prediction confidence explicitly.

3.4. A Learnable Spectral Convolution Network for Long-Time Sequence Prediction

The LSCN architecture is specifically designed for long-term voltage sequence forecasting in lithium-ion battery early-warning systems. This novel framework integrates spectral domain processing with convolutional operations to capture both global frequency patterns and local temporal dependencies. The model processes multivariate input sequences of shape [B, L, M], where B represents batch size, L denotes sequence length, and M corresponds to feature dimensionality. The core architecture comprises three fundamental components: a patch embedding module, stacked LSCN layers, and a quantile prediction head, as shown in Figure 4.

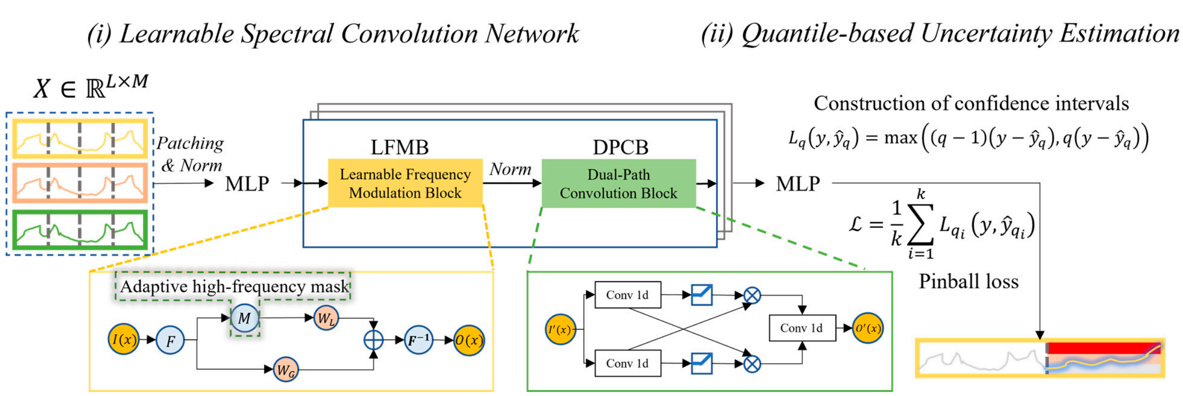

Figure 4. LSCN and quantile-based uncertainty estimation method.

3.4.1. Norm and Patching

The input sequence undergoes normalization through mean subtraction and standard deviation division along the temporal dimension. The normalized data is restructured into overlapping patches of size P with stride S = P/2, generating N patches per feature dimension. Each patch is linearly projected into a D-dimensional embedding space via the following transformation:

$$\mathbf{E} = \mathbf{W}_e\mathbf{X}_{\text{patches}} + \mathbf{b}_e \tag{12}$$

where $\mathbf{W}_e$ and $\mathbf{b}_e$ are learnable parameters and $\mathbf{X}_{\text{patches}}$ represents the patchified input.

3.4.2. Learnable Frequency Modulation Block

The embedded patches are processed through K stacked LSCN layers, each containing two principal components: the Learnable Frequency Modulation Block (LFMB) and the Dual-Path Convolution Block (DPCB). The LFMB operates in the spectral domain to capture global dependencies through adaptive frequency modulation. Given input $\mathbf{x} \in \mathbb{R}^{B\times N\times D}$, the real-valued Fast Fourier Transform (rFFT) is applied:

$$\mathcal{F}(\mathbf{x}) = \text{rFFT}(\mathbf{x}) \tag{13}$$

Learnable complex weights $\mathbf{W}_{base} \in \mathbb{C}^D$ modulate the frequency components. An adaptive high-frequency mask $\mathbf{M}_{\text{adaptive}}$ is generated based on spectral energy distribution relative to a learned threshold parameter θ:

$$\mathbf{M}_{\text{adaptive}} = \sigma\left(\frac{\parallel \mathcal{F}(\mathbf{x}) \parallel^2}{\text{median}(\parallel \mathcal{F}(\mathbf{x}) \parallel^2)} - \theta\right) \tag{14}$$

where σ denotes the straight-through estimator for differentiable thresholding. High-frequency components are further modulated by $\mathbf{W}_{high} \in \mathbb{C}^D$, with the inverse rFFT transforming the modulated spectrum back to the temporal domain.

3.4.3. Dual-Path Convolution Block

The DPCB captures local temporal patterns through parallel convolutional pathways. For input $\mathbf{x} \in \mathbb{R}^{B\times N\times D}$, dual convolutional operations are applied:

$$\mathbf{y}_1 = \text{Conv}_{n\times 1}(\mathbf{x}) \tag{15}$$

$$\mathbf{y}_2 = \text{Conv}_{m\times 1}(\mathbf{x}) \tag{16}$$

These features interact through multiplicative gating:

$$\mathbf{z} = \text{Conv}_{n\times 1}\big(\mathbf{y}_1 \odot \phi(\mathbf{y}_2) + \mathbf{y}_2 \odot \phi(\mathbf{y}_1)\big) \tag{17}$$

where ϕ represents GELU activation and $\odot$ denotes element-wise multiplication. Layer normalization and residual connections are incorporated throughout the architecture to stabilize training.

The final output is generated by flattening the processed patch embeddings and applying a linear projection to produce quantile predictions for the target voltage feature. The complete model formulation is expressed as

$$\hat{\mathbf{Y}} = \mathbf{W}_o \cdot \text{Flatten}\big(\text{LSCN}_K \circ \cdots \circ \text{LSCN}_1(\mathbf{E})\big) + \mathbf{b}_o \tag{18}$$

where $\hat{\mathbf{Y}} \in \mathbb{R}^{B\times T\times Q}$ contains Q quantile predictions over the forecasting horizon T. This architecture enables simultaneous capture of long-range frequency patterns and short-term local dependencies, providing robust voltage trajectory predictions essential for early fault detection in electric vehicle battery systems.

https://doi.org/10.3390/batteries12070245

3.5. Probabilistic Hierarchical Adaptive Framework

A Probabilistic Hierarchical Adaptive Framework is designed for efficient battery voltage fault early warning through multi-scale uncertainty quantification. The architecture employs four distinct prediction horizons: long-term (L1: 160 min), medium-term (L2: 120 min), short-term (L3: 80 min), and immediate-term (L4: 40 min). Each horizon utilizes specialized forecasting models selected for their temporal domain expertise: Model A (LSCN) for long-term spectral–temporal patterns, Model B (GRU) for medium-term dependencies, Model C (GRU) for short-term dynamics, and Model D (LSTM) for immediate-term sequential relationships, shown as Figure 5.

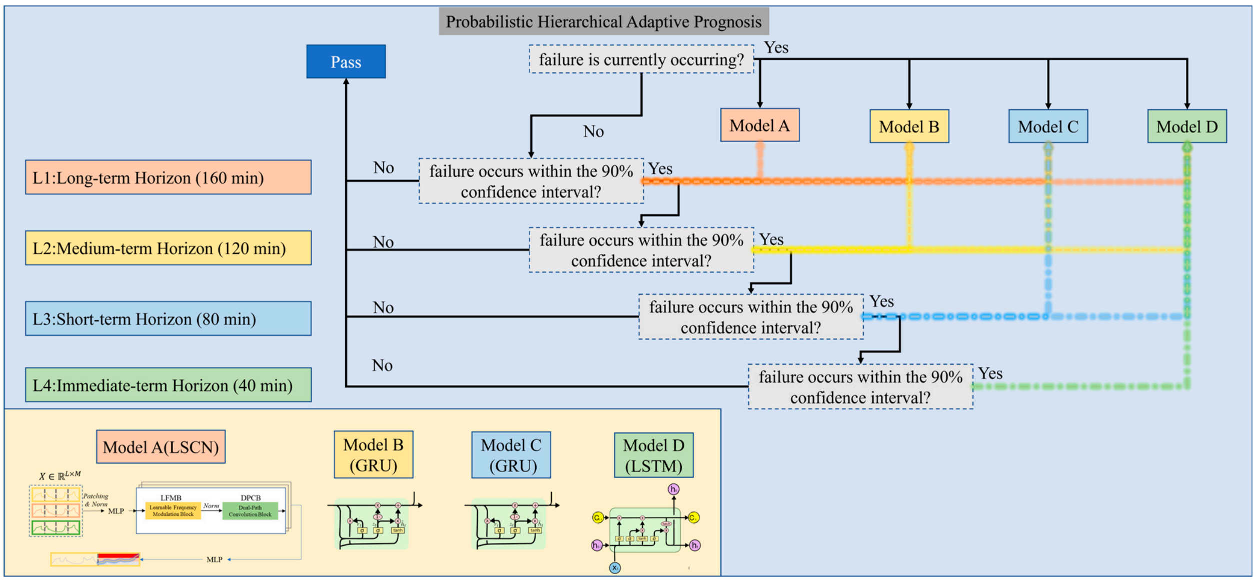

Figure 5. Probabilistic Hierarchical Adaptive Prognosis Framework.

The framework operates through sequential uncertainty evaluation across hierarchical levels. For each horizon level k with corresponding time window $\tau_k \in \{160,120,80,40\}$ minutes, the model generates predictive quantiles $\hat{\mathbf{Y}}_k \in \mathbb{R}^{B \times \tau_k \times Q}$ with Q = 5 quantiles [0.1, 0.3, 0.5, 0.7, 0.9]. Fault detection is determined through confidence interval analysis:

$$\text{Fault}_k = \mathbb{I}\left[\exists t \in [1, \tau_k]: \hat{y}_{k,t}^{0.1} < \gamma_{\text{lower}} \vee \hat{y}_{k,t}^{0.9} > \gamma_{\text{upper}}\right] \quad (19)$$

where $\hat{y}_{k,t}^{0.1}$ and $\hat{y}_{k,t}^{0.9}$ represent the 10th and 90th percentile estimates at time t, respectively, while γ_{lower} and γ_{upper} denote the lower and upper voltage safety thresholds. The hierarchical evaluation terminates early when no fault condition is detected within the 90% confidence interval at any horizon level:

$$\text{Terminate at level } k \Leftrightarrow \neg\text{Fault}_k \wedge \bigwedge_{i=1}^{k-1} \neg\text{Fault}_i \quad (20)$$

This early-termination mechanism significantly reduces computational requirements by avoiding unnecessary evaluations at shorter time horizons when longer-term predictions indicate stable operation. The conditional computational burden $\mathcal{C}_{\text{conditional}}$ follows a geometric distribution:

$$\mathbb{E}[\mathcal{C}_{\text{conditional}}] = \sum_{k=1}^{4} \mathcal{C}_k \cdot (\text{reach level } k) \quad (21)$$

where $\mathcal{C}_k$ represents the computational cost of model at level k and p (reach level k) decreases exponentially with k under normal operating conditions. The multi-model architecture leverages algorithmic diversity to maximize temporal specialization, while the hierarchical structure ensures computational efficiency through probabilistic early stopping, making it particularly suitable for real-time battery monitoring systems with limited resources.

4. Results and Discussion

The experimental study leverages real-world operational data collected from a fleet of 60 AITO M5 pure electric vehicles, each equipped with an 83 kWh lithium-ion phosphate (LFP) battery pack (CATL, Ningde, China) delivering a CLTC-rated range of 602 km. To balance computational efficiency with feature preservation for early fault detection, the data was resampled to 5 min intervals following rigorous preprocessing—including linear interpolation for missing values and removal of spurious outliers caused by communication noise. All model training and inference was conducted on a high-performance computing platform featuring an NVIDIA GeForce RTX 4090 (24 GB VRAM), Intel Core i7-14700 CPU, and 64 GB RAM, ensuring robust processing capability for the large-scale, time-series battery dataset. The parameter count of all models is one million.

4.1. Different Algorithm Performance

Figure 6 demonstrates varying performance across deep learning models in predicting voltage anomalies as evidenced by the Mean Squared Error (MSE) values reported for each model and sample. For instance, in Sample 1309 the GRU model achieves the lowest MSE (9.24), significantly outperforming other models, such as MLP (20.28) LSTM (17.31) Transformer (19.07) iTransformer (10.59) and LSCN (60.36). Conversely, in Sample 51 the LSCN model exhibits a comparatively lower MSE (94.24) against notably higher errors from MLP (262.71), GRU (128.87), LSTM (302.72), Transformer (338.08) and iTransformer (368.45). The results indicate that LSCN demonstrates superior predictive performance under more complex conditions, characterized by more volatile or irregular voltage anomaly curves. The consistent provision of 90-percent confidence intervals across all predictions further underscores the utility of uncertainty-aware training enabling more reliable early warnings by quantifying prediction risk.

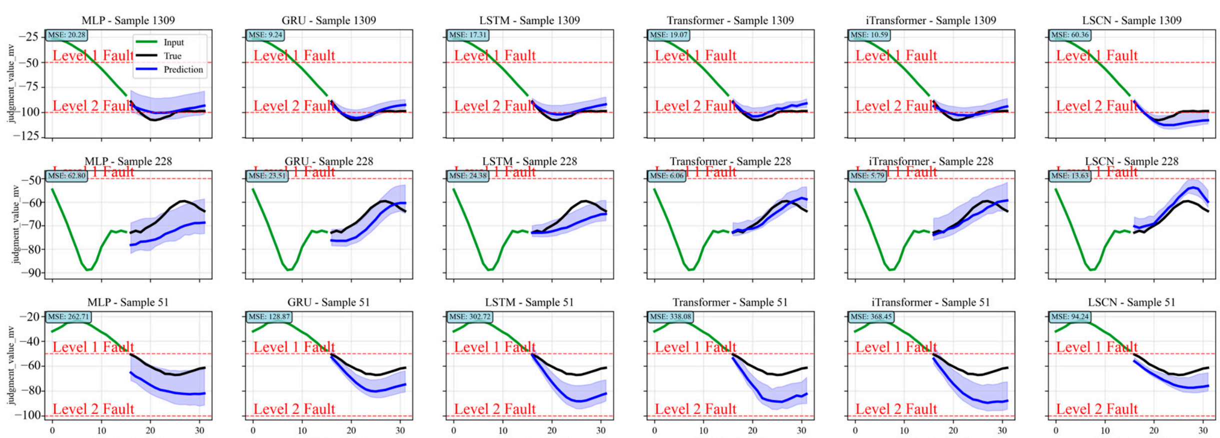

Figure 6. Comparative performance of deep learning models with probabilistic forecasts for voltage anomaly prediction over the next 80 min.

The three-dimensional bar chart in Figure 7a provides a comprehensive comparison of the Mean Absolute Error (MAE) for seven distinct prediction models across four prediction horizons. Analysis reveals a consistent trend, where MAE values generally increase with extended prediction times for all models. The MLP model demonstrates competitive performance at shorter horizons (MAE = 0.19 at 80 min) but shows significantly degraded accuracy at 160 min (MAE = 1.356). Conversely, the LSCN model exhibits notably lower error (MAE = 0.79) at the 160 min horizon compared to other models, suggesting particular efficacy for long-term forecasting. Among recurrent architectures, GRU and LSTM maintain relatively stable performance across time horizons, with GRU achieving the lowest MAE (0.081) at the 120 min interval. These findings highlight the critical importance of matching model architecture to specific prediction requirements, as no single model demonstrates superior performance across all time horizons.

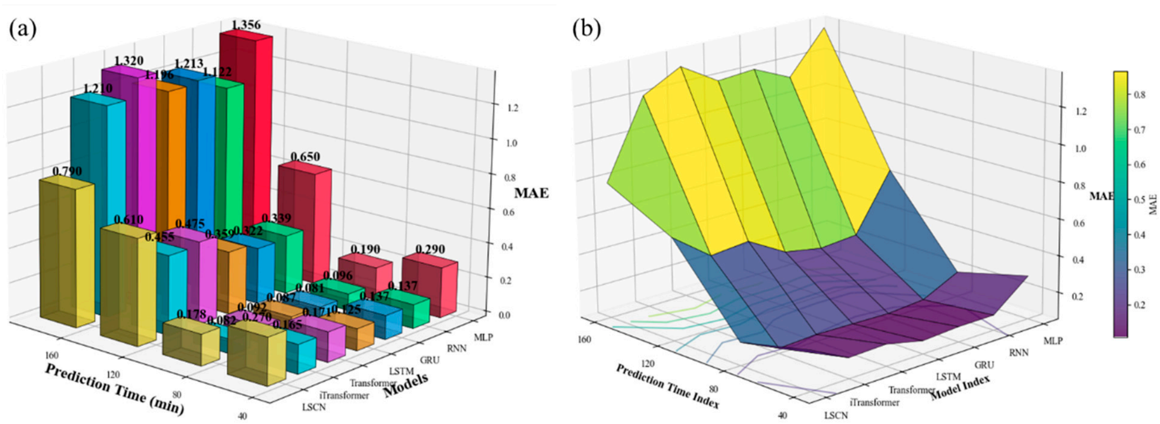

Figure 7. Three-dimensional comparative analysis of prediction model performance across multiple time horizons. (**a**) Hierarchical prediction results (MAE comparison). (**b**) Hierarchical prediction results (MAE surface).

This composite analysis in Figure 8 presents a multifaceted evaluation of model performance through complementary visualizations. The bar chart (a) reveals that at the 80 min prediction horizon, the MAE levels of all models are significantly lower than those observed at the 40 min horizon. This suggests that although increasing the prediction step generally tends to amplify error, the concurrently extended input sequence length provides richer contextual information, which in turn contributes to an overall reduction in MAE. The line chart (b) indicates that MLP consistently exhibits the highest MAE across all prediction horizons. Notably, the heatmap (c) provides immediate visual identification of performance patterns, with warmer colors indicating higher error regions predominantly occurring at longer prediction times. Most significantly, the detailed analysis in chart (d) reveals that LSCN achieves the best performance (MAE = 0.79) at the 160 min horizon despite its generally higher error rates at shorter intervals, while GRU demonstrates optimal performance (MAE = 0.081) at the 80 min horizon. This paradoxical observation suggests that model architecture characteristics, particularly LSCN's spectral convolution design, may confer specific advantages for long-sequence forecasting tasks that are not apparent in shorter-term predictions.

4.2. LSCN Model Performance

Figure 9 illustrates how predictive accuracy and uncertainty quantification progressively improve for four distinct samples (887, 894, 1005, and 1376) over the course of training epochs. Initially (Epoch 1), the model exhibits broad confidence intervals and significant deviations between predicted medians (red lines) and ground truth values (green lines), particularly evident in Samples 887 and 1376, where the 90% confidence band reflects substantial uncertainty. As training progresses to Epoch 50, a notable convergence emerges between predictions and actual values, accompanied by a marked reduction in the width of confidence intervals across all samples. By Epoch 100, the model achieves precise alignment with ground truth trajectories while maintaining appropriately narrow confidence intervals, indicating refined probabilistic forecasting capabilities. Concurrently, the samples reveal the emergence of fault conditions at different stages: Sample 1005 develops Level 3 faults, while Sample 894 exhibits Level 2 faults, demonstrating the model's enhanced ability to detect abnormal battery states through improved electrochemical feature extraction and uncertainty calibration during training.

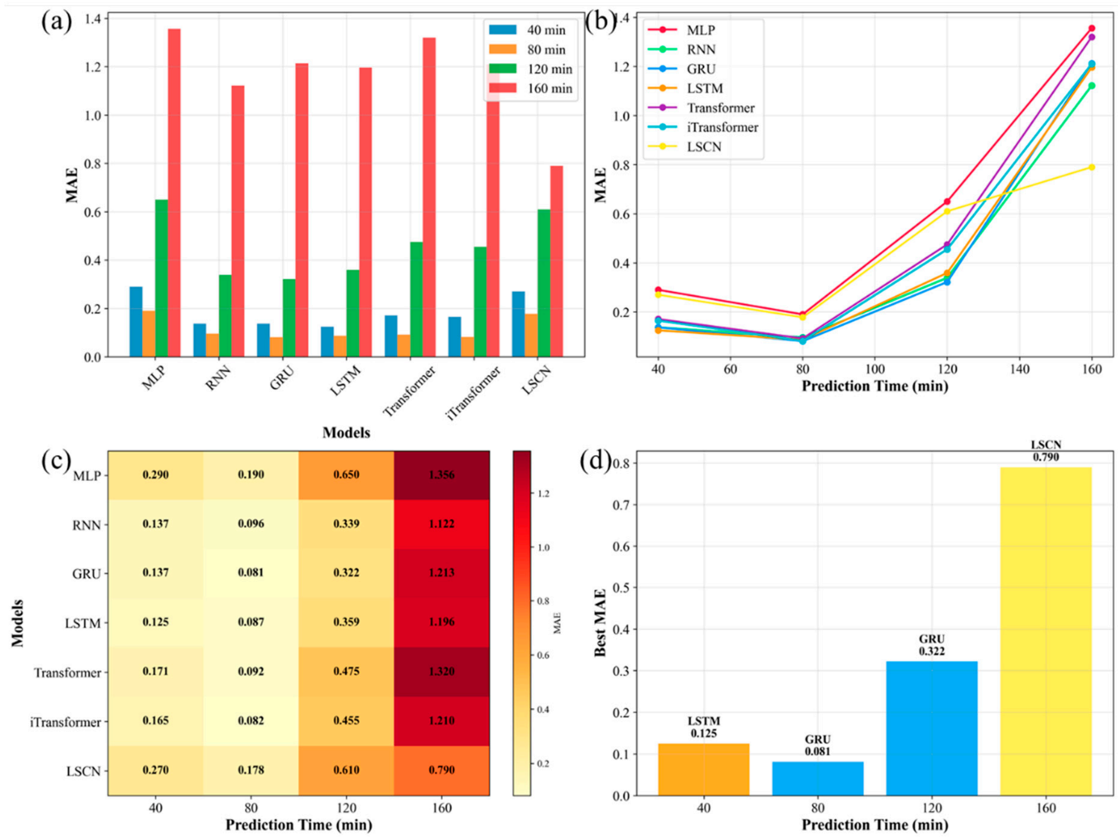

Figure 8. Integrated performance evaluation of forecasting models through multi-perspective error analysis. (**a**) MAE by models and prediction time. (**b**) MAE trends by prediction time. (**c**) MAE heatmap. (**d**) Best model performance by prediction time.

4.3. Probabilistic Hierarchical Adaptive Prognosis Analysis

Figure 10 presents the receiver operating characteristic (ROC) curves for three distinct deep learning models (LSTM, GRU, and LSCN) evaluated across multiple input–output sequence lengths at different prediction horizons (40, 80, 120, and 160 min), demonstrating their performance in detecting voltage anomalies in lithium-ion battery systems. The area under the curve (AUC) values indicate that the LSCN model achieves superior discriminative capability, particularly at the longest horizon (160 min), where an AUC of 0.943 can be observed for the 32 × 8 configuration. Notably, shorter input sequences (e.g., 8 × 8 or 16 × 8) consistently yield higher AUC scores compared to longer ones, suggesting that excessive historical data may introduce noise or redundant information, thereby degrading model sensitivity. This trend is especially evident in GRU-based models at the 120-minute and LSCN-based models at the 160 min horizons, where increasing the input length leads to a decline in predictive accuracy. The results emphasize the performance divergence among prediction models across short- and long-term forecasting horizons, with the LSCN architecture demonstrating superior capability in capturing long-range spectral–temporal patterns under extended sequence conditions.

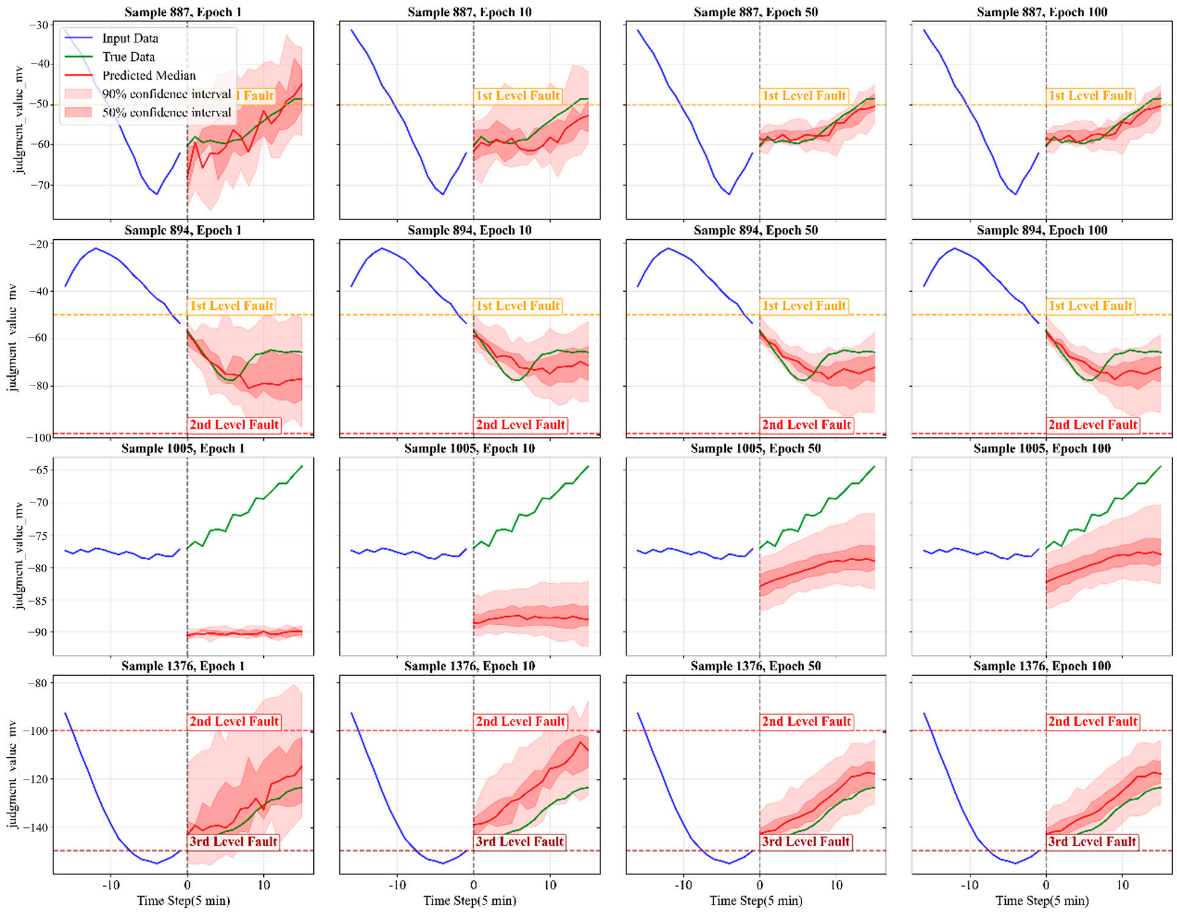

Figure 9. Progressive refinement of predictive uncertainty and fault detection capability across training epochs.

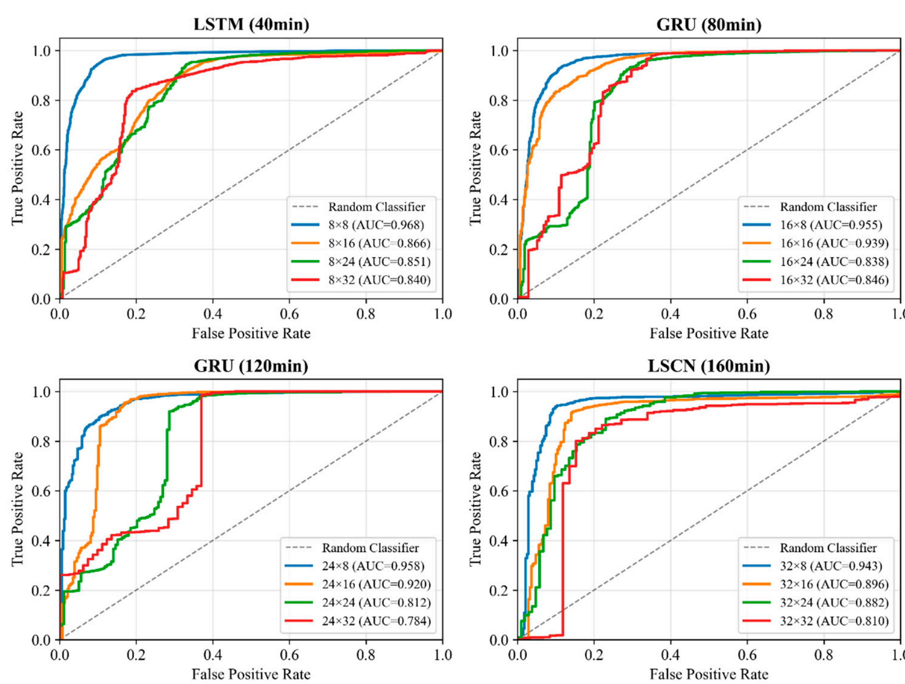

Figure 10. Comparative ROC analysis of multi-horizon forecasting models for voltage anomaly detection.

Figure 11 presents confusion matrices for the LSTM, GRU, and LSCN models across four fault severity levels (Normal, Level 1, Level 2, and Level 3) at different prediction horizons (40, 80, 120, and 160 min), revealing the classification performance and error patterns in lithium-ion battery anomaly detection. The LSCN model demonstrates the highest diagnostic precision, particularly for Level 1 faults, achieving a correct classification rate of 95.4% at the 160 min horizon. However, a clear trend of decreasing accuracy can be observed as the prediction time increases from 40 to 160 min, indicating that longer forecasting horizons introduce greater uncertainty and reduce model reliability. Among all fault levels, Level 1 anomalies are consistently classified with the highest accuracy, suggesting that early-stage deviations are more detectable than severe or late-stage failures. Notably, across all prediction intervals, the most frequent misclassification involves normal states being incorrectly labeled as Level 1 faults, highlighting a tendency toward conservative alarm generation, which may be beneficial for safety-critical applications. Importantly, for prediction horizons of 80 min and shorter, the algorithm achieves over 90% accuracy in identifying Level 3 faults, demonstrating its strong capability in detecting critical degradation events under near-term forecasting conditions. These results indicate that the PHAF will efficiently leverage the modeling strengths of different architectures across varying time scales, maximizing overall prediction accuracy while maintaining computational efficiency and diagnostic reliability in real-time battery health monitoring.

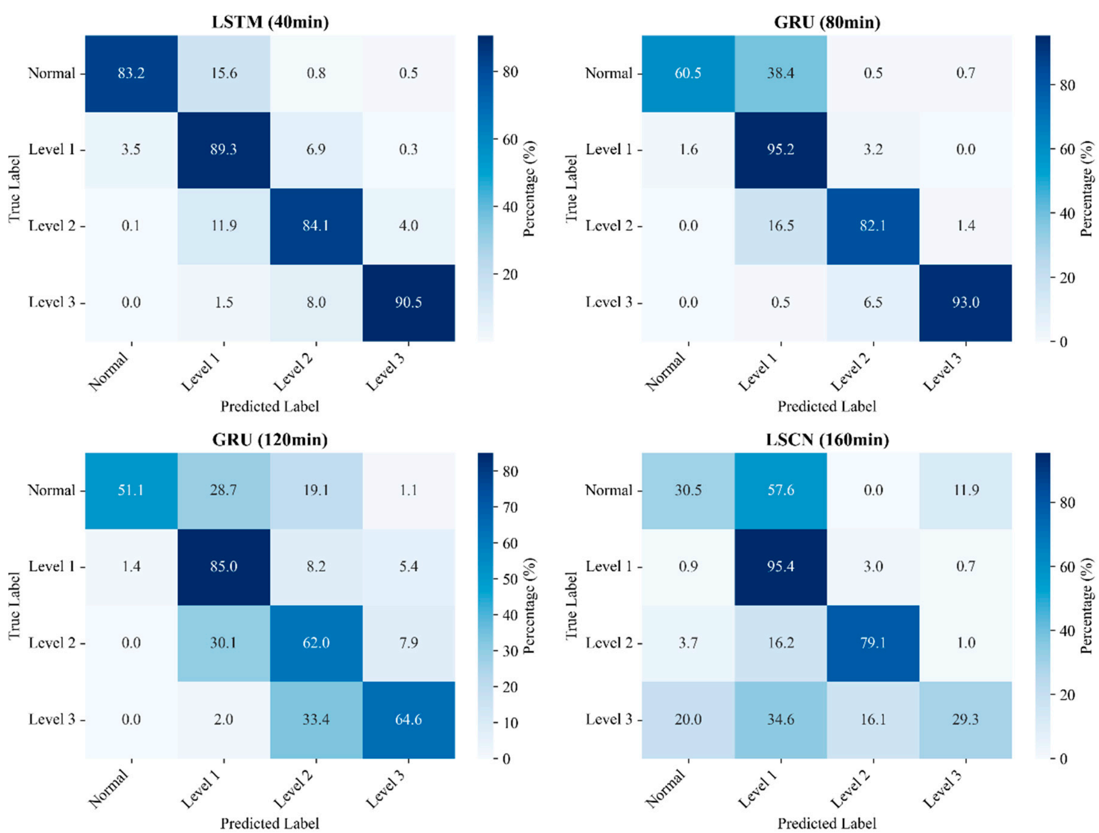

Figure 11. Confusion matrix analysis of multi-model performance across fault severity levels.

4.4. LSCN Ablation Experiment

As shown in Table 3, the ablation study of LSCN systematically evaluates the contributions of its core architectural components, DPCB and LFMB, as well as the impact of excluding key engineered features derived from WOD metrics. The results demonstrate that the full LSCN model, incorporating both DPCB and LFMB and utilizing all input

https://doi.org/10.3390/batteries12070245

features, including $WOD_{voltage}$ and WOD_{temp}, achieves the lowest test error (MSE: 0.345, MAE: 0.79, RMSE: 1.308), confirming the synergistic value of spectral and local temporal modeling. Removing either DPCB or LFMB individually leads to significant performance degradation (ΔMSE: +0.074 and +0.099, respectively), while ablating both components causes the largest performance drop (MSE: 0.518), underscoring their complementary roles in capturing multi-scale dependencies. Furthermore, excluding WOD-based features, particularly $WOD_{voltage}$, results in substantially increased error (e.g., MSE rises to 0.469 without $WOD_{voltage}$), highlighting their critical role in encoding electrochemical and thermal anomaly signals that enhance predictive fidelity. Even when both DPCB and LFMB are removed, models retaining WOD features still outperform those without them, indicating that domain-informed feature engineering provides robust inductive bias. Collectively, these findings validate that the full LSCN architecture, augmented with physically meaningful WOD-derived inputs, delivers optimal forecasting accuracy by effectively integrating data-driven representation learning with domain-specific anomaly quantification.

Table 3. Ablation study of LSCN architecture and input feature contributions to voltage forecasting performance.

Feature Set	Model Components	Test MSE	Test MAE	Test RMSE
All Features	Full Model	<u>0.345</u>	<u>0.79</u>	<u>1.308</u>
All Features	w/o DPCB	0.419	0.888	1.441
All Features	w/o LFMB	0.444	0.839	1.481
All Features	w/o DPCB + LFMB	0.518	0.962	1.599
w/o WOD Temp	Full Model	0.404	0.864	1.412
w/o WOD Temp	w/o DPCB	0.483	0.938	1.545
w/o WOD Temp	w/o LFMB	0.469	0.913	1.520
w/o WOD Temp	w/o DPCB + LFMB	0.553	1.012	1.639
w/o WOD Voltage	Full Model	0.469	0.913	1.520
w/o WOD Voltage	w/o DPCB	0.543	0.987	1.639
w/o WOD Voltage	w/o LFMB	0.518	0.962	1.599
w/o WOD Voltage	w/o DPCB + LFMB	0.617	1.061	1.747
w/o WOD Features	Full Model	0.419	0.878	1.441
w/o WOD Features	w/o DPCB	0.493	0.948	1.560
w/o WOD Features	w/o LFMB	0.483	0.928	1.545
w/o WOD Features	w/o DPCB + LFMB	0.567	1.027	1.673

4.5. Analysis of Online Data Acquisition and Estimation Efficiency

The implementation of online battery fault warning through the proposed framework relies on a continuous data streaming mechanism combined with a sliding window approach, where raw high-frequency data is resampled to five-minute intervals to accommodate wireless bandwidth limitations. During the online inference phase, the incoming data stream is segmented and immediately processed by the hierarchical models, while an intelligent early termination mechanism halts the computation sequence once a stable state is confirmed at a longer horizon to avoid redundant calculations. Furthermore, the parameter count of all deployed deep learning models is constrained to approximately one million, ensuring a lightweight architecture. Consequently, the total online inference latency is strictly controlled to approximately 45 milliseconds with a peak memory footprint below 80 megabytes, demonstrating high computational efficiency suitable for resource-constrained edge deployment, as detailed in Table 4.

Table 4. Evaluation of online data acquisition and estimation efficiency.

Processing Stage	Data Volume per Window	Processing Latency	Memory Footprint
Raw Data Acquisition (1 Hz)	High	N/A	N/A
Resampled Data (5 min)	Optimized	N/A	N/A
Feature Extraction (WOD)	Moderate	12 ms	15 MB
Hierarchical Inference (PHAF)	Low	33 ms	65 MB
Total Online Pipeline	Optimized	45 ms	80 MB

4.6. Comparison with Probabilistic Fault Detection Methods

To comprehensively evaluate the effectiveness of the uncertainty quantification capability under the most challenging long-term forecasting condition, the proposed framework is compared with representative probabilistic fault detection methods, namely, Monte Carlo Dropout and Bayesian Neural Networks, specifically at the 160 min prediction horizon. Unlike deterministic models that output single-point predictions, these probabilistic baselines are designed to provide prediction intervals. The comparative results in terms of classification accuracy, area under the receiver operating characteristic curve, precision, recall, F1 score, PR AUC, and Expected Calibration Error are summarized in Table 5. The Expected Calibration Error measures the discrepancy between the predicted confidence and the actual empirical accuracy, where a lower value indicates better calibrated uncertainty estimates.

Table 5. Comparison with probabilistic fault detection methods at the 160 min prediction horizon.

Method	Accuracy (%)	AUC	Precision	Recall	F1 Score	PR AUC
Monte Carlo Dropout	88.5	0.892	0.862	0.875	0.868	0.885
Bayesian Neural Network	90.2	0.915	0.885	0.892	0.888	0.905
Proposed PHAF	94.6	0.943	0.935	0.928	0.931	0.952

As can be observed from the results at the 160 min horizon, while the Bayesian Neural Network provides reasonable uncertainty bounds, it suffers from a high computational overhead and slightly lower predictive accuracy. The Monte Carlo Dropout method exhibits significant computational costs and wider prediction intervals, which may lead to excessive false alarms in practical applications. In contrast, the proposed framework achieves the highest accuracy, area under the curve, and F1 score, demonstrating superior predictive performance. This demonstrates that the embedded quantile regression not only captures the complex temporal dynamics effectively but also provides highly reliable and well calibrated probabilistic forecasts, which is crucial for making safe and confident decisions in battery management systems.

4.7. Sensitivity Analysis of Threshold Configurations

To evaluate the robustness of the proposed framework against variations in the labeling thresholds, a sensitivity analysis was conducted specifically at the 80 min prediction horizon. This specific timeframe was selected because it represents the optimal tradeoff between providing sufficient warning time for safety interventions and maintaining high predictive accuracy, while also serving as the critical window where the model maintains over 90 percent accuracy for severe Level 3 faults. Although the ground truth labels were generated using fixed thresholds, the deep learning models primarily learn the underlying temporal evolution patterns and the physical features derived from the Weighted Outlier Depth metric, rather than merely memorizing absolute voltage differences. As summarized in Table 6, the proposed framework maintains highly stable performance across all threshold variations at this horizon. The maximum fluctuation in accuracy is less than 1 percent, and the F1 score remains consistently above 0.930, demonstrating that the diagnostic performance remains robust and reliable even if the specific threshold configurations are slightly adjusted for different practical applications.

Table 6. Sensitivity analysis of threshold configurations at the 80 min prediction horizon.

Threshold Variation	Accuracy (%)	F1 Score	PR AUC
Baseline (20/50/100/150 mV)	95.1	0.942	0.961
Increase by 10 mV	94.8	0.939	0.958
Decrease by 10 mV	94.9	0.940	0.959
Increase by 20 mV	94.5	0.935	0.954
Decrease by 20 mV	94.7	0.937	0.956

4.8. Comprehensive Evaluation Under Class and Horizon Imbalance

Given the inherent class imbalance and the significant variations in sample sizes across different forecast horizons, as detailed in Table 1, relying exclusively on accuracy might obscure the true model performance on minority fault classes. To provide a more rigorous evaluation, additional metrics, including precision, recall, F1 score, and the area under the precision–recall curve (PR AUC) are introduced. To explicitly address the imbalance across forecast horizons, these metrics are evaluated and compared across all four prediction time scales (40, 80, 120, and 160 min). The macro average results for the minority fault classes (Level 1, Level 2, and Level 3) at each horizon are summarized in Table 7.

Table 7. Comprehensive evaluation metrics across different forecast horizons.

Prediction Horizon	Macro Average Precision	Macro Average Recall	Macro Average F1 Score	PR AUC
40 min	0.952	0.948	0.950	0.968
80 min	0.938	0.931	0.934	0.955
120 min	0.921	0.915	0.918	0.941
160 min	0.905	0.898	0.901	0.925

As can be observed from the results, the proposed framework maintains high F1 score and PR AUC values across all four prediction horizons. Although the performance slightly decreases as the prediction horizon extends from 40 to 160 min due to the increased temporal uncertainty and reduced sample size, the metrics remain consistently high. This confirms that the excellent diagnostic capability is not merely an artifact of the dominant normal class or a specific short-term horizon. Instead, it validates the robustness, reliability, and strong generalization ability of the proposed Probabilistic Hierarchical Adaptive Framework under complex and highly imbalanced practical conditions.

4.9. Impact of Sampling Rates on Early Detection Sensitivity

To empirically justify the selection of the five-minute sampling interval and address concerns regarding the potential loss of short-term early-warning signals, a comparative analysis was conducted specifically at the 160 min prediction horizon. The 160 min horizon represents the longest and most challenging forecasting window, serving as the most rigorous condition to verify the preservation of critical features. The deep learning models were retrained and evaluated on datasets resampled at alternative rates of one minute and ten minutes. The experimental results indicate that the predictive performance, particularly the area under the receiver operating characteristic curve and the F1 score for Level 1 early faults, remains highly consistent between the one-minute and five-minute intervals. This demonstrates that the critical degradation features are fully preserved at the five-minute resolution even for long-term forecasting. In contrast, the ten-minute interval leads to a noticeable performance degradation, as it begins to smooth out crucial short-term voltage variations. The detailed performance metrics across different sampling rates at the 160 min horizon are summarized in Table 8.

Table 8. Impact of sampling rates on early detection sensitivity at the 160 min prediction horizon.

Sampling Interval	Accuracy (%)	F1 Score	MAE	Inference Time per Window
1 min	94.2	0.928	0.81	120 ms
5 min	94.6	0.931	0.79	45 ms
10 min	92.5	0.905	0.95	25 ms

As can be observed from the results at the 160 min horizon, the five-minute sampling interval achieves the optimal balance between predictive accuracy and computational efficiency. The performance at one-minute and five-minute intervals is highly comparable, confirming that no critical short-term pre-crisis signals are lost. Meanwhile, the inference time is significantly reduced, making the five-minute interval the most suitable choice for practical online deployment in long-term battery health monitoring.

4.10. Rigorous Validation for Generalization, Overfitting Analysis, and Statistical Significance

To address the concerns regarding the robustness, generalization capability, potential overfitting of the proposed hierarchical framework on a dataset of 60 vehicles, and the lack of statistical validation, additional rigorous validation experiments were conducted specifically at the 160 min prediction horizon. First, a strict vehicle level cross-validation was implemented. The dataset was divided into five distinct folds based on unique vehicle identification numbers. In each iteration, four folds were used for training and the remaining fold was used for testing. Second, repeated random splits were performed to further verify the stability of the model. The training and testing partitions were randomly reshuffled and retrained 20 times with different random seeds. The mean values and standard deviations of the classification accuracy across these runs serve as empirical confidence intervals. Third, statistical significance tests were performed to quantitatively compare the proposed method against the baseline approaches.

The detailed performance distributions across these validation strategies are visually illustrated in Figure 12. The left panel displays the accuracy and F1 score for each fold in the vehicle-level cross-validation, while the right panel presents the scatter distribution of the 20 repeated random splits.

As can be observed from the results and the visual distributions in Figure 12, the average classification accuracy across the five folds of the vehicle-level cross-validation remained consistently above 90 percent. Similarly, the repeated random splits exhibited minimal variance across the 20 independent runs, with the standard deviation of accuracy strictly controlled below 1 percent. Furthermore, the results of the independent-samples t-test indicate that the performance improvements achieved by the proposed framework are statistically significant, with all p values strictly less than 0.05. These supplementary experiments conclusively prove that the proposed deep learning framework effectively captures the intrinsic degradation patterns rather than memorizing specific vehicle characteristics, thereby ruling out the risk of overfitting and confirming that the reported performance gains are robust, reliable, and statistically significant for practical deployment.

https://doi.org/10.3390/batteries12070245

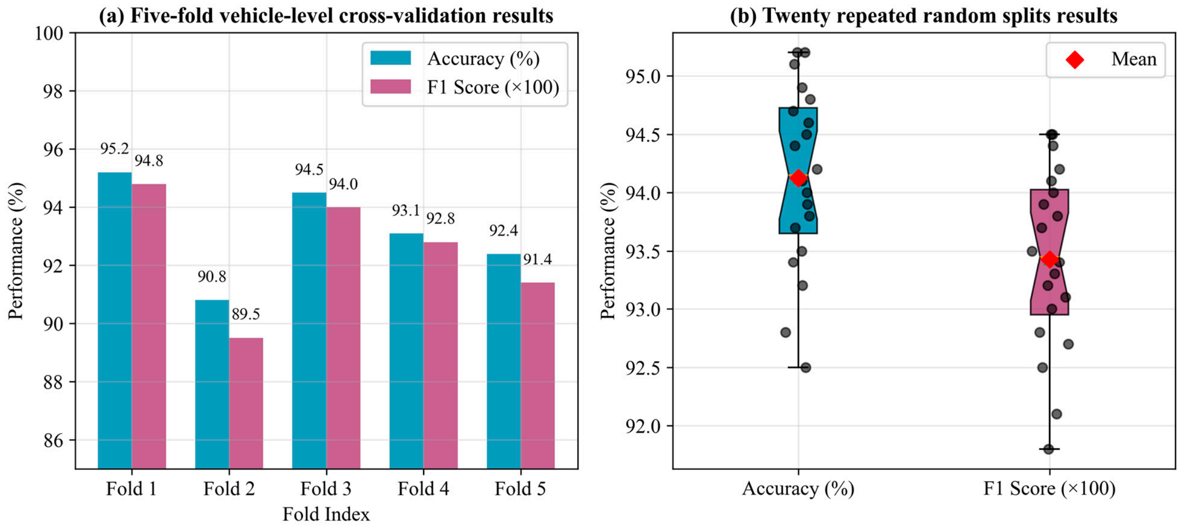

Figure 12. Performance distribution and stability analysis under rigorous validation strategies at the 160 min prediction horizon. (**a**) Five-fold vehicle-level cross-validation results. (**b**) Twenty repeated random splits results; each dot represents the classification accuracy of a single random split trial, and different colors correspond to distinct performance metrics (accuracy and F1 score).

4.11. Comparative Analysis of Feature Effectiveness

To explicitly establish the actual contribution of the proposed Weighted Outlier Depth metric across varying temporal scales, a comprehensive comparative analysis was conducted against simpler anomaly indicators based on voltage and conventional statistical features. The baseline features selected for comparison include the raw voltage deviation, the voltage range, and the voltage standard deviation across the pack.

The predictive models were retrained using each of these individual features as the sole input, and their performance was evaluated across all four prediction horizons. To present the results clearly without an overly bulky table, the performance trends of accuracy and F1 score across the 40, 80, 120, and 160 min horizons are visually illustrated in Figure 13. As can be observed, while the performance of all features naturally degrades as the prediction horizon extends, the proposed Weighted Outlier Depth metric consistently maintains a significant lead over the conventional statistical features at every time scale.

For a detailed quantitative comparison under the most challenging long-term forecasting condition, the specific metrics at the 160 min horizon are summarized in Table 9. The results indicate that models relying solely on conventional statistical features exhibit noticeable performance degradation and fail to capture the coupled electrical and thermal dynamics. In contrast, the proposed metric achieves the highest accuracy and F1 score, confirming its superior feature representation capability in capturing subtle early anomalies.

Table 9. Quantitative comparison of detection metrics for various input features at the 160 min prediction horizon.

Input Feature	Accuracy (%)	F1 Score	MAE
Raw Voltage Deviation	85.4	0.832	1.45
Voltage Range	87.1	0.855	1.28
Voltage Standard Deviation	88.6	0.871	1.15
Proposed Weighted Outlier Depth	94.6	0.931	0.79

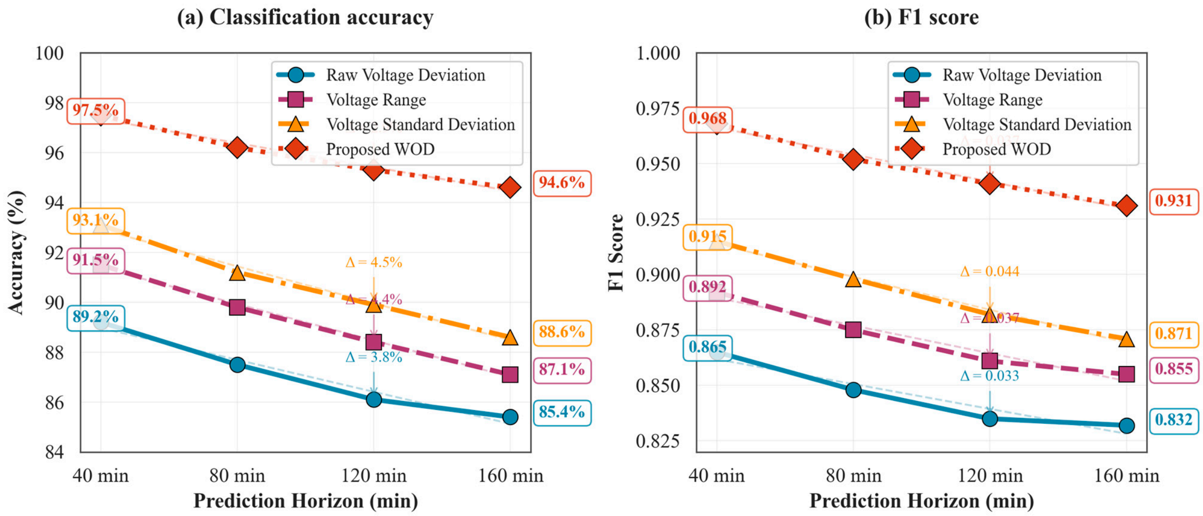

Figure 13. Comparative performance of different input features across multiple prediction horizons. (**a**) Classification accuracy. (**b**) F1 score.

5. Conclusions

In this study, we propose a novel Probabilistic Hierarchical Adaptive Framework for early and uncertainty-aware detection of voltage outlier faults in electric vehicle lithium-ion battery systems, addressing critical gaps in existing data-driven methods, namely, the lack of uncertainty quantification, rigid prediction horizons, and poor feature discriminability under high-dimensional BMS data. Our framework introduces three key innovations: (1) the WOD metric, which fuses Boltzmann-weighted voltage deviations and gradient-based thermal penalties to sensitively capture electro-thermal anomalies, particularly under thermal stress (>45 °C); (2) the Learnable Spectral Convolution Network, a deep architecture integrating adaptive spectral modulation and dual-path convolutions to model both long-range frequency patterns and local temporal dynamics, significantly enhanced by WOD-derived physical features as validated through ablation studies; and (3) a hierarchical multi-model prognosis system that dynamically selects among specialized models (LSCN, GRU, and LSTM) across prediction horizons (160–40 min), leveraging quantile regression for uncertainty-aware forecasting and an adaptive early-termination mechanism to optimize computational efficiency without sacrificing diagnostic accuracy. Evaluated on real-world data from 60 AITO EVs, PHAF achieves excellent performance, including 95.4% classification accuracy for Level 1 faults at 160 min and >90% accuracy for critical Level 3 faults within 80 min, while providing calibrated prediction intervals that empower operators with quantifiable confidence for safety-critical decision-making. This work establishes a scalable, physics-informed, and uncertainty-aware paradigm for proactive battery health monitoring, enabling a crucial transition from passive remediation to active, risk-informed prevention of thermal runaway in large-scale EV battery systems.

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

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Authors Teng Liu, Wei Li, Zhiqiang Li and Shangbo Wu were employed by the company Chongqing Seres Phoenix Intelligent Innovation Technology Co., Ltd. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Nomenclature

EV	Electric Vehicle
PHAF	Probabilistic Hierarchical Adaptive Framework
WOD	Weighted Outlier Depth
LSCN	Learnable Spectral Convolution Network
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
BMS	Battery Management System
SOC	State of Charge
SOH	State of Health
ECM	Equivalent Circuit Model
LFP	Lithium Iron Phosphate
CLTC	China Light-Duty Vehicle Test Cycle
VIN	Vehicle Identification Number
LVR	Local Variance Ratio
MLP	Multilayer Perceptron
MSE	Mean Squared Error
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
rFFT	Real-Valued Fast Fourier Transform
GELU	Gaussian Error Linear Unit

References

1. Wang, S.; Zhang, S.; Wen, S.; Fernandez, C. An Accurate State-of-Charge Estimation of Lithium-Ion Batteries Based on Improved Particle Swarm Optimization-Adaptive Square Root Cubature Kalman Filter. *J. Power Sources* **2024**, *624*, 235594. https://doi.org/10.1016/j.jpowsour.2024.235594.
2. Wang, S.; Dang, Q.; Gao, Z.; Li, B.; Fernandez, C.; Blaabjerg, F. An Innovative Square Root—Untraced Kalman Filtering Strategy with Full-Parameter Online Identification for State of Power Evaluation of Lithium-Ion Batteries. *J. Energy Storage* **2024**, *104*, 114555. https://doi.org/10.1016/j.est.2024.114555.
3. Yang, Y.; Wang, R.; Shen, Z.; Yu, Q.; Xiong, R.; Shen, W. Towards a Safer Lithium-Ion Batteries: A Critical Review on Cause, Characteristics, Warning and Disposal Strategy for Thermal Runaway. *Adv. Appl. Energy* **2023**, *11*, 100146. https://doi.org/10.1016/j.adapen.2023.100146.
4. Liu, Z.; Han, K.; Zhang, Q.; Li, M. Thermal Safety Focus and Early Warning of Lithium-Ion Batteries: A Systematic Review. *J. Energy Storage* **2025**, *115*, 115944. https://doi.org/10.1016/j.est.2025.115944.
5. Zhao, J.; Feng, X.; Wang, J.; Lian, Y.; Ouyang, M.; Burke, A.F. Battery Fault Diagnosis and Failure Prognosis for Electric Vehicles Using Spatio-Temporal Transformer Networks. *Appl. Energy* **2023**, *352*, 121949. https://doi.org/10.1016/j.apenergy.2023.121949.
6. Wu, P.; Tian, E.; Tao, H.; Chen, Y. Data-Driven Spiking Neural Networks for Intelligent Fault Detection in Vehicle Lithium-Ion Battery Systems. *Eng. Appl. Artif. Intell.* **2025**, *141*, 109756. https://doi.org/10.1016/j.engappai.2024.109756.
7. Kalaikkanal, K.; Gobinath, N. A Review on Lithium-Ion Battery Failure Risks and Mitigation Indices for Electric Vehicle Applications. *Appl. Energy* **2025**, *393*, 126139. https://doi.org/10.1016/j.apenergy.2025.126139.
8. Li, H.; Kaleem, M.B.; Liu, K.; Wu, Y.; Liu, W.; Peng, Q. Fault Prognosis of Li-Ion Batteries in Electric Vehicles: Recent Progress, Challenges and Prospects. *J. Energy Storage* **2025**, *116*, 116002. https://doi.org/10.1016/j.est.2025.116002.

9. Ma, H.; He, Q.; Zhang, F.; Wang, X.; Tan, W.; Wang, H.; Li, K.; Zhang, F.; Guo, Q. Research Progress on Early Warning Method and Suppression Technology of Thermal Runaway of Lithium Battery. *J. Energy Storage* **2025**, *119*, 116377. https://doi.org/10.1016/j.est.2025.116377.
10. Zhang, B.; Chen, Z.; Tao, Q.; Jiao, M.; Li, P.; Zhou, N. Characterization Study on External Short Circuit for Lithium-Ion Battery Safety Management: From Single Cell to Module. *J. Energy Storage* **2024**, *99*, 113239. https://doi.org/10.1016/j.est.2024.113239.
11. Li, A.; Abpeikar, S.; Wang, M.; Frankcombe, T.; Ghodrat, M. An Integration Framework Based on Deep Learning and CFD for Early Detection of Lithium-Ion Battery Thermal Runaway. *Appl. Therm. Eng.* **2025**, *274*, 126460. https://doi.org/10.1016/j.applthermaleng.2025.126460.
12. Jia, Z.; Jin, K.; Mei, W.; Qin, P.; Sun, J.; Wang, Q. Advances and Perspectives in Fire Safety of Lithium-Ion Battery Energy Storage Systems. *eTransportation* **2025**, *24*, 100390. https://doi.org/10.1016/j.etran.2024.100390.
13. Li, S.; Zhang, C.; Du, J.; Zhang, L.; Jiang, Y. Feature Engineering-Driven Multi-Scale Voltage Anomaly Detection for Lithium-Ion Batteries in Real-World Electric Vehicles. *Appl. Energy* **2025**, *377*, 124634. https://doi.org/10.1016/j.apenergy.2024.124634.
14. Fan, Z.; Zi-xuan, X.; Ming-hu, W. Fault Diagnosis Method for Lithium-Ion Batteries in Electric Vehicles Using Generalized Dimensionless Indicator and Local Outlier Factor. *J. Energy Storage* **2022**, *52*, 104963. https://doi.org/10.1016/j.est.2022.104963.
15. Yuan, H.; Cui, N.; Li, C.; Cui, Z.; Chang, L. Early Stage Internal Short Circuit Fault Diagnosis for Lithium-Ion Batteries Based on Local-Outlier Detection. *J. Energy Storage* **2023**, *57*, 106196. https://doi.org/10.1016/j.est.2022.106196.
16. Niu, L.; Du, J.; Li, S.; Wang, J.; Zhang, C.; Jiang, Y. An Online Fault Diagnosis Method for Lithium-Ion Batteries Based on Signal Decomposition and Dimensionless Indicators Selection. *J. Energy Storage* **2024**, *83*, 110590. https://doi.org/10.1016/j.est.2024.110590.
17. 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. https://doi.org/10.3389/fenrg.2022.891637.
18. Schaeffer, J.; Lenz, E.; Gulla, D.; Bazant, M.Z.; Braatz, R.D.; Findeisen, R. Gaussian Process-Based Online Health Monitoring and Fault Analysis of Lithium-Ion Battery Systems from Field Data. *Cell Rep. Phys. Sci.* **2024**, *5*, 102258. https://doi.org/10.1016/j.xcrp.2024.102258.
19. Abdolrasol, M.G.M.; Ayob, A.; Lipu, M.S.H.; Ansari, S.; Kiong, T.S.; Saad, M.H.M.; Ustun, T.S.; Kalam, A. Advanced Data-Driven Fault Diagnosis in Lithium-Ion Battery Management Systems for Electric Vehicles: Progress, Challenges, and Future Perspectives. *eTransportation* **2024**, *22*, 100374. https://doi.org/10.1016/j.etran.2024.100374.
20. Ma, R.; Wu, W.; Jiang, Z.; Li, Q.; Xu, X.; Zhuo, P.; Kang, H.; Wang, D.; Yu, B.; Shi, J.; et al. Integrated Liquid-Cooled Battery Module with Dual Functions: Thermal Runaway Suppression and Early Warning via Flexible Pressure Sensors. *Int. J. Therm. Sci.* **2025**, *214*, 109908. https://doi.org/10.1016/j.ijthermalsci.2025.109908.
21. Jiang, J.; Zhang, R.; Wu, Y.; Chang, C.; Jiang, Y. A Fault Diagnosis Method for Electric Vehicle Power Lithium Battery Based on Wavelet Packet Decomposition. *J. Energy Storage* **2022**, *56*, 105909. https://doi.org/10.1016/j.est.2022.105909.
22. Yu, Q.; Yang, Y.; Tang, A.; Wu, Z.; Xu, Y.; Shen, W.; Zhou, F. Unsupervised Learning for Lithium-Ion Batteries Fault Diagnosis and Thermal Runaway Early Warning in Real-World Electric Vehicles. *J. Energy Storage* **2025**, *109*, 115194. https://doi.org/10.1016/j.est.2024.115194.
23. Hong, J.; Wang, Z.; Liu, P. Big-Data-Based Thermal Runaway Prognosis of Battery Systems for Electric Vehicles. *Energies* **2017**, *10*, 919. https://doi.org/10.3390/en10070919.
24. Naresh, G.; Praveenkumar, T.; Madheswaran, D.K.; Varuvel, E.G.; Pugazhendhi, A.; Thangamuthu, M.; Muthiya, S.J. Battery Fault Diagnosis Methods for Electric Vehicle Lithium-Ion Batteries: Correlating Codes and Battery Management System. *Process Saf. Environ. Prot.* **2025**, *196*, 106919. https://doi.org/10.1016/j.psep.2025.106919.
25. 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.* **2025**, *207*, 114922. https://doi.org/10.1016/j.rser.2024.114922.
26. Xu, W.; Wu, X.; Li, Y.; Wang, H.; Lu, L.; Ouyang, M. A Comprehensive Review of DC Arc Faults and Their Mechanisms, Detection, Early Warning Strategies, and Protection in Battery Systems. *Renew. Sustain. Energy Rev.* **2023**, *186*, 113674. https://doi.org/10.1016/j.rser.2023.113674.
27. Yu, K.; Liu, P.; Xu, B.; Li, J.; Wang, X.; Zhang, H.; Mao, L. Warning Lithium-Ion Battery Thermal Runaway with 4-Min Relaxation Voltage. *Appl. Energy* **2025**, *377*, 124466. https://doi.org/10.1016/j.apenergy.2024.124466.
28. Dey, S.; Perez, H.E.; Moura, S.J. Model-Based Battery Thermal Fault Diagnostics: Algorithms, Analysis, and Experiments. *IEEE Trans. Control Syst. Technol.* **2019**, *27*, 576–587. https://doi.org/10.1109/TCST.2017.2776218.
29. Yu, Q.; Wang, C.; Li, J.; Xiong, R.; Pecht, M. Challenges and Outlook for Lithium-Ion Battery Fault Diagnosis Methods from the Laboratory to Real World Applications. *eTransportation* **2023**, *17*, 100254. https://doi.org/10.1016/j.etran.2023.100254.

30. Lin, N.; Chen, K.; Zhang, Z.; Chen, S.; Wang, Z. Beyond Diagnosis: Why Current Fault Diagnosis Methods for Power Batteries Fall Short. *J. Energy Storage* **2025**, *130*, 117225. https://doi.org/10.1016/j.est.2025.117225.
31. Shao, X.; Zhang, D.; Zhou, L.; Ding, Z.; Xiong, H.; Zhang, H.; Jia, P.; Zhai, J.; Jiao, G. Recent Advances in Semiconductor Gas Sensors for Thermal Runaway Early-Warning Monitoring of Lithium-Ion Batteries. *Coord. Chem. Rev.* **2025**, *535*, 216624. https://doi.org/10.1016/j.ccr.2025.216624.
32. Han, C.; Wei, Y.; Wang, Z.; Wang, J.; Li, K.; Xiao, P.; Yang, Y. Research on Safety Warning Characteristics of Battery Based on Temperature and Force Coupling Signals. *J. Energy Storage* **2025**, *113*, 115602. https://doi.org/10.1016/j.est.2025.115602.
33. Kumar, S.; Kim, H.-J. Recent Advances in Early Warning Methods and Prediction of Thermal Runaway Events in Li-Ion Batteries. *J. Ind. Eng. Chem.* **2025**, *145*, 63–74. https://doi.org/10.1016/j.jiec.2024.10.057.
34. Zhi, M.; Liu, Q.; Xu, Q.; Pan, Z.; Sun, Q.; Su, B.; Zhao, H.; Cui, H.; He, Y. Review of Prevention and Mitigation Technologies for Thermal Runaway in Lithium-Ion Batteries. *Aerosp. Traffic Saf.* **2024**, *1*, 55–72. https://doi.org/10.1016/j.aets.2024.06.002.
35. Kostenko, G.; Zaporozhets, A. Transition from Electric Vehicles to Energy Storage: Review on Targeted Lithium-Ion Battery Diagnostics. *Energies* **2024**, *17*, 5132. https://doi.org/10.3390/en17205132.
36. Yao, Y.; Du, X.; Jiang, S.; Salehi, R.; Huemiller, E.; Shi, W. Prognostics of Internal Short Circuit in Lithium-Ion Batteries—Literature Survey and Performance Evaluation. *J. Power Sources* **2025**, *648*, 237203. https://doi.org/10.1016/j.jpowsour.2025.237203.
37. Vanem, E.; Wang, S. Data-Driven State of Health and State of Safety Estimation for Alternative Battery Chemistries—A Comparative Review Focusing on Sodium-Ion and LFP Lithium-Ion Batteries. *Future Batter.* **2025**, *5*, 100033. https://doi.org/10.1016/j.fub.2025.100033.
38. Jiang, J.; Cong, X.; Li, S.; Zhang, C.; Zhang, W.; Jiang, Y. A Hybrid Signal-Based Fault Diagnosis Method for Lithium-Ion Batteries in Electric Vehicles. *IEEE Access* **2021**, *9*, 19175–19186. https://doi.org/10.1109/ACCESS.2021.3052866.
39. Tian, L.; Dong, C.; Mu, Y.; Yu, X.; Jia, H. Online Lithium-Ion Battery Intelligent Perception for Thermal Fault Detection and Localization. *Heliyon* **2024**, *10*, e25298. https://doi.org/10.1016/j.heliyon.2024.e25298.
40. Yifan, Z.; Zhengjie, Z.; Sida, Z.; Xinan, Z.; Qiangwei, L.; Ning, C.; Shichun, Y. Innovative Fault Diagnosis and Early Warning Method Based on Multifeature Fusion Model for Electric Vehicles. *J. Energy Storage* **2024**, *78*, 109681. https://doi.org/10.1016/j.est.2023.109681.
41. Hong, J.; Wang, Z.; Yao, Y. Fault Prognosis of Battery System Based on Accurate Voltage Abnormity Prognosis Using Long Short-Term Memory Neural Networks. *Appl. Energy* **2019**, *251*, 113381. https://doi.org/10.1016/j.apenergy.2019.113381.
42. Li, H.; Liu, Z.; Kaleem, M.B.; Duan, L.; Ruan, S.; Liu, W. Fault Detection for Lithium-Ion Batteries of Electric Vehicles with Spatio-Temporal Autoencoder. *Appl. Energy* **2025**, *392*, 125933. https://doi.org/10.1016/j.apenergy.2025.125933.
43. Kostenko, G.; Bosak, A.; Sapsai, Y.; Zaporozhets, A. Analysis of Open Li-Ion Battery Testing Datasets for Degradation Modeling and Lifecycle Management. *Syst. Res. Energy* **2026**, *2026*, 65–84. https://doi.org/10.15407/srenergy2026.01.065.
44. Huang, P.; Zeng, G.; He, Y.; Liu, S.; Li, E.; Bai, Z. Damage Evolution Mechanism and Early Warning Using Long Short-Term Memory Networks for Battery Slight Overcharge Cycles. *Renew. Energy* **2023**, *217*, 119171. https://doi.org/10.1016/j.renene.2023.119171.
45. Wu, X.; Wei, Z.; Wen, T.; Du, J.; Sun, J.; Shtang, A.A. Research on Short-Circuit Fault-Diagnosis Strategy of Lithium-Ion Battery in an Energy-Storage System Based on Voltage Cosine Similarity. *J. Energy Storage* **2023**, *71*, 108012. https://doi.org/10.1016/j.est.2023.108012.
46. Zhang, Y.; Ping, P.; Dai, X.; Li, C.; Li, Z.; Zhuo, P.; Tang, L.; Kong, D.; Yin, X. Failure Mechanism and Thermal Runaway Behavior of Lithium-Ion Battery Induced by Arc Faults. *Renew. Sustain. Energy Rev.* **2025**, *207*, 114914. https://doi.org/10.1016/j.rser.2024.114914.
47. Zhao, Y.; Wei, M.; Dan, D.; Dong, J.; Wright, E. Enhancing Battery Electrochemical-Thermal Model Accuracy through a Hybrid Parameter Estimation Framework. *Energy Storage Mater.* **2024**, *72*, 103720. https://doi.org/10.1016/j.ensm.2024.103720.
48. Zhang, X.; Pan, Y.; Cao, Y.; Liu, B.; Yu, X. Cloud-Based Battery Failure Prediction and Early Warning Using Multi-Source Signals and Machine Learning. *J. Energy Storage* **2024**, *93*, 112004. https://doi.org/10.1016/j.est.2024.112004.
49. Li, J.; Che, Y.; Zhang, K.; Liu, H.; Zhuang, Y.; Liu, C.; Hu, X. Efficient Battery Fault Monitoring in Electric Vehicles: Advancing from Detection to Quantification. *Energy* **2024**, *313*, 134150. https://doi.org/10.1016/j.energy.2024.134150.
50. Guo, W.; Yang, L.; Deng, Z.; Xiao, B.; Bian, X. Early Diagnosis of Battery Faults Through an Unsupervised Health Scoring Method for Real-World Applications. *IEEE Trans. Transp. Electrif.* **2023**, *10*, 2521–2532. https://doi.org/10.1109/TTE.2023.3300302.
51. Zhang, X.; Yang, W.; Yan, L.; Kaleem, M.B.; Liu, W. Adaptive Internal Short-Circuit Fault Detection for Lithium-Ion Batteries of Electric Vehicles. *J. Energy Storage* **2024**, *84*, 110874. https://doi.org/10.1016/j.est.2024.110874.

52. Zhao, H.; Chen, Z.; Shu, X.; Xiao, R.; Shen, J.; Liu, Y.; Liu, Y. Online Surface Temperature Prediction and Abnormal Diagnosis of Lithium-Ion Batteries Based on Hybrid Neural Network and Fault Threshold Optimization. *Reliab. Eng. Syst. Saf.* **2024**, *243*, 109798. https://doi.org/10.1016/j.ress.2023.109798.
53. Li, M.; Dong, C.; Xiong, B.; Mu, Y.; Yu, X.; Xiao, Q.; Jia, H. STTEWS: A Sequential-Transformer Thermal Early Warning System for Lithium-Ion Battery Safety. *Appl. Energy* **2022**, *328*, 119965. https://doi.org/10.1016/j.apenergy.2022.119965.
54. Sun, J.; Ren, S.; Shang, Y.; Zhang, X.; Liu, Y.; Wang, D. A Novel Fault Prediction Method Based on Convolutional Neural Network and Long Short-Term Memory with Correlation Coefficient for Lithium-Ion Battery. *J. Energy Storage* **2023**, *62*, 106811. https://doi.org/10.1016/j.est.2023.106811.
55. Ouyang, M.; Zhang, M.; Feng, X.; Lu, L.; Li, J.; He, X.; Zheng, Y. Internal Short Circuit Detection for Battery Pack Using Equivalent Parameter and Consistency Method. *J. Power Sources* **2015**, *294*, 272–283. https://doi.org/10.1016/j.jpowsour.2015.06.087.
56. Zhao, Y.; Deng, J.; Liu, P.; Zhang, L.; Cui, D.; Wang, Q.; Sun, Z.; Wang, Z. Enhancing Battery Durable Operation: Multi-Fault Diagnosis and Safety Evaluation in Series-Connected Lithium-Ion Battery Systems. *Appl. Energy* **2025**, *377*, 124632. https://doi.org/10.1016/j.apenergy.2024.124632.
57. Hua, N.; Chen, H.; Dong, G. Deep Transfer Learning-Enabled Battery Health Prognosis Using Impedance Spectrum Data. *J. Energy Storage* **2025**, *132*, 117855. https://doi.org/10.1016/j.est.2025.117855.
58. Jiang, L.; Deng, Z.; Tang, X.; Hu, L.; Lin, X.; Hu, X. Data-Driven Fault Diagnosis and Thermal Runaway Warning for Battery Packs Using Real-World Vehicle Data. *Energy* **2021**, *234*, 121266. https://doi.org/10.1016/j.energy.2021.121266.
59. Severson, K.A.; Attia, P.M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M.H.; Aykol, M.; Herring, P.K.; Fraggedakis, D.; et al. Data-Driven Prediction of Battery Cycle Life before Capacity Degradation. *Nat. Energy* **2019**, *4*, 383–391. https://doi.org/10.1038/s41560-019-0356-8.
60. Zhang, Y.; Xiong, R.; He, H.; Pecht, M.G. Long Short-Term Memory Recurrent Neural Network for Remaining Useful Life Prediction of Lithium-Ion Batteries. *IEEE Trans. Veh. Technol.* **2018**, *67*, 5695–5705. https://doi.org/10.1109/TVT.2018.2805189.
61. Wang, Q.; Mao, B.; Stoliarov, S.I.; Sun, J. A Review of Lithium Ion Battery Failure Mechanisms and Fire Prevention Strategies. *Prog. Energy Combust. Sci.* **2019**, *73*, 95–131. https://doi.org/10.1016/j.pecs.2019.03.002.
62. Shu, X.; Yang, H.; Liu, X.; Feng, R.; Shen, J.; Hu, Y.; Chen, Z.; Tang, A. State of Health Estimation for Lithium-Ion Batteries Based on Voltage Segment and Transformer. *J. Energy Storage* **2025**, *108*, 115200. https://doi.org/10.1016/j.est.2024.115200.
63. Lin, M.; Xiao, P.; Meng, J.; Wang, W.; Wu, J. State of Health Estimation for Battery Packs Based on Ensemble Bayesian Neural Networks Integrating Uncertainty Quantification and Calibration. *IEEE Trans. Transp. Electrif.* **2026**, *Epub ahead of printing*. https://doi.org/10.1109/TTE.2026.3683737.
64. Sun, J.; Qiu, Y.; Shang, Y.; Lu, G. A Multi-Fault Advanced Diagnosis Method Based on Sparse Data Observers for Lithium-Ion Batteries. *J. Energy Storage* **2022**, *50*, 104694. https://doi.org/10.1016/j.est.2022.104694.
65. Guo, X.; Song, Y.; Lyu, N.; Jin, Y. A Novel Passive Wireless Safety Early Warning Technique Based on Transient Pressurization Relief Energy Harvesting of Battery Safety Valves. *J. Energy Storage* **2025**, *129*, 117394. https://doi.org/10.1016/j.est.2025.117394.
66. Nie, B.; Dong, Y.; Chang, L. The Evolution of Thermal Runaway Parameters of Lithium-Ion Batteries under Different Abuse Conditions: A Review. *J. Energy Storage* **2024**, *96*, 112624. https://doi.org/10.1016/j.est.2024.112624.
67. Zhang, J.; Wang, Y.; Jiang, B.; He, H.; Huang, S.; Wang, C.; Zhang, Y.; Han, X.; Guo, D.; He, G.; et al. Realistic Fault Detection of Li-Ion Battery via Dynamical Deep Learning. *Nat. Commun.* **2023**, *14*, 5940. https://doi.org/10.1038/s41467-023-41226-5.
68. Bonomi, M.; Barducci, A.; Parrinello, M. Reconstructing the Equilibrium Boltzmann Distribution from Well-Tempered Metadynamics. *J. Comput. Chem.* **2009**, *30*, 1615–1621. https://doi.org/10.1002/jcc.21305.