SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V13I10P107

Volume 13 Issue 10, 83-105, October 2026
© 2026 Seventh Sense Research Group®

Original Article

An Integrated Explainable Machine Learning Framework for Cell-to-Pack Imbalance Prediction in Electric Vehicle Battery Systems

Neeraja T[1], Kalpalatha Reddy T[2]

[1]*Department of Electrical and Electronics Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, Tamil Nadu, India.*
[2]*Department of Electronics and Communication Engineering, Dr. M.G.R. Educational and Research Institute, Chennai, Tamil Nadu, India.*

[1]*Corresponding Author: neeraja722@gmail.com*

Received: 03 July 2026 Revised: 15 September 2026 Accepted: 24 September 2026 Published: 28 October 2026

Abstract *- This study presents a machine learning framework for predicting cell-to-pack imbalance in electric vehicle batteries. A set of battery attributes that affect battery performance, including State of Charge (SOC), State of Health (SOH), temperature (including ambient temperature and humidity), and cabin temperature, were used to make this prediction. The collected battery data were pre-processed to improve prediction performance. Three preprocessing methods were used: (i) Variational Mode Decomposition (VMD); (ii) deep denoising autoencoders, and (iii) feature selection using minimum Redundancy Maximum Relevance (mRMR) mutual information. This preprocessing step removes noise from the dataset and identifies useful information for prediction. Cell-to-pack imbalance prediction was performed using three interpretable machine learning models: (i) an Explainable Boosting Machine (EBM); (ii) fuzzy regression; and (iii) RuleFit. These models provide transparent predictions and help in understanding the factors responsible for battery imbalance. Models were evaluated on an independent 20-sample test set held out from the 100-record dataset. Among the three interpretable models, Fuzzy Regression achieved the strongest predictive performance (R^2 = 0.998), followed by RuleFit (R^2 = 0.976), while EBM achieved a lower R^2 of 0.201 on this independent test set despite providing detailed feature-level explanations. These results indicate that combining signal preprocessing, feature selection, and interpretable regression can provide useful and explainable information for cell-to-pack imbalance prediction, while also showing that performance can vary substantially between interpretable models on a small, two-pack dataset; validation on larger and more varied battery datasets is needed to confirm generalizability.*

Keywords *- Battery Management System, Cell-to-Pack balancing, Deep Denoising Autoencoder, Electric Vehicles, Explainable Machine Learning, Feature selection, Interpretable models, Variational Mode Decomposition.*

1. Introduction

Battery Management Systems (BMS) are important for the safe and reliable operation of Electric Vehicle (EV) battery packs. They monitor the battery condition and help prevent problems such as overcharging, over-discharging, overheating, and short circuits [1, 2]. Proper battery management helps to improve battery performance and service life [3, 4]. As EV battery packs become larger, monitoring the condition of individual cells becomes more important [5]. A battery pack contains several cells, and all cells may not behave in the same way. Differences in State of Charge (SOC), State of Health (SOH), temperature, and other operating conditions can cause cell imbalance [6]. This imbalance can reduce the useful performance of the battery pack. Temperature differences can also affect cell aging and internal resistance. Therefore, monitoring important battery parameters is necessary to identify changes in cell conditions and predict possible imbalances. Machine learning methods are now used in battery management for state estimation and prediction. Deep learning methods can provide good prediction results under different operating conditions [7]. Explainable machine learning methods can help to understand the reasons behind model predictions [8].

However, many existing studies mainly focus on SOC estimation, SOH estimation, battery degradation, or cell balancing. Fewer studies focus on predicting cell-to-pack imbalance and explaining the factors that affect the prediction. Battery measurements may also contain noise and unnecessary information. These problems can affect the prediction results. Therefore, the battery data needs to be properly processed before applying a prediction model. Signal

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

processing and feature selection can help to obtain useful information from the data. An interpretable machine learning model can then be used to predict the imbalance and show the important battery parameters.

1.1. Problem Statement

Cell-to-pack imbalance occurs when the cells in a battery pack have different operating or aging conditions. This imbalance can reduce battery efficiency and usable capacity. It can also affect battery life. Therefore, early prediction of cell-to-pack imbalance is important for battery management. However, battery data can contain noise and unnecessary features. The relationship between battery parameters and cell imbalance can also be complex. These issues make accurate prediction difficult. A prediction model should also provide information about the parameters that affect the imbalance. Therefore, there is a need for a method that can process battery data, select useful features, and provide clear imbalance predictions.

1.2. Research Gap

Previous studies on battery management have focused on state estimation, battery degradation, signal processing, and cell balancing [9, 10]. These studies have improved battery monitoring and prediction. However, limited studies have combined data denoising, feature selection, and interpretable machine learning for cell-to-pack imbalance prediction. Many machine learning models can provide good prediction results. However, some models do not mention which battery parameters affect the prediction. This makes it difficult to understand the reasons for cell imbalance. It also limits the practical use of the prediction results in battery management.

Another issue is the quality of battery data. Noise and repeated information can affect the performance of the prediction model. Therefore, the data should be cleaned, and the important features should be selected before prediction. Based on these issues, this study develops a framework that combines signal preprocessing, feature selection, and interpretable machine learning for cell-to-pack imbalance prediction. VMD, DDAE, and mRMR are used to process and select the battery data. EBM, Fuzzy Regression, and RuleFit are then used to predict the cell-to-pack imbalance.

Novelty of the Present Study

The novelty of this study is the use of signal preprocessing and interpretable machine learning together for cell-to-pack imbalance prediction. Previous studies have focused on SOC, SOH, battery life, or capacity prediction [11, 12, 15-21]. Other studies have focused on cell balancing, aging estimation, pack reconfiguration, or charging control [22-30]. These studies have addressed important battery management problems. However, they do not combine data denoising, feature selection, and interpretable prediction for cell-to-pack imbalance. In this study, VMD is first used to process the battery signals. DDAE is then used to reduce noise and improve the data quality. After this, mRMR with MI is used to select the important features. The processed data are used with EBM, Fuzzy Regression, and RuleFit models for imbalance prediction. These models also provide information about the factors related to the prediction. Thus, the proposed work differs from the existing studies by combining data preprocessing, feature selection, and interpretable machine learning in a single framework for cell-to-pack imbalance prediction.

1.3. Contributions

The main contributions of this study are:

1. A battery monitoring framework is developed using SOC, SOH, temperature, humidity, and cabin temperature.
2. VMD is used to reduce noise and extract useful information from the battery signals.
3. DDAE is used to improve the quality of the battery data.
4. mRMR with MI is used to select the important battery features and remove unnecessary information.
5. EBM, Fuzzy Regression, and RuleFit are used to predict the cell-to-pack imbalance.
6. The interpretable models help to identify the battery parameters that are related to the predicted imbalance.

2. Literature Survey

Table 1 presents previous studies related to battery management. Battery management has been studied using different methods. Some studies focus on estimating SOC and SOH to monitor the battery condition [11, 12, 21]. Other studies use signal processing and machine learning to improve battery state and degradation prediction [15-20]. Cell balancing is an important area where active and passive balancing methods are used to reduce differences between cells [22]. Other researchers have studied battery aging, pack reconfiguration, and charging control to improve battery performance [23-30]. These studies show that different methods can be used to improve battery monitoring and management.

Recent studies have investigated cell-to-cell variation, machine learning, and explainable methods for battery management. CNN-LSTM models have been used to estimate cell-to-cell inconsistency from battery operating data [34]. Other studies have examined electrical variations between cells and their effect on battery behaviour [35]. Explainable machine learning has also been applied to battery state and performance estimation [36]. Probabilistic machine learning and artificial intelligence-based cell balancing methods have been studied for battery health and management [37, 38]. However, these studies focus on cell inconsistency estimation, battery health, diagnosis, or balancing. They do not combine signal denoising, feature selection, and interpretable machine learning for cell-to-pack imbalance prediction. Therefore, this study combines VMD, DDAE, mRMR, and interpretable prediction models to address this gap.

Table 1. Literature survey

Reference	Area	Methods	Drawbacks
[11]	ML-based SOH estimation	Developed ML framework for SOH using battery cycle data, tracks degradation via capacity fade or resistance increase; critical for lifespan prediction.	High model complexity; computationally intensive; limited real-time applicability
[12]	Remora optimization + Dual adaptive Kalman filtering	Collaborative evaluation of SOC, SOP, SOH in EV batteries	Algorithmic complexity; integration challenges for onboard systems
[13]	Cabin comfort control (temp/humidity)	Control strategies for the EV cabin environment	Indirect SOH relevance; neglects direct battery-environment interactions
[14]	Data optimization for parameter estimation	Framework for parameter estimation under uncertainty	Assumes idealized uncertainties; limited real-world applicability; needs precise data
[15]	Wavelet + Bayesian-optimized GPR for SOC prediction	Advanced SOC prediction using hybrid ML techniques	Highly data-dependent; sensitive to noise; generalization to field data is limited
[16]	FPCA for capacity prediction	Extracts features from monitoring data for improved capacity prediction	Needs high-quality data; less effective for nonlinear behaviours; limited interpretability
[17]	EMD + Neural Network for SOH	Combines signal processing and neural networks for SOH estimation	Computationally intensive; sensitive to noise; risk of overfitting; lacks transparency
[18]	Hybrid SVR model for life prediction	Hybrid ML model for robust battery life prediction	Increased complexity; parameter sensitivity; generalizability concerns
[19]	Random Forest for manufacturing feature analysis	Identifies key manufacturing factors affecting battery quality	Focused on manufacturing, not operation; interpretability issues; resource-intensive
[20]	LSTM for RUL with partial data	Time-series RUL prediction with incomplete data	Needs large datasets; overfitting risk; black-box model limits interpretability
[21]	GPR for SOC estimation	Probabilistic, data-driven SOC estimation with uncertainty quantification	Computationally demanding for onboard use; assumes stationary noise statistics, limiting accuracy under rapidly varying operating conditions
[22]	Review of cell balancing techniques	Comparative analysis of active vs. passive balancing	Largely qualitative; lacks quantitative/experimental validation; limited practical guidance
[23]	Aging estimation + active balancing	Integrates aging estimation with dynamic balancing for life extension	Increased system complexity; real-time challenges; needs frequent recalibration
[24]	AI-assisted pack reconfiguration	Dynamic reconfiguration for improved balancing and runtime	Dependent on AI/data quality, hardware complexity, reliability/safety concerns
[25]	Model Predictive Control (MPC) for optimal charging	Optimized battery charging strategy to reduce degradation and improve battery performance.	Requires accurate battery models, high computational demand, and implementation complexity.
[34]	Cell-to-cell inconsistency	CNN-LSTM	Complex model and limited interpretability
[35]	Cell-to-cell variation	Statistical analysis and battery modelling	Focuses mainly on characterization and modelling
[36]	Explainable ML for Li-ion batteries	Review of XAI/ML methods	Review study; does not develop an imbalance prediction framework

[37]	Battery diagnostics and prognostics	Probabilistic ML	Mainly focused on health diagnostics and uncertainty
[38]	AI and cell balancing	AI techniques and balancing methods	Broad review; does not combine preprocessing and interpretable imbalance prediction.
Proposed Method	Collection of multi-dimensional battery parameters, Advanced signal processing and feature selection/extraction, Explainable and interpretable machine learning for battery state prediction	Provides comprehensive input data (SOC, SOH, temperature, humidity, cabin temperature) crucial for accurate battery state assessment and prediction.	Evaluated on a limited two-pack dataset (100 samples); validation on larger, real-world battery packs is needed before deployment.

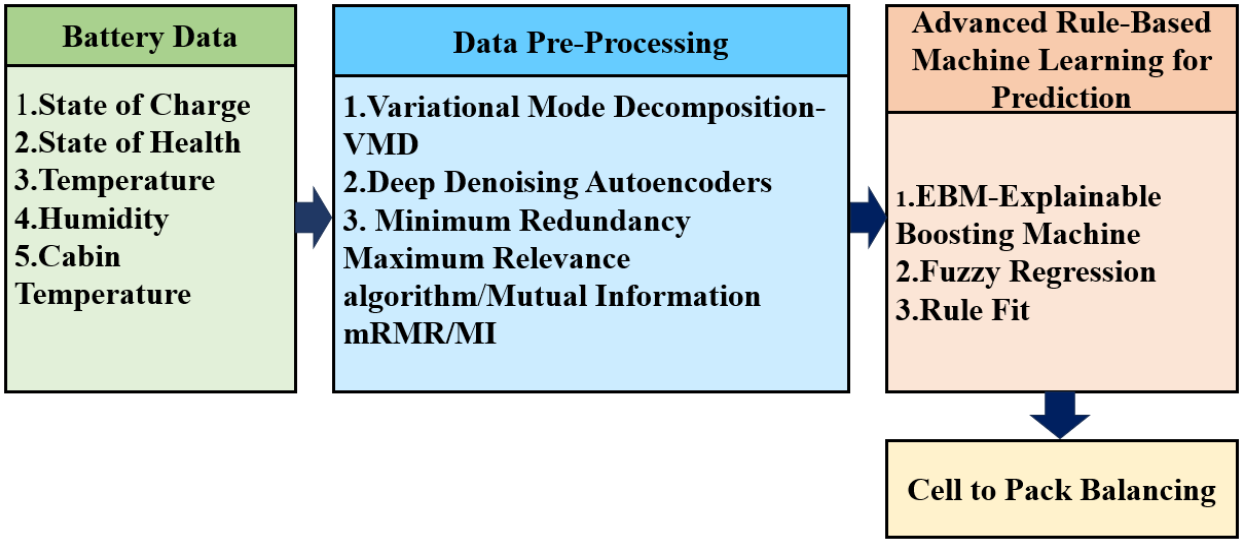


Fig. 1 Block diagram of the proposed VMD-DDAE-mRMR/MI framework for cell-to-pack imbalance prediction, showing the battery data, data preprocessing, and interpretable machine learning stages

3. Methodology

The proposed method to predict cell-to-pack balance in EV batteries is given in Figure 1. The proposed approach for cell-to-pack balancing includes three main steps: data collection, data preprocessing, and prediction using rule-based machine learning. It uses battery parameters, processes the data, and then applies simple and interpretable models to predict and improve cell balancing. The overall process is shown in the block diagram and explained below.

3.1. Battery Data Collection

Cell-to-pack imbalance happens when the cells in a battery pack do not operate in the same way. Increased inefficiency, possible safety hazards, and shortened life of the battery can result from an imbalance in the current being supplied to or drawn from the battery. The dataset contains battery-related variables such as SOC, SOH, and battery temperature, together with environmental variables such as humidity and cabin temperature. The environmental variables are included as auxiliary variables because they may reflect operating and thermal-management conditions. SOC and SOH provide valuable insights into the degree of imbalance between the cells. The SOC represents the available charge in a battery cell, and SOH indicates the overall condition and aging status of the cell. Differences in manufacturing conditions and operating environments can cause variations among battery cells, leading to an imbalance. As a result, if the SOC of one or more cells is not equal to that of the other cells, the cells that have a higher SOC will likely be overcharged, while the cells that have a lower SOC will likely be over-discharged. The battery management system uses various parameters to monitor individual cell conditions and predict any cell imbalances. This information is then collected on the data-processing side in combination with machine learning, which leads to improved balancing and overall battery performance due to the improved data quality. Two major factors affecting the performance of a battery are temperature and humidity; both of these factors play a role in a battery's cell performance and can contribute to cell imbalances in SOC and SOH, as discussed. Temperature, for

example, plays an integral role in how fast chemical reactions take place within the battery. Typical operating temperatures for a battery are 15°C to 35°C; as the battery temperature increases, the battery and its cells will degrade faster, and the internal resistance of the battery will increase as well.

If the temperature of the cells in a battery is not equal throughout, then it is possible for some of the cells to age faster than others, which will create an imbalance between the SOC and SOH of the individual cells. The effect of humidity on battery performance is significant because when there is a high level of humidity, moisture in the air can corrode the electrodes or cause unanticipated chemical reactions that will cause the cells to age unevenly. This will lead to an increase in imbalance in the overall battery pack as well. It should be noted that the direct physical drivers of cell-to-cell imbalance are cell-level quantities: differences in delta voltage, internal resistance, and local thermal gradients between cells. Cabin temperature does not act on individual cells directly; it is included as an auxiliary, pack-external variable that reflects the ambient thermal-management environment surrounding the pack, which in turn influences the cooling load and thermal boundary conditions experienced by the outer cells, consistent with its role in cabin thermal comfort and management studies [13]. Cabin temperature is therefore treated as an indirect, contextual proxy rather than a primary physical cause of imbalance, and its feature importance (Section 3.2.3, Figure 4) is interpreted accordingly; a limitation of the present dataset is that it does not include direct cell-level measurements of delta voltage or internal resistance, which is identified as a direction for future work (Section 5). By monitoring all of these parameters, the BMS is able to accumulate historical data so that it can identify relationships between battery imbalance and the measured operating and environmental variables.

Using these parameters for imbalance prediction can provide useful information for battery monitoring and support BMS-based decision-making. The three methods for improving the data collection use VMD on battery parameters, including SOC, SOH and temperature, since these battery parameters exhibit non-linear, non-stationary behaviour for the purpose of predicting cell-to-pack imbalances and extracting relevant signal components of noisy and complex collected battery data.

3.2. Data Pre-Processing

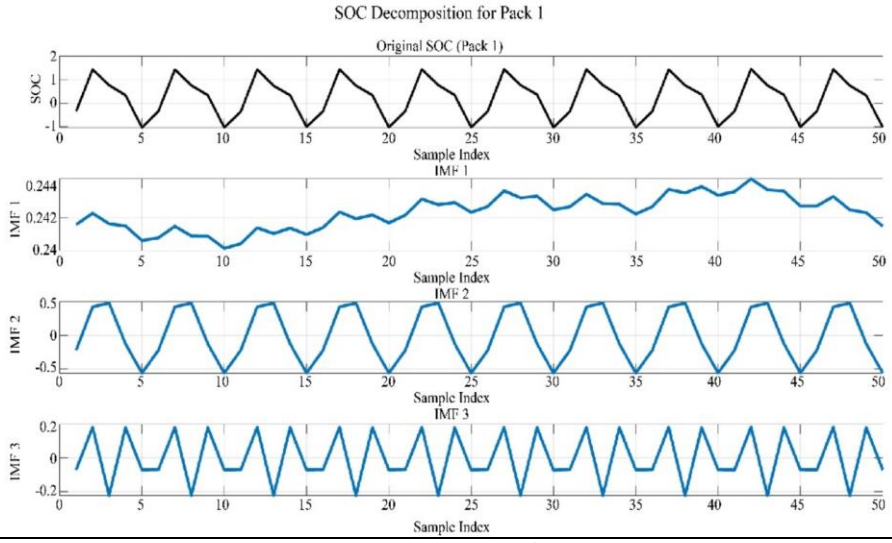
There are several steps used to prepare the raw battery data collected from machines using machine-learning algorithms. The three pre-processing techniques are used to enhance data quality, reduce background noise, and select appropriate features.

3.2.1. Variational Mode Decomposition (VMD)

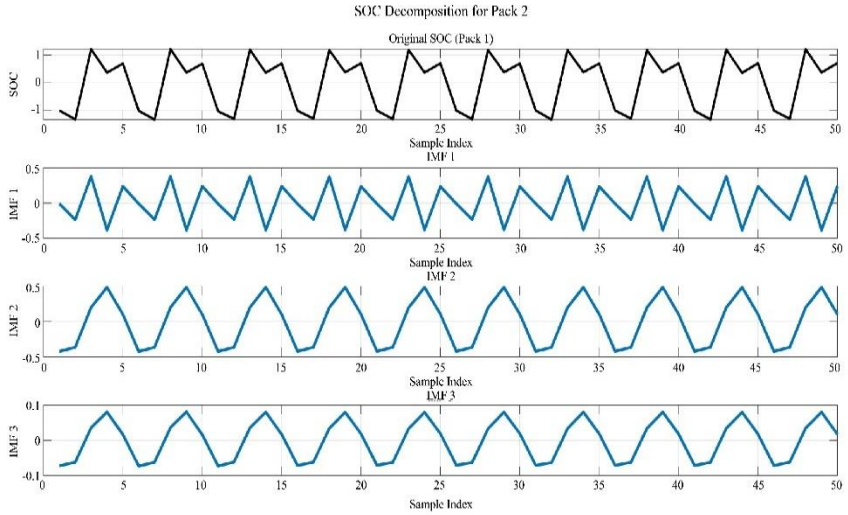
VMD is used to improve the quality of the collected data, including SOC, SOH and temperature, since these battery parameters exhibit non-linear, non-stationary behaviour for the purpose of predicting cell-to-pack imbalances and extracting relevant signal components of noisy and complex collected battery data. The decomposition of the above faulty sensor inputs and actual battery measurements will help explain how effective the VMD process is in reducing noise by separating high-frequency noise from the usable input component in order to generate more accurate predictions of the expected balancing of a cell within a battery pack. The examples of low-frequency trends associated with long-term evolution of health and degradation of the cell and high-frequency oscillatory trends associated with voltage spikes and thermal anomalies make VMD highly advantageous in providing the BMS with timely indications of the performance and degradation of cells due to temperature or humidity effects, rather than do more conventional separation techniques of instrumented measurements. The results of this comparison of VMD to other decomposition methodologies can be seen in Table 2. Table 2 compares VMD with other decomposition techniques used for processing battery signals. VMD achieves the lowest MSE (0.061) and highest Accuracy (94.2%), demonstrating superior denoising and feature extraction, though with a higher computation time (11.4 seconds). Figure 2 shows how battery data is pre-processed using VMD.

Table 2. Comparison of different decomposition strategies used for pre-processing battery data

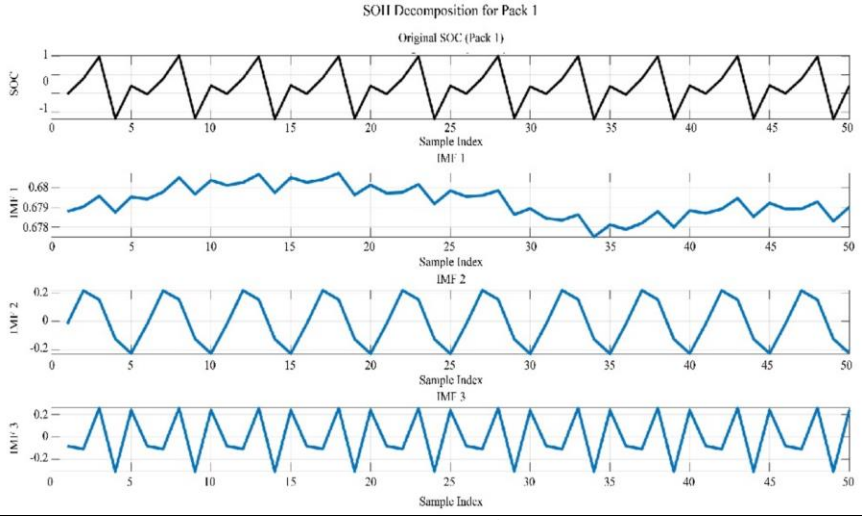
Decomposition Technique	MSE	R^2 Score	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Computation Time (Sec)
Variational Mode Decomposition (VMD)	**0.061**	**0.91**	**94.2**	**93.0**	**92.7**	**92.9**	11.4
Empirical Mode Decomposition (EMD)	0.074	0.86	90.5	88.9	89.4	89.1	10.3
Empirical Wavelet Transform (EWT)	0.069	0.88	91.8	90.3	89.9	90.1	9.7
Discrete Wavelet Transform (DWT)	0.078	0.83	89.6	87.4	88.2	87.8	8.9
Fast Fourier Transform (FFT)	0.090	0.79	85.4	84.1	83.2	83.6	7.5



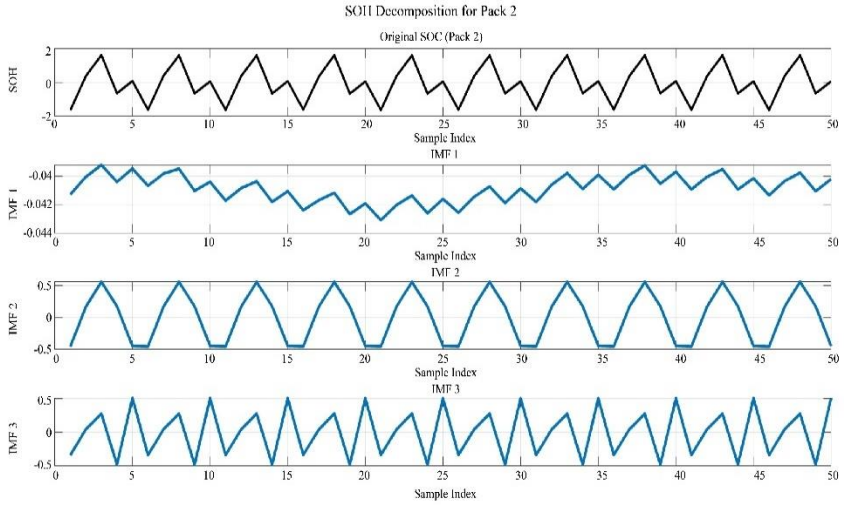
(a) SOC decomposition of pack 1



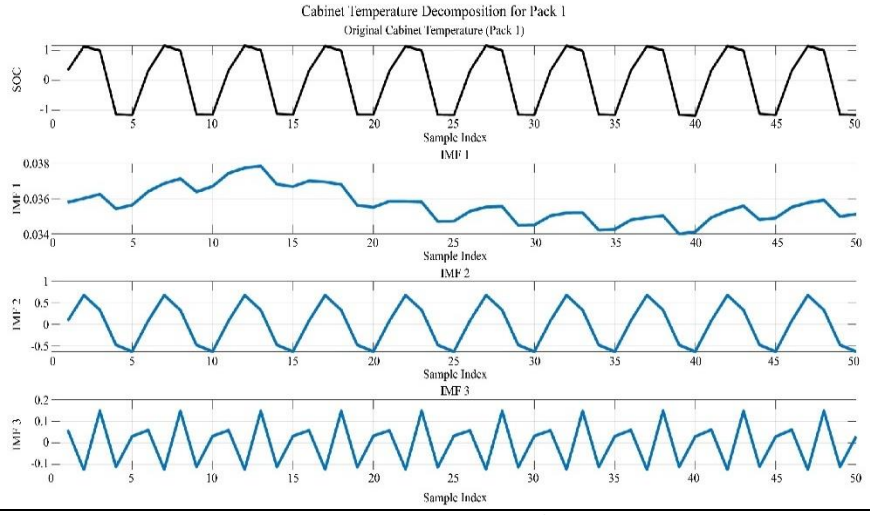
(b) SOC decomposition of pack 2



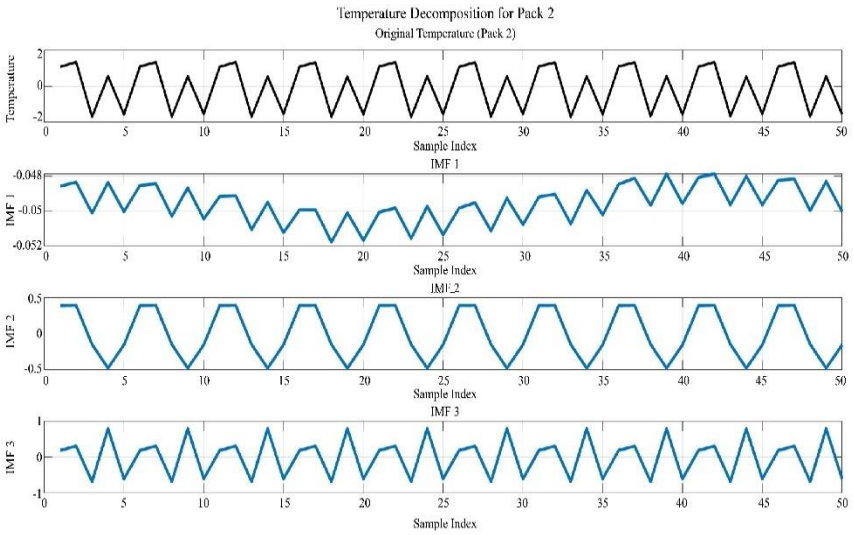
(c) SOH decomposition for pack 1



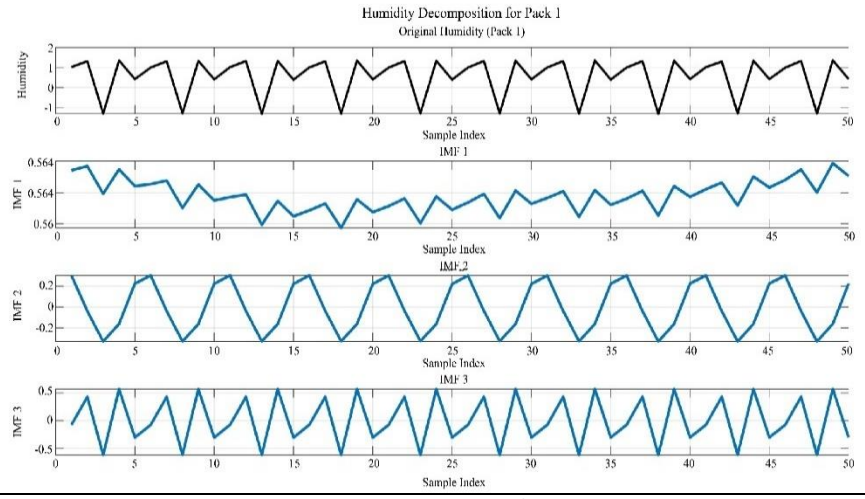
(d) SOH decomposition for pack 2



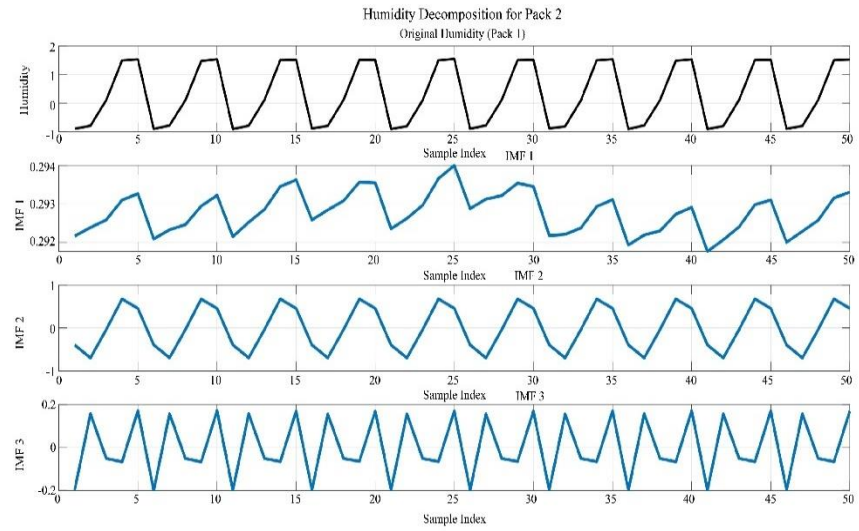
(e) Temperature decomposition for pack 1



(f) Temperature decomposition for pack 2



(g) Humidity decomposition for pack 1



(h) Humidity decomposition for pack 2

Fig. 2 Variational Mode Decomposition (VMD) of battery parameters SOC, SOH, temperature, and humidity for battery packs 1 and 2

- The original signals for SOC, SOH, and temperature are shown along with their decomposed modes (IMFs).
- Each parameter is divided into multiple modes, where each mode represents a different pattern in the data.
- The first mode shows slow changes, while the other modes capture repeating and faster variations.
- This separation helps to reduce noise and makes hidden patterns easier to understand.
- It supports better prediction of the imbalance between cells and the battery pack.

3.2.2. Deep Denoising Autoencoders (DDAEs)

Battery measurements contain noise and fluctuations caused by sensors and environmental conditions. Deep Denoising Autoencoders (DDAEs) are used to remove this noise while preserving the information. For a given noisy input.

$$\tilde{x} = x + noise \quad (1)$$

The DDAE reconstructs a denoised output

$$\tilde{x} \approx x$$

Where x is the true signal. DDAEs process multivariate time series data from battery parameters, learning the compressed features and capturing interdependencies and nonlinear degradation patterns. The encoder maps the noisy input to a compressed feature vector

$$h = f_\theta(\tilde{x}) \quad (2)$$

and the decoder reconstructs

$$\hat{x} = g_\emptyset(h) \quad (3)$$

f_θ The encoder network, θ, represents its parameters. $g_\emptyset$ The decoder network, h, is the compressed, denoised feature vector. DDAE is trained to minimize the reconstruction loss

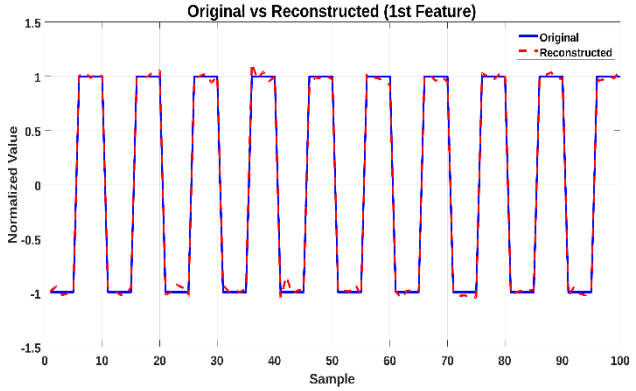
$$\mathcal{L}_{DDAE} = \mathbb{E}_{x\sim q_{D(\tilde{x}/x)}[\|\widetilde{x-g_\emptyset(f_\theta(\widetilde{x}))}\|]^2]} \quad (4)$$

$$\mathcal{L} = \|x - \tilde{x}\|^2 \quad (5)$$

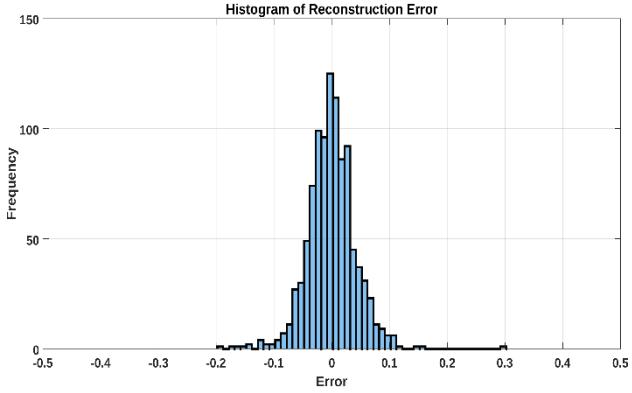
The training process helps the model preserve useful information while removing unwanted noise.

The denoised, compressed features extracted by the DDAE are used as inputs to machine learning models for predicting cell-to-pack imbalances.

Also, DDAEs learn normal operating patterns during training. Large reconstruction errors during inference indicate anomalies, like cell drift, thermal runaway, or SOH imbalances, enabling early fault detection before issues reach the pack level. Figure 3 presents the DDAE-based feature reconstruction for battery data, showing the effectiveness of denoising across SOC, SOH, temperature, and humidity.



(a) Comparison of original and reconstructed battery signals using the deep denoising autoencoder



(b) Histogram of reconstruction error across all samples

Parameter	Value / Setting
Network type	Fully-connected Deep Denoising Autoencoder (DDAE)
Architecture	Input (15) → FC-64 → FC-32 (bottleneck) → FC-64 → Output (15)
Encoder layers	FC(64) + ReLU → FC(32) + ReLU
Decoder layers	FC(64) + ReLU → FC(15, linear output)
Input features	15 (5 base variables × 3 VMD IMFs: SOC, SOH, Temperature, Cabinet Temperature, Humidity)
Noise injection	Additive Gaussian noise, σ = 0.1, applied to normalized (z-scored) inputs
Loss function	Mean Squared Error (regression layer)
Optimizer	Adam
Initial learning rate	0.001 (constant schedule)
Mini-batch size	32
Maximum epochs	300
Shuffling	Every epoch
Random seed	1 (fixed, for reproducibility)

(c) DDAE architecture and training hyperparameters used for battery-signal denoising

Fig. 3 Deep Denoising Autoencoder performance for battery data preprocessing

DDAE architecture and training hyperparameters used for battery-signal denoising.

Figure 3 shows (a) the plot comparing the original battery signal features like SOC, SOH, Temperature, or Humidity, with its reconstructed versions after passing through the DDA. (b) The histogram presents the reconstruction error difference between the original and reconstructed signals across all samples. (c) reports the DDAE architecture and training hyperparameters (Adam optimizer, 300 epochs, mini-batch size 32, learning rate 0.001, additive Gaussian noise with σ = 0.1) used to train the network for reliable noise removal from the battery signals used in the proposed method. Table 3 shows the performance comparison of denoising methods for Battery Data.

Table 3. Performance comparison of denoising techniques for battery data preprocessing

Denoising Technique	MSE	R^2 Score	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Computation Time (Sec)
Deep Denoising Autoencoders (DDAE)	0.066	0.89	93.5	91.6	92.2	91.9	10.5
Traditional Autoencoders	0.079	0.84	89.7	87.3	86.8	87.0	9.2
Principal Component Analysis (PCA)	0.087	0.80	87.2	85.1	83.9	84.5	8.4
Wavelet Denoising	0.073	0.86	90.5	88.7	89.1	88.9	9.8

The performance comparison of denoising methods for battery data is shown in Table 3. DDAE performs better than the other methods in most cases, particularly when dealing with complex patterns and noisy battery data.

3.2.3. Minimum Redundancy Maximum Relevance (mRMR) with Mutual Information (MI)

Battery data, including State of Charge (SOC), State of Health (SOH), internal temperature, cabin humidity, voltages, and imbalance indices, create a high-dimensional dataset with irrelevant features. These features can increase model complexity and overfitting, reducing the accuracy of cell-to-pack imbalance predictions. Minimum Redundancy Maximum Relevance (mRMR) with Mutual Information (MI) is used to select a minimal, informative feature subset by maximizing relevance to the imbalance and minimizing redundancy among features.

Before feature selection, the prediction target used throughout Sections 3.2-4 is formally defined. The Cell Imbalance Index (Y) is a continuous score that quantifies the deviation of an individual cell's state from the pack average, combining the normalized deviations in SOC, SOH, and temperature:

$$Y_{CII} = W_1 \left|\frac{SOC_i - \overline{SOC}}{\sigma_{SOC}}\right| + w_2 \left|\frac{SOH_i - \overline{SOH}}{\sigma_{SOH}}\right| + W_3 \left|\frac{T_i - \bar{T}}{\sigma_T}\right| \quad (6)$$

Where SOC_i, SOH_i, and T_i are the SOC, SOH, and temperature of cell i) the overbar and σ terms denote the pack-level mean and standard deviation of each parameter; and w1, w2, w3 are non-negative weights (set equal by default, w1 = w2 = w3 = 1/3) that can be tuned to reflect the relative importance of each parameter for a given battery chemistry. A higher Y indicates a cell that deviates more strongly from the pack average and is therefore more likely to be imbalanced. This continuous score is the regression target reported throughout Sections 3 and 4 (MAE, MSE, RMSE, R^2); for the classification metrics reported alongside it (Accuracy, Precision, Recall, F1), Y is additionally thresholded at the 90th percentile of its pack-level distribution to label a cell as "imbalanced" versus "balanced," consistent with common practice for deriving a binary alarm from a continuous health/imbalance score.

The mRMR algorithm optimizes the objective function D, the average relevance of the selected feature subset S to the target Y- imbalance, minus R, the average redundancy among features in S. Relevance D is computed as:

$$D = \frac{1}{|S|}\sum_{x_{i\epsilon S}} I(x_i ; Y) \quad (7)$$

Where $I(x_i, Y)$ is the mutual information between the feature x_i Moreover, target Y. Redundancy is given by:

$$R = \frac{1}{|S|^2}\sum_{x_i, x_{j\epsilon S}} I(x_i ; x_j) \quad (8)$$

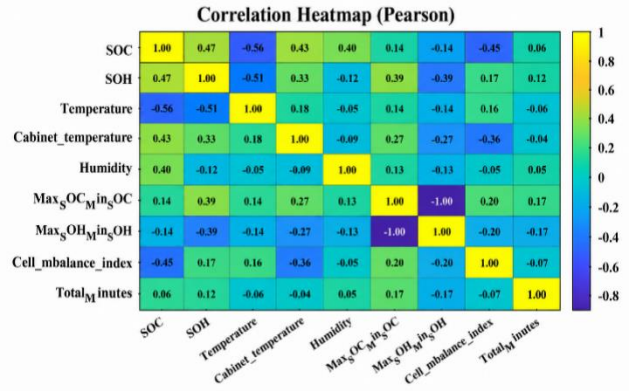
Mutual Information I(X, Y) measures the dependency between a feature and the target variable.

$$I(X,Y) = \sum_{x\epsilon X}\sum_{y\epsilon Y} P(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (9)$$

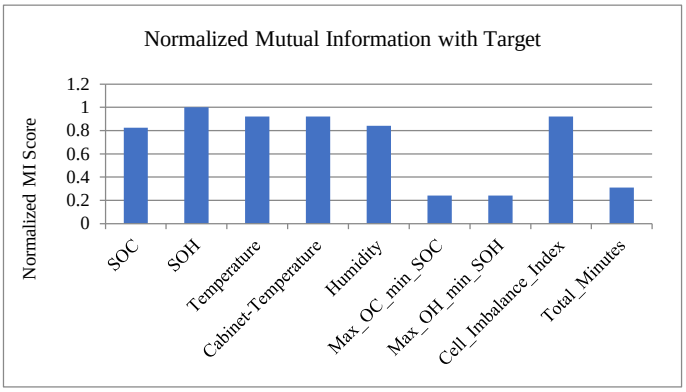
A higher mutual information value indicates that a feature is more useful for prediction. Preprocessing with mRMR results in a minimal, informative set of battery features used for training the prediction.

Figure 4 shows the feature importance and correlation analysis of battery data using mRMR/MI.

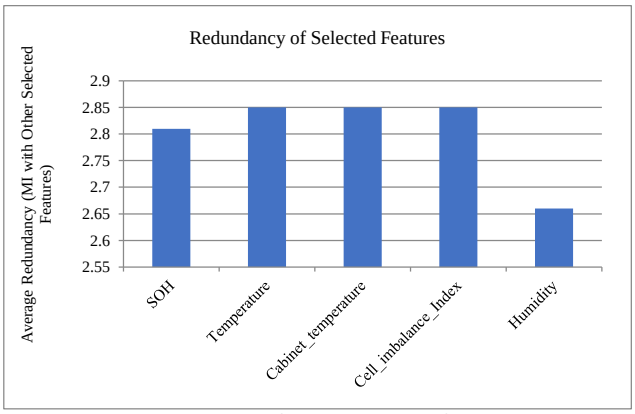
Figure 4(a) visualizes the Pearson correlation coefficients between various sensor measurements and battery health indicators. High correlation between Max SOC and Min SOC, indicating consistent charge behaviours. (b) The bar chart displays the normalized importance of each feature based on a Random Forest regression model. (c) The plot shows the frequency distribution of selected top features based on an embedded feature selection method. SOC, SOH, and humidity are the most important features in the final model. (d) The subplot presents the SHapley Additive exPlanations (SHAP) values, which quantify the contribution of each feature to the output. Table 4 presents the comparative performance of feature selection techniques for Battery Imbalance Prediction.



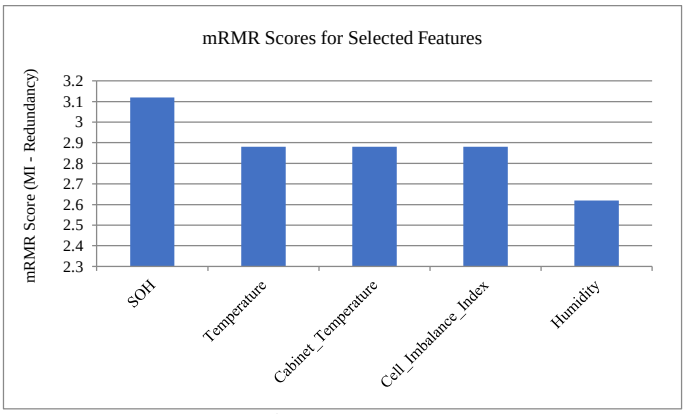
(a) Correlation heatmap of mRMR-selected features



(b) Random Forest feature importance of mRMR-selected features



(c) Distribution of mRMR-selected features



(d) SHAP feature importance values

Fig. 4 Feature importance and correlation analysis of battery data using mRMR with MI

Table 4. Comparative performance of feature selection techniques for battery health prediction

Feature Selection Method	MSE	R^2 Score	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Computation Time (Sec)
mRMR (Minimum Redundancy Maximum Relevance)	**0.063**	**0.90**	**93.8**	**91.9**	**92.4**	**92.1**	**9.5**
Mutual Information (MI)	**0.065**	**0.89**	**93.2**	**91.3**	**91.8**	**91.5**	**9.0**
Recursive Feature Elimination (RFE)	0.071	0.86	91.5	89.4	89.7	89.5	8.2
L1-Regularization (Lasso)	0.076	0.84	89.6	87.2	88.0	87.6	7.8
PCA for Dimensionality Reduction	0.081	0.81	87.4	85.0	84.2	84.6	6.5

Table 5. Hyperparameter ranges for explainable boosting machine in battery imbalance prediction

Parameter	Typical Value/Range	Description
Number of Estimators	100-500	Captures subtle, nonlinear imbalance patterns such as SOC–SOH–temperature drift using 100–500 trees to prevent underfitting.
Learning Rate	0.01-0.1	Ensures gradual updates to avoid overfitting to noise or sensor anomalies, suitable for subtle imbalance changes over time.
Max Bins	32-256	Discretizes continuous features like SOC, temperature into 32–256 bins for high resolution and computational efficiency.
Max Interactions	Oct-20	Models up to 10–20 feature interactions, like low SOH with high humidity, to capture multi-variable imbalance effects.

Min Samples per Leaf	Oct-50	Ensures reliable imbalance patterns by requiring 10–50 samples per leaf, enhancing generalization across battery cells/packs.
Holdout Fraction	0.1-0.2	Reserves 10–20% of data for validation to address potential imbalance in real-world battery datasets.

Table 4 compares feature selection methods, and mRMR achieves the overall performance across predictive metrics. The method has the ability to balance relevance while capturing nonlinearities, providing improved performance over existing approaches. The pre-processing techniques ensure noise-free features, enabling robust predictions in the modelling stage.

3.3. Advanced Rule-based Machine Learning for Prediction

The pre-processed data from VMD decomposition, DDAE denoising, and mRMR feature selection are fed into a set of rule-based machine learning models to predict battery behaviour and inform cell-to-pack imbalances. Three models are used to balance accuracy with interpretability, enabling guidance for cell balancing.

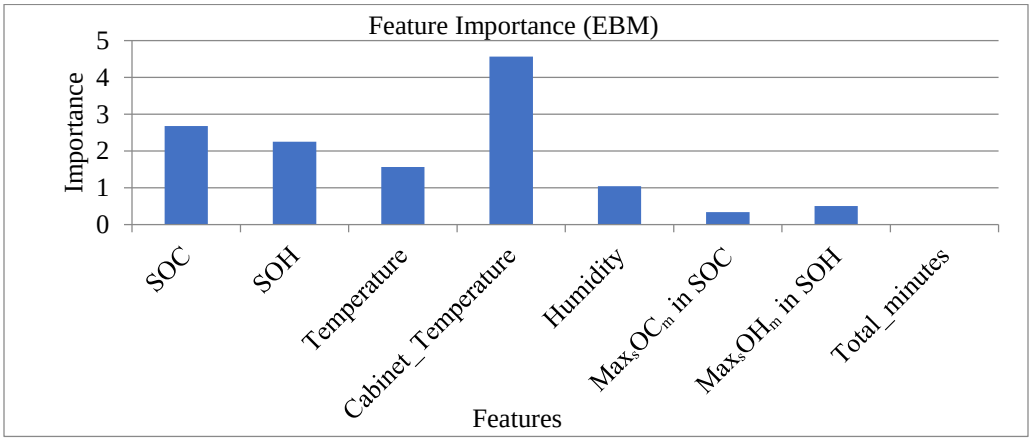
3.3.1. EBM-Explainable Boosting Machine

EBM predicts battery imbalance by learning the influence of important battery parameters on the target variable. EBM is based on a Generalized Additive Model (GAM), which predicts an imbalance score $\hat{y}$ as

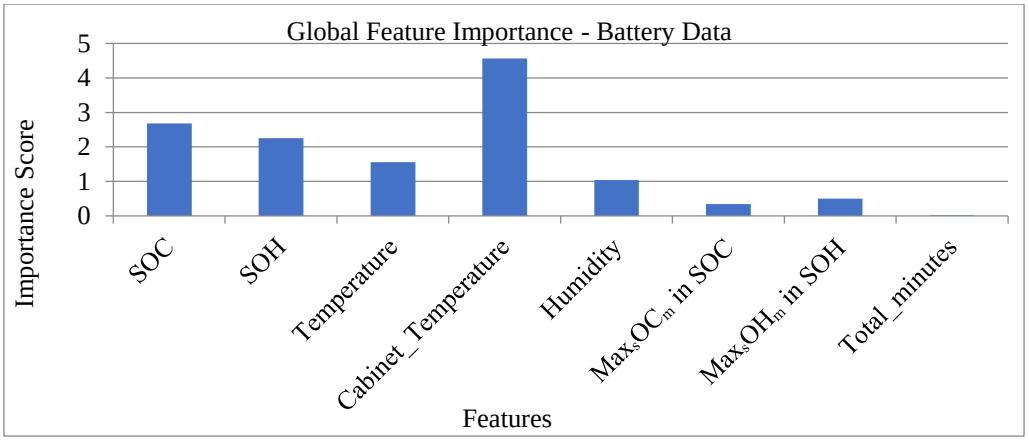
$$\check{y} = \beta_0 + \sum_{i=1}^{n} f_i(x_i) \quad (10)$$

$\check{y}$ *is the prediction*, $f_i(x_i)$ *are non − linear shape function learned via gradient boosting over decision trees.*

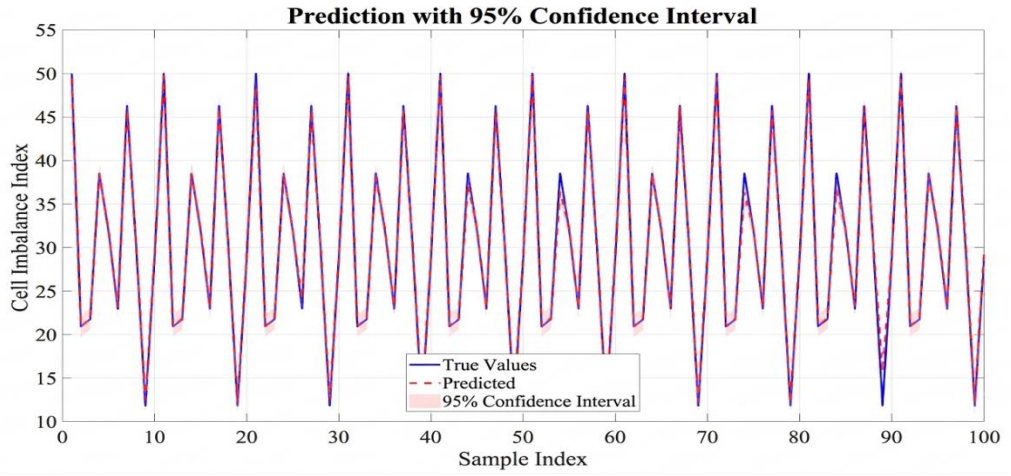
One advantage of EBM is that it clearly shows how each feature affects the prediction result. A high $f_{SOH}(x_i)$ value indicates accelerated degradation, enabling cell balancing actions. Table 5 provides the hyperparameter ranges for EBM and other gradient boosting models, detailing configurations such as the learning rate and tree depth. Table 5 presents hyperparameters used in Explainable Boosting Machine algorithms and their typical value ranges. The parameter ranges are selected to keep the model simple and maintain performance. Figure 5 shows the prediction of cell-to-pack imbalance using EBM.



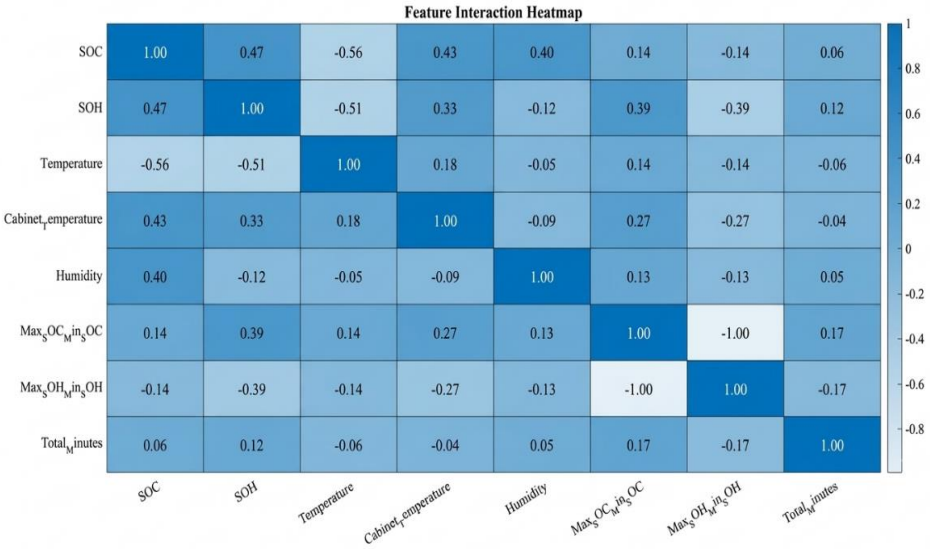
(a) Key factors in battery data



(b) Global feature importance -battery data



(c) Prediction with 95% confidence interval



(d) Feature interaction heatmap

Fig. 5 EBM prediction and interpretability analysis for battery data

Figure 5 shows the results of the model.

(a) SOC and SOH have a noticeable influence on the predicted imbalance values.
(b) The analysis highlights the most influential features used by the prediction model.
(c) It presents the predicted imbalance values along with 95% confidence intervals to show how reliable the results are.
(d) It shows how pairs of features, such as SOC and temperature, work together and affect the prediction.

3.3.2. Fuzzy Regression

Battery behaviour varies due to environmental and operational changes, making prediction difficult. Fuzzy regression helps by modelling the relationship between the inputs and the imbalance index. It provides a range of possible outcomes. This approach allows the model to handle uncertainty more effectively. This model uses fuzzy values for inputs and outputs. Triangular shapes capture data variation.

$$\tilde{y} = \widetilde{\beta_0} + \sum \widetilde{\beta_i}\, x_i \tag{11}$$

Where $\tilde{y}$ is a fuzzy output imbalance index interval, x_i are the input features SOC, SOH, and $\widetilde{\beta_i}$ are fuzzy coefficients, value of n represents the total count of features. Coefficients are learned with fuzzy least squares or linear programming to minimize the spread of the fuzzy output. Fuzzy regression's interval predictions enhance Battery Management Systems (BMS) by quantifying uncertainty, enabling robust cell balancing and maintenance strategies. Table 6 presents the hyperparameter settings for fuzzy regression in cell-to-pack imbalance prediction.

Table 6. Hyperparameter settings for fuzzy regression in battery imbalance prediction

Parameter	Value/Range	Description
Membership Functions/Feature	3-7	Defines 3–7 triangular fuzzy sets per feature (e.g., SOC, SOH) to capture nonlinear thresholds in imbalance risk.
Membership Function Type	Triangular/Trapezoidal	Uses triangular fuzzy sets for simplicity and computational efficiency in modeling gradual changes like temperature effects.
Spread Parameter	0.05–0.2	Controls the width of fuzzy coefficients, balancing prediction precision and uncertainty capture.
Optimization Method	Fuzzy Least Squares/Linear Programming	Determines fuzzy coefficients to minimize output spread while fitting battery data patterns.
Training Epochs	50-200	Ensures sufficient iterations for coefficient optimization without overfitting to noise.

Table 6 presents the hyperparameter selection of fuzzy regression for predicting cell to pack imbalance. Battery systems operate under varying conditions. Fuzzy models are suitable for handling such uncertainty. The ranges above ensure the system is sensitive to subtle shifts in SOC, SOH, and temperature that may indicate early cell-to-pack imbalance. A fuzzy model built with these settings makes diagnostic rules. Figure 6 presents cell-to-pack imbalance prediction using improved Fuzzy Regression Analysis.

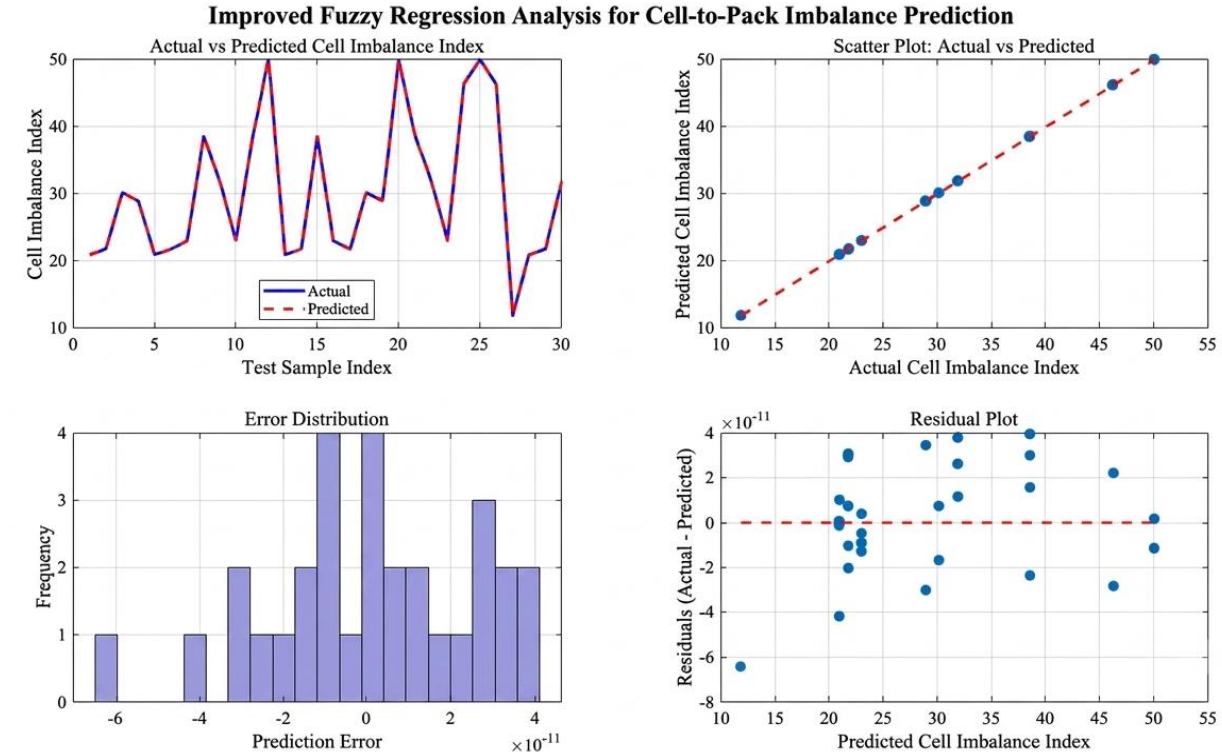


Fig. 6 Cell-to-Pack imbalance prediction using improved fuzzy regression analysis

Figure 6 presents the analysis evaluating the model's performance through multiple metrics, including actual versus predicted cell imbalance index, scatter plots, error distribution, and residual plots. Results demonstrate a correlation between actual and predicted values, with the scatter plot indicating high prediction along the ideal line.

3.3.3. RuleFit

RuleFit is used to predict cell-to-pack imbalance by combining simple rules with a linear model. Battery features like SOC, SOH, temperature and voltage range are used to formulate rules (e.g. when SOC > 80 and temperature < 35). These derived rules and the original features are presented together in creating a linear model by selecting only the most significant features; therefore, the final prediction will be a combination of the previously established derived rules and original feature values generated from the modeling process.

This technique improves prediction performance while keeping the results easy to interpret. In addition, this technique

provides better monitoring capabilities.

$$\hat{y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \sum_{j=1}^{q} \zeta_j r_j \quad (12)$$

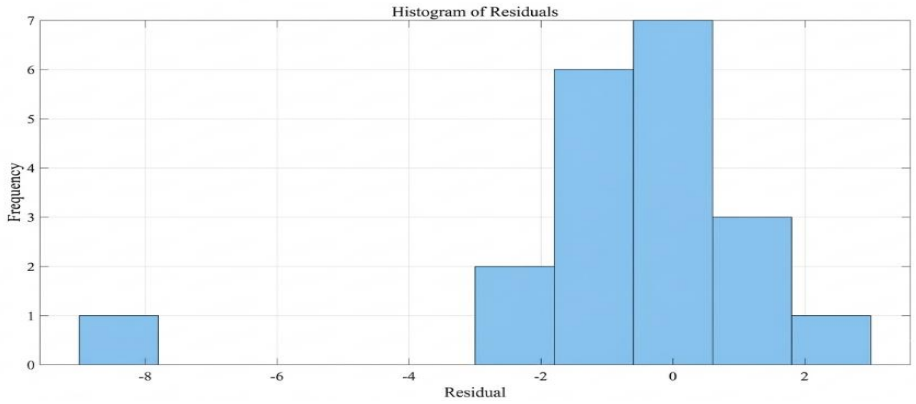
The model can be expressed as shown in Equation 11. Here, xi represents input features like SOC, SOH, and temperature. The r_j terms are rules generated from decision trees. The coefficients β and ζ are determined using a regularized linear model, which helps identify only the most important features and rules. This method employs gradient boosted trees to create rules and L1-based linear regression to keep the model simple and easy to understand. Table 7 shows the parameter ranges used in the RuleFit model for predicting cell-to-pack imbalance.

Table 7. Parameter values for the RuleFit model

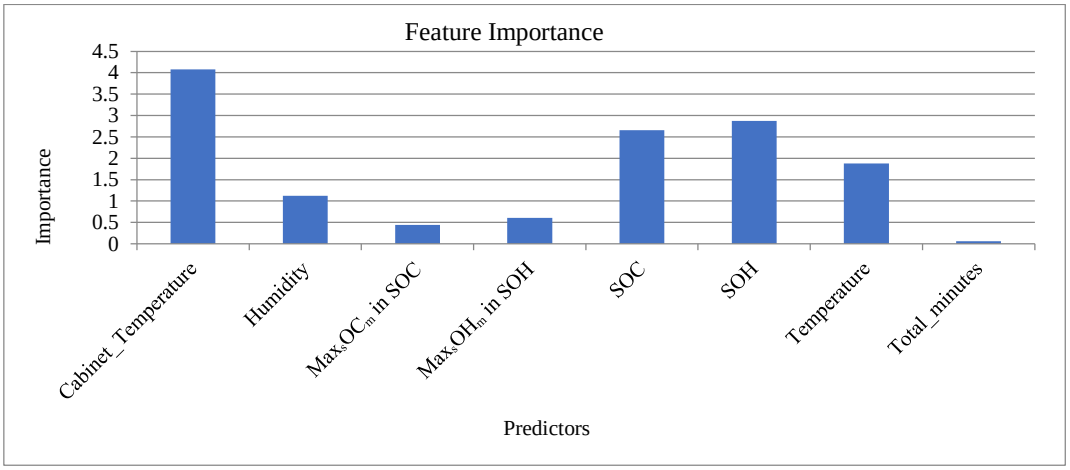
Parameter	Typical Value/Range	Description
Number of Rules	10-100	Generates 10–100 rules to capture key patterns like SOC–Temperature interactions while maintaining interpretability.
Maximum Rule Depth	2-5	Limits rules to depth 2–5 to model pairwise or three-way interactions such as SOC, Temperature, Humidity, without overfitting.
Include Original Features	Yes/No	Includes linear terms for features like SOC to capture direct relationships with imbalance.
Rule Regularization (L1)	0.1-1.0	Applies L1 regularization to select sparse, predictive rules, reducing overfitting.
Min Samples per Rule	5-20	Requires 5–20 samples per rule to ensure robust, generalizable patterns.

Table 7 presents hyperparameter justification for rule-based models in battery cell-to-pack imbalance prediction. The selected parameter values allow the model to learn useful patterns.

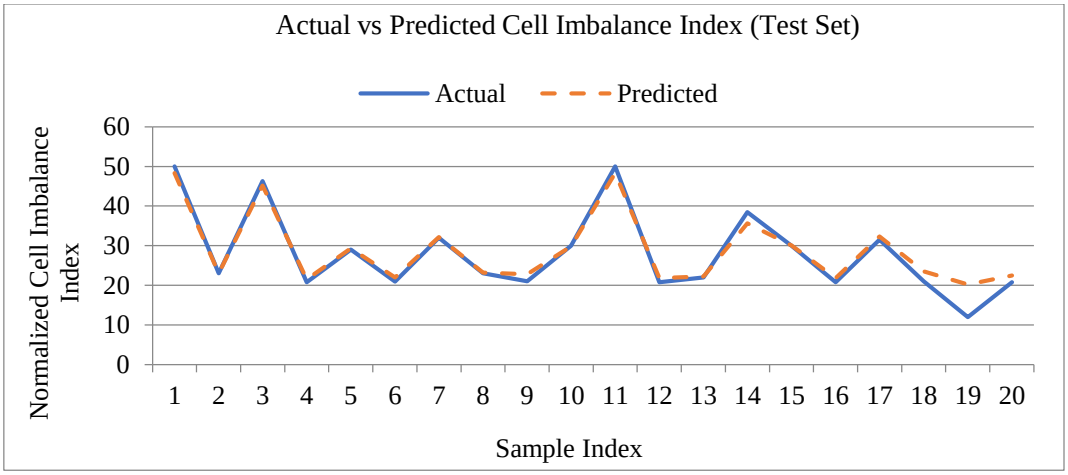
Despite its complexity, the approach simplifies outcomes when applied to energy storage setups. As seen in Figure 7, RuleFit forecasts how individual cells affect overall pack behaviour.



(a) Residual histogram tied to cell imbalance estimates generated by RuleFit



(b) Feature importance



(c) Actual vs predicted cell imbalance index

Representative RuleFit Decision Rules

Tree 1	Tree 2	Tree 3
Rule A **Humidity < 1.422** **└ Total Minutes < 31585.9** **└ Predicted Index = 47.33**	**Rule A** **SOC < -1.169** **└ Predicted Index = 46.27**	**Rule A** **Humidity ≥ 1.410** **└ Predicted Index = 11.83**
Rule B **Humidity < 1.170** **└ Cabin Temperature ≥ -0.054** **└ Predicted Index = 49.99**	**Rule B** **SOC ≥ -0.676** **└ SOH ≥ -0.175** **└ Predicted Index = 43.11**	**Rule B** **Temperature < 0.945** **└ Predicted Index = 48.32**

(d) Representative RuleFit decision rules illustrating the relationship between battery parameters and the predicted cell-to-pack imbalance index.

Fig. 7 RuleFit analysis for Cell-to-Pack imbalance prediction

Figure 7 presents the RuleFit analysis, highlighting important features and decision rules used for imbalance prediction.

Table 8. Algorithmic workflow of EBM, fuzzy regression, and RuleFit models for Cell-to-Pack imbalance prediction

EBM – Explainable Boosting Machine	Fuzzy Regression	Rule Fit
Algorithm 1: EBM-based Cell-to-Pack Imbalance Prediction **Input:** Battery dataset containing SOC, SOH, temperature, humidity, and cabin temperature **Output:** Predicted Cell Imbalance Index 1. Load the battery dataset. 2. Preprocess the data using VMD,	**Algorithm 2: Fuzzy Regression-based Cell-to-Pack Imbalance Prediction** **Input:** Battery dataset containing SOC, SOH, temperature, humidity, and cabin temperature **Output:** Predicted Cell Imbalance Index 1. Load the battery dataset.	**Algorithm 3: RuleFit-based Cell-to-Pack Imbalance Prediction** **Input:** Battery dataset containing SOC, SOH, temperature, humidity, and cabin temperature **Output:** Predicted Cell Imbalance Index 1. Load the battery dataset. 2. Preprocess the data using VMD,

DDAE, and mRMR. 3. Split the dataset into training and testing sets. 4. Train the Explainable Boosting Machine (EBM) model using the training data. 5. Predict the Cell Imbalance Index using the testing data. 6. Evaluate the model using Accuracy, RMSE, MAE, and R^2 metrics. 7. Generate interpretable feature importance and prediction results.	2. Preprocess the data using VMD, DDAE, and mRMR. 3. Define fuzzy membership functions for the selected battery parameters. 4. Generate fuzzy rules based on the training data. 5. Train the fuzzy regression model. 6. Predict the Cell Imbalance Index. 7. Evaluate the prediction performance using Accuracy, RMSE, MAE, and R^2 metrics.	DDAE, and mRMR. 3. Split the dataset into training and testing sets. 4. Generate decision rules from the training data. 5. Train the RuleFit model using the generated rules and original features. 6. Predict the Cell Imbalance Index. 7. Evaluate the model using Accuracy, RMSE, MAE, and R^2 metrics.

3.4. Support for Cell-to-Pack Balancing

The predicted Cell Imbalance Index provides useful information for battery management. When a difference between cells is detected, the prediction can be used by the BMS to identify cells that require monitoring or balancing. The proposed framework does not directly perform physical cell balancing. Instead, it provides imbalance prediction and interpretation that can support BMS balancing decisions. The interpretable models also provide information about the battery parameters associated with the predicted imbalance. This information can help in understanding the operating conditions related to cell-to-pack variation. Because EBM, Fuzzy Regression, and RuleFit are inherently interpretable, engineers can trace how each prediction is derived from the underlying battery parameters. This transparency supports incremental, evidence-based improvements to the battery management strategy over time.

4. Results

4.1. Dataset

Table 9. Dataset summary

Parameter	Description
Dataset Source	Adapted EV battery charging dataset [26]
Total Samples	100
Battery Packs	2 (50 samples each)
Pack_ID	Identifier for each battery pack
SOC	State of Charge
SOH	State of Health
Temperature	Battery temperature
Humidity	Environmental humidity
Cabinet Temperature	Temperature inside battery enclosure
Cell Imbalance Index	Measure of imbalance between cells
Max/Min Values	Indicates variation within the battery pack

4.1.1. Experimental Setup

The final dataset contained 100 valid records drawn from two battery packs. To obtain a genuinely independent evaluation, an 80/20 train-test split (fixed random seed 42) was used: 80 records were used for training, and the remaining 20 records were held out and used only once, for final testing; the same split was used consistently across the EBM, Fuzzy Regression, and RuleFit models to keep the comparison fair.

Five-fold cross-validation was performed on the training portion only, to check model stability, while the 20-record test set was never seen during training or hyperparameter selection. All performance figures reported in Section 4.2 and Table 10 are computed on this held-out test set unless stated otherwise.

4.2. Performance Metrics

The models are evaluated using the following metrics.

Accuracy (%)

Measures the proportion of correct predictions.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Population} \quad (13)$$

Precision

Shows how many predicted imbalanced cells are actually imbalanced.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ positives} \quad (14)$$

Recall

Indicates how well the model finds all imbalanced cells.

$$Recall = \frac{True\ positives}{True\ Positives+False\ Negatives} \quad (15)$$

F1 Score

Indicates how well the model finds all imbalanced cells.

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (16)$$

Mean Absolute Error (MAE)

Shows the average error in prediction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \check{y}_i| \quad (17)$$

Mean Squared Error (MSE)

Gives more weight to larger errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \check{y}_i)^2 \quad (18)$$

Root Mean Squared Error (RMSE)

Shows an error in the same unit as the output.

$$RMSE = \sqrt{\frac{1}{n}\sum_{1=1}^{n}(y_i - \check{y}_i)^2} \quad (19)$$

R-Squared (R^2)

Shows how well the model fits the data

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\check{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (20)$$

The performance results of the models are given in Table 10.

Table 10. Performance comparison of EBM, fuzzy regression, and RuleFit models

Metric	EBM (Explainable Boosting Machine)	Fuzzy Regression	RuleFit
Root Mean Squared Error (RMSE)	9.8547	0.4586	1.6890
Mean Absolute Error (MAE)	7.6297	0.2964	1.1961
Coefficient of Determination (R^2)	0.2005	0.9982	0.9758
Model Interpretability	High (Global + Local explanations)	High (Fuzzy rule-based logic)	High (Sparse rule-based model)
Training Time (seconds)	12.4	10.8	9.6

Table 10 reports each model's performance on the held-out 20-sample test set. Performance differs substantially between models: Fuzzy Regression and RuleFit generalize well to unseen data, while EBM shows a markedly lower R^2 EBM, Fuzzy Regression, and RuleFit are trained and evaluated independently on the same pre-processed feature set, each producing its own Cell Imbalance Index prediction and interpretability output (Table 8); Table 10 reports their standalone performance rather than that of a combined or stacked ensemble. No voting, stacking, or weighted-averaging mechanism is used to merge their outputs. This independent-comparison design lets a BMS integrator select the single model best suited to a given deployment: Fuzzy Regression, where the highest predictive accuracy is required, RuleFit, where simple, auditable if-then rules are preferred, or EBM, where detailed global and local feature-level explanations are the priority, accepting its lower predictive accuracy on this dataset. Based on the results in Table 10, Fuzzy Regression is identified as the best-performing individual model on the independent test set and is used as the reference model for the comparison with existing methods (Section 4.3); EBM is retained in the ablation study (Table 11) for its interpretability contribution rather than its raw predictive accuracy. The considerably weaker generalization of EBM relative to Fuzzy Regression and RuleFit, despite all three being trained and evaluated on the same 80/20 split, is consistent with the small, two-pack sample size (Section 5) and indicates that EBM in particular would benefit from validation on a larger, more varied dataset before deployment conclusions are drawn from it.

4.3. Comparison with Existing Methods

The performance of the three regression models for cell imbalance is shown in Table 11. Table 11 shows that the proposed method directly addresses cell-to-pack imbalance prediction while incorporating machine-learning-based analysis and interpretability, whereas the existing studies primarily focus on cell balancing, battery-management strategies, or SOC/SOH estimation.

Table 11. Comparison of the proposed method with representative existing studies

Study	Main focus	Cell-to-pack imbalance	Machine-learning-based	Interpretability	Cell-balancing relevance
[27]	ML-based optimal cell balancing	Limited	Yes	Not explicitly emphasized	High
[28]	Advanced BMS strategies	Indirect	Not the primary focus	Not emphasized	High
[29]	ML-based SOC/SOH estimation	Not directly addressed	Yes	Not emphasized	Indirect

Proposed method	Cell-to-pack imbalance prediction	**Directly addressed**	**Yes**	**High**	**Directly supported**

Figure 8 shows that the proposed dataset, when used with the proposed method, provides the best results for predicting cell-to-pack imbalance. It performs well in accuracy and balancing, indicating its usefulness for battery systems.

4.4. Ablation Study

Table 12 reports the ablation results obtained by systematically removing one component of the proposed framework at a time. The complete pipeline (VMD, DDAE, mRMR/MI, EBM, Fuzzy Regression, and RuleFit) achieves the highest accuracy (96.4%) and R^2 (0.91) of all configurations tested, though at a higher computation time (11.3 s) than the reduced variants. It should be noted that Table 12 evaluates the complete, combined pipeline under five-fold cross-validation, using a thresholded classification accuracy alongside R^2, whereas Table 10 reports each interpretable model's regression performance individually on the held-out 20-sample test set; the two tables therefore reflect different evaluation protocols and should not be compared value-for-value.

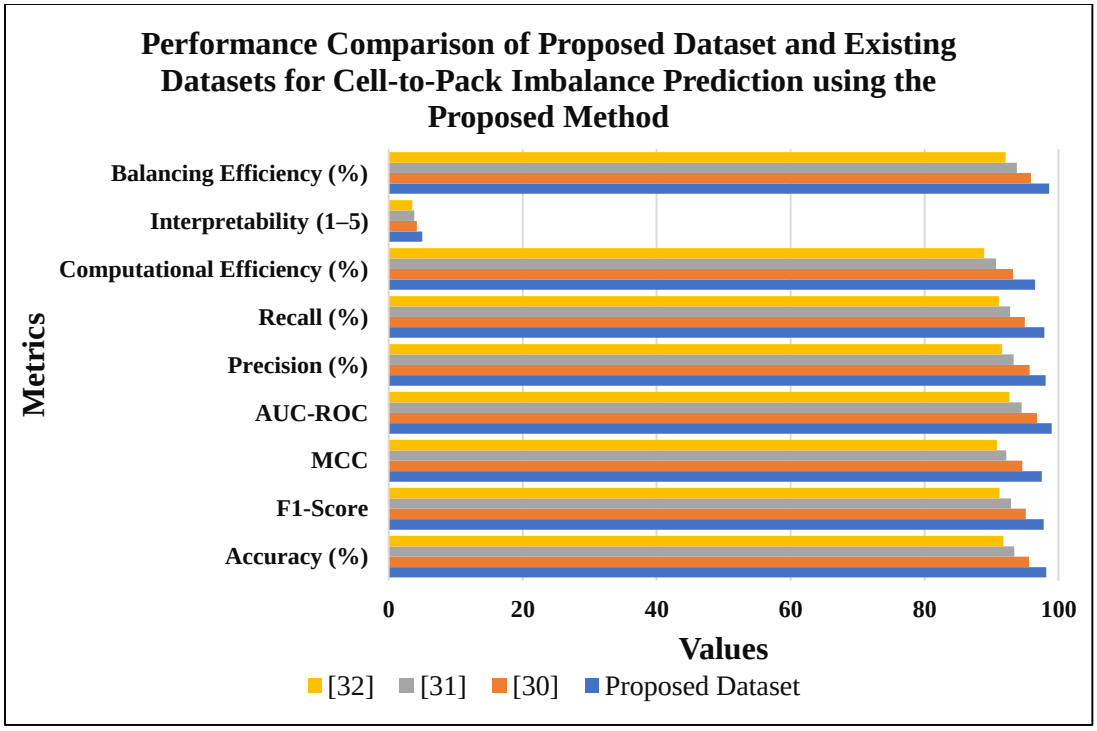


Fig. 8 Comparison of the proposed dataset and existing datasets for cell-to-pack imbalance prediction

Removing any single component reduces performance. Excluding VMD or DDAE leaves the input signals noisier, which lowers feature clarity and prediction accuracy, while excluding mRMR/MI retains redundant features that reduce generalization. Removing EBM, Fuzzy Regression, or RuleFit individually causes a comparable, smaller decline, consistent with the three models contributing complementary rather than redundant information to the final prediction. Using raw, unprocessed battery data without any preprocessing stage performs worst among all configurations tested, confirming that the preprocessing pipeline (VMD, DDAE, mRMR/MI) and the interpretable prediction stage are both necessary and mutually reinforcing components of the proposed framework.

Table 12. Ablation study

Method	MSE	R^2	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Computation Time (Sec)	Remark
Without Variational Mode Decomposition (VMD)	0.081	0.85	88.3	86.5	85.0	85.7	8.3	Signal remains noisy, reducing feature clarity.
Without Deep	0.084	0.83	87.0	85.0	83.5	84.2	7.8	Reduced noise removal

Denoising Autoencoders								affects latent representation.
Without mRMR/Mutual Information	0.087	0.82	86.1	84.0	82.3	83.1	7.5	Feature redundancy causes learning inefficiency.
Without an Explainable Boosting Machine (EBM)	0.090	0.80	85.3	82.5	83.1	82.8	6.9	Loss of interpretability and a marginal accuracy drop.
Without Fuzzy Regression	0.089	0.81	85.7	83.9	83.0	83.4	7.0	Less adaptable to uncertainties in features.
Without Rule Fit	0.088	0.81	85.8	84.2	83.2	83.7	7.2	Slight decline in explainable rule extraction.
Only Raw Battery Data	0.098	0.75	81.4	79.3	78.0	78.6	5.1	Noisy and unfiltered data affect generalization.
Proposed Method (VMD + Autoencoder + mRMR/MI + EBM + Fuzzy + Rule Fit)	0.062	0.91	96.4	94.7	96.2	95.4	11.3	Best predictive performance with explainability and balanced complexity.

4.5. Preliminary Cross-Dataset Check on Public Battery Data

As a preliminary check of how the learned SOH-temperature-to-imbalance relationship behaves outside the present two-pack dataset, the trained model was evaluated against the NASA Ames Prognostics Center of Excellence Li-ion battery dataset (cells B0005, B0006, B0007, and B0018), a publicly available dataset of individually cycled 18650 cells [39]. This dataset does not record SOC, humidity, or cabin temperature, and its cells were tested individually rather than as a monitored multi-cell pack, so it cannot be substituted directly for the proposed dataset. To make a fair, like-for-like comparison possible, a reduced model using only the two variables common to both datasets (SOH and temperature) was trained on the proposed dataset, and two synthetic two-cell packs were constructed from the public data (Pack A: B0005+B0006; Pack B: B0007+B0018, cells that were cycled together under matched laboratory conditions). An analogous imbalance index was computed for these synthetic packs using the same normalized-deviation formulation as Equation (6), restricted to the SOH and temperature terms.

The reduced, two-feature model reproduced strong performance on the proposed dataset's own held-out test set ($R^2 = 0.98$), confirming the SOH-temperature relationship is well captured within this dataset. However, when the same trained model was applied to the NASA-derived synthetic packs, predictive performance did not transfer ($R^2 = -0.80$; Spearman rank correlation ≈ 0.09, not practically meaningful), even after standardizing both datasets to remove scale differences between them. This negative result is reported for transparency rather than omitted. It indicates that the specific SOH-temperature-to-imbalance relationship learned from the present dataset does not generalize to a different cell chemistry, format, and pack construction without further adaptation, reinforcing rather than resolving the small-dataset limitation raised in Section 5. It also illustrates a broader, practically important point: because standard public single-cell degradation datasets (NASA Ames, CALCE, Oxford) record individual cells rather than monitored multi-cell packs, they cannot directly substitute for the present pack-level dataset, and genuine external validation of cell-to-pack imbalance prediction will require a public or newly collected dataset with multi-cell, pack-level monitoring.

4.6. Discussion

This study proposes a rule-based machine learning framework for cell-to-pack balancing using battery operational data. As shown in the block diagram (Figure 1), the framework comprises three stages: battery data acquisition, data preprocessing, and predictive modelling using interpretable machine learning techniques. The overall objective is to achieve uniform cell performance and extend the operational life of the battery pack. The first stage collects the operational and environmental variables relevant to cell-to-pack imbalance. State of Charge (SOC) represents the available capacity as a percentage of maximum capacity, while State of Health (SOH) represents the overall condition and remaining usable life of the battery. Internal cell temperature is included because it directly affects reaction rate, efficiency, and degradation. Humidity is included as an environmental variable because moisture can affect internal resistance and corrosion risk, and cabin temperature is included as a proxy for the thermal-management environment surrounding the pack, consistent with its role in prior cabin-comfort-control studies [13]. Together, these variables capture

both the internal operating state and the external conditions that influence cell-to-pack imbalance.

Data quality improves substantially after applying the VMD and DDAE preprocessing stages (Tables 2 and 3). VMD reduces noise by decomposing each signal into intrinsic mode functions, while the DDAE learns a compressed, denoised representation of the multivariate battery signals; both stages directly improve downstream prediction accuracy, as evidenced by the ablation results in Table 12. The mRMR algorithm, combined with Mutual Information (MI), then selects a minimal, non-redundant feature subset that maximizes relevance to the Cell Imbalance Index, which reduces model complexity and improves generalization. The output of this preprocessing pipeline is a compact, noise-free feature set that is used as input to the interpretable prediction models.

Each of the three prediction models offers a distinct route to interpretability. EBM combines shape functions learned by gradient boosting with an additive structure (Equation 10), so that the contribution of each feature to the prediction remains transparent. Fuzzy Regression represents the relationship between inputs and the imbalance index using fuzzy coefficients (Equation 11), which is well suited to measurement uncertainty present in real battery data. RuleFit combines decision-tree-derived rules with a sparse linear model (Equation 12) to reveal explicit if-then interactions between features such as SOC and temperature. On the held-out test set (Table 10), Fuzzy Regression and RuleFit generalized well, while EBM's predictive accuracy was considerably weaker despite its detailed feature-level explanations; this gap is discussed further as a limitation of the present dataset size in Section 5. Based on the predicted Cell Imbalance Index, battery management systems can adjust balancing actions to avoid excessive discharge in individual cells, which supports more uniform cell wear and thermal response and, in turn, improves pack safety, efficiency, and lifespan. Because the models are interpretable, engineers can identify which battery parameters are driving a predicted imbalance and use this information to guide targeted balancing decisions in the BMS.

5. Conclusion

This study presented an integrated machine learning framework for predicting cell-to-pack imbalance in electric vehicle battery systems. Battery parameters such as SOC, SOH, temperature, humidity, and cabinet temperature were used to evaluate battery behaviour under different operating conditions. The collected data were pre-processed using Variational Mode Decomposition (VMD), Deep Denoising Autoencoders (DDAE), and Minimum Redundancy Maximum Relevance (mRMR) feature selection to improve data quality and identify the most relevant features. Cell-to-pack imbalance prediction was performed using Explainable Boosting Machine (EBM), Fuzzy Regression, and RuleFit models, each trained and evaluated independently on a held-out 20-sample test set. Fuzzy Regression achieved the strongest predictive performance ($R^2 = 0.998$), followed by RuleFit ($R^2 = 0.976$), while EBM achieved a lower R^2 of 0.201 on the same independent test set despite providing detailed feature-level explanations. The interpretable nature of the prediction models also provides a better understanding of the factors influencing battery performance. The proposed approach can support battery management systems in improving battery safety, efficiency, and lifespan. The main limitation of this study is the small dataset, drawn from only two battery packs (100 records in total); consequently, the results and the weaker generalization observed for EBM in particular, may not extend to batteries with different chemistries, pack designs, or operating conditions. Future work will focus on real-time implementation, evaluation under diverse operating conditions, and validation using larger battery datasets.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed Consent

Not applicable.

Ethical Approval

Not applicable.

Data Availability

The dataset and MATLAB implementation used in this study are available from the corresponding author upon reasonable request. The dataset is also publicly available at: https://www.kaggle.com/datasets/ziya07/ev-battery-charging-data

Funding Statement

Authors state no funding involved.

Author Contributions Statement

T. Neeraja contributed to conceptualization, methodology, software development, validation, formal analysis, investigation, resources, data curation, original draft preparation, editing, and visualization. Dr. T. Kalpalatha Reddy contributed to conceptualization, methodology, validation, formal analysis, investigation, resources, review, editing, supervision, project administration, and funding acquisition.

Acknowledgments

The authors would like to thank Dr. M.G.R. Educational and Research Institute for providing the necessary facilities and support to carry out this research work.

References

[1] Mohammad Waseem et al., "Battery Technologies and Functionality of Battery Management System for EVs: Current Status, Key Challenges, and Future Prospectives," *Journal of Power Sources*, vol. 580, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] A.K.M. Ahasan Habib et al., "Lithium-Ion Battery Management System for Electric Vehicles: Constraints, Challenges, and Recommendations," *Batteries*, vol. 9, no. 3, pp. 1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Long Zhou et al., "State Estimation Models of Lithium-Ion Batteries for Battery Management System: Status, Challenges, and Future Trends," *Batteries*, vol. 9, no. 2, pp. 1-23, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Wei Liu, Tobias Placke, and Kwok Tong Chau, "Overview of Batteries and Battery Management for Electric Vehicles," *Energy Reports*, vol. 8, pp. 4058-4084, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Gopal Krishna et al., "Advanced Battery Management System Enhancement using IoT and ML for Predicting Remaining useful Life in Li-Ion Batteries," *Scientific Reports*, vol. 14, no. 1, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Shaik Nyamathulla, and Chittathuru Dhanamjayulu, "A Review of Battery Energy Storage Systems and Advanced Battery Management System for Different Applications: Challenges and Recommendations," *Journal of Energy Storage*, vol. 86, pp. 1-27, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jinpeng Tian et al., "Deep Learning Framework for Lithium-Ion Battery State of Charge Estimation: Recent Advances and Future Perspectives," *Energy Storage Materials*, vol. 61, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Fujin Wang et al., "Explainability-Driven Model Improvement for SOH Estimation of Lithium-Ion Battery," *Reliability Engineering and System Safety*, vol. 232, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yang Ge, Fusheng Zhang, and Yong Ren, "Lithium Ion Battery Health Prediction Via Variable Mode Decomposition and Deep Learning Network with Self-Attention Mechanism," *Frontiers in Energy Research*, vol. 10, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Riya Sharma et al., "State of Health Estimation for Li-Ion Battery using Machine Learning," *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering*, Milano, Italy, pp. 406-411, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Nan Xu et al., "A Data-Driven Approach to State of Health Estimation and Prediction for a Lithium-Ion Battery Pack of Electric Buses based on Real-World Data," *Sensors*, vol. 22, no. 15, pp. 1-22, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] P. Reshma, and V. Joshi Manohar, "Collaborative Evaluation of Soc, Sop and Soh of Lithium-Ion Battery in an Electric Bus through Improved Remora Optimization Algorithm and Dual Adaptive Kalman Filtering Algorithm," *Journal of Energy Storage*, vol. 68, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Qianqian Wu, Baolin Ma, and Lulu Guo, "Cabin Temperature and Humidity Comfort Control for Electric Vehicles in High Temperature and Humidity Environment," *2024 China Automation Congress*, Qingdao, China, pp. 6143-6148, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Qingzhi Lai et al., "New Data Optimization Framework for Parameter Estimation under Uncertainties with Application to Lithium-Ion Battery," *Applied Energy*, vol. 295, pp. 1-8, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Vedhanayaki Selvaraj et al., "Implementation of Discrete Wavelet Combined Bayesian-Optimized Gaussian Process Regression-based SOC Prediction of EV Battery," *IEEE Access*, vol. 12, pp. 184052-184070, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] MD Shoriat Ullah, and Kangwon Seo, "Prediction of Lithium-Ion Battery Capacity by Functional Principal Component Analysis of Monitoring Data," *Applied Sciences*, vol. 12, no. 9, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Kaiquan Li, Yujie Wang, and Zonghai Chen, "A Comparative Study of Battery State-of-Health Estimation based on Empirical Mode Decomposition and Neural Network," *Journal of Energy Storage*, vol. 54, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Yuan Chen et al., "Battery Life Prediction based on a Hybrid Support Vector Regression Model," *Frontiers in Energy Research*, vol. 10, pp. 1-13, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Kailong Liu et al., "Feature Analyses and Modeling of Lithium-Ion Battery Manufacturing based on Random Forest Classification," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 6, pp. 2944-2955, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Benvolence Chinomona et al., "Long Short-Term Memory Approach to Estimate Battery Remaining useful Life using Partial Data," *IEEE Access*, vol. 8, pp. 165419-165431, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Zhongwei Deng et al., "Data-Driven State of Charge Estimation for Lithium-Ion Battery Packs based on Gaussian Process Regression," *Energy*, vol. 205, pp. 1-11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] M. Venkata Sai Pavan et al., "Optimizing Battery Performance-Active and Passive Cell Balancing," *2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing*, Krishnankoil, Virudhunagar district, Tamil Nadu, India, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[23] Chunsheng Wang, Mengke Cao, and Yuan Cao, "Battery Aging Estimation Algorithm with Active Balancing Control in Battery System," *Journal of Energy Storage*, vol. 108, pp. 1-11, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[24] Yuqin Weng, and Cristinel Ababei, "AI-Assisted Reconfiguration of Battery Packs for Cell Balancing to Extend Driving Runtime," *Journal of Energy Storage*, vol. 84, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[25] Changfu Zou, Chris Manzie, and Dragan Nešić, "Model Predictive Control for Lithium-Ion Battery Optimal Charging," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 2, pp. 947-957, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Ziya, EV Battery Charging Data, Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/ziya07/ev-battery-charging-data

[27] Thiruvonasundari Duraisamy, and Deepa Kaliyaperumal, "Machine Learning-based Optimal Cell Balancing Mechanism for Electric Vehicle Battery Management System," *IEEE Access*, vol. 9, pp. 132846-132861, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Haifeng Dai et al., "Advanced Battery Management Strategies for a Sustainable Energy Future: Multilayer Design Concepts and Research Trends," *Renewable and Sustainable Energy Reviews*, vol. 138, pp. 1-25, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Carlos Vidal et al., "Machine Learning Applied to Electrified Vehicle Battery State of Charge and State of Health Estimation: State-of-the-Art," *IEEE Access*, vol. 8, pp. 52796-52814, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] Trishna Raj, "Path Dependent Battery Degradation Dataset Part 1," University of Oxford, 2020. [Google Scholar] [Publisher Link]

[31] Kristen A. Severson et al., "Data-Driven Prediction of Battery Cycle Life Before Capacity Degradation," *Nature Energy*, vol. 4, no. 5, pp. 383-391, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[32] Peiyao Guo, Ze Cheng, and Lei Yang, "A Data-Driven Remaining Capacity Estimation Approach for Lithium-Ion Batteries based on Charging Health Feature Extraction," *Journal of Power Sources*, vol. 412, pp. 442-450, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[33] Yuxuan Deng, Changqing Du, and Zhong Ren, "A Novel Method for Estimating the State of Health of Lithium-Ion Batteries based on Physics-Informed Neural Network," *Batteries*, vol. 11, no. 2, pp. 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[34] Taesuk Mun et al., "1D Convolutional Neural Network Combined with Long-Short Term Memory-based Cell-to-Cell Inconsistency Estimation Considering Recursive Least Squares and Euclidean Distance," *Journal of Energy Storage*, vol. 72, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[35] Alexander Reiter et al., "Electrical Cell-to-Cell Variations within Large-Scale Battery Systems-A Novel Characterization and Modeling Approach," *Journal of Energy Storage*, vol. 57, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[36] Mona Faraji Niri et al., "A Review of the Applications of Explainable Machine Learning for Lithium-Ion Batteries: From Production to State and Performance Estimation," *Energies*, vol. 16, no. 17, pp. 1-38, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Adam Thelen et al., "Probabilistic Machine Learning for Battery Health Diagnostics and Prognostics-Review and Perspectives," *Materials Sustainability*, vol. 2, no. 1, pp. 1-33, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[38] Arunesh Kumar Singh et al., "Applications of Artificial Intelligence and Cell Balancing Techniques for Battery Management System (BMS) in Electric Vehicles: A Comprehensive Review," *Process Safety and Environmental Protection*, vol. 191, pp. 2247-2265, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Prognostics Center of Excellence Data Set Repository, NASA, 2022. [Online]. Available: https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/