Big Data Analysis Is Used To Analyze The Health State Of The Lithium Batteries In New Energy Vehicles

Xiaojing Chu[1], Chenhong Feng[1*], and Yongming Ding[2]

[1] School of Automotive Engineering,Tianjin Vocational Institute
[2] Tianjin Shitong Network Technology Co., Ltd
[*] Corresponding author. E-mail: 17720072187@163.com

Received: Apr. 25, 2026; Accepted: May 27, 2026

This paper proposes a unique approach based on big data analysis to address the challenge of predicting the state of health (SOH) of lithium batteries for future energy vehicles. By combining a convolutional neural network (CNN) with a bidirectional long and short-term memory neural network (BiLSTM) and automatically modifying the model parameters using the Bayesian optimization (BO) technique, an accurate forecast of the battery state of health is obtained. Several health parameters are extracted from the NASA-provided public lithium battery dataset, such as the constant current charging time (CCCT) and discharge voltage plateau time (DVPT). The next step is to employ wavelet packet transform to lessen data noise. The experimental results show that the proposed method outperforms both a single neural network model and the conventional approach, with both average absolute error and root mean square error being less than 1%. Thus, this work offers a technical route and theoretical support for the safe and dependable functioning of electric car power batteries in addition to an effective SOH estimation method. Specifically, the proposed BO-CNN-BiLSTM model achieves a minimum MAE of 0.288% and RMSE of 0.365%, significantly lower than baseline models such as CNN and LSTM.

Keywords: lithium battery; SOH; big data analytics; Bi-LSTM; Bayesian optimization algorithm.

http://dx.doi.org/10.6180/jase.202611_34.057

1. Introduction

Over time, the performance of power batteries in electric vehicles degrades, making accurate battery health monitoring essential for safe and efficient operation [1]. Although the experimental validation is conducted on a publicly available dataset of limited size, the proposed framework is designed within a big data analytics paradigm focusing on high-dimensional feature extraction and scalable model architecture. State of health (SOH) is an important indicator used by battery management systems (BMS) to evaluate battery degradation under varying operating conditions. Since direct capacity estimation is difficult under real vehicle conditions, more practical SOH estimation methods are required. Existing SOH estimation techniques are broadly classified into data-driven, model-based, and experimental methods [2]. Experimental approaches are accurate but unsuitable for online estimation, whereas model-based methods, including electrochemical and equivalent circuit models, estimate battery health using battery characteristics and operating parameters [3–6].

Data-driven methods estimate battery SOH using historical charging and discharging data such as current, voltage, and temperature without relying on internal battery models [7]. Relevant features are extracted from battery charging behaviour to train machine learning models for SOH prediction. Previous studies have employed CC-CV charging features, segmented charging data, and machine learning techniques such as support vector machines and random forests to improve prediction accuracy [8–11].

Deep learning approaches further enhance feature extraction and correlation learning from vehicle operation data, although data quality and feature extraction remain challenging under real driving conditions [12]. The proposed BO-CNN-BiLSTM framework combines feature denoising, spatial–temporal learning, and hyperparameter optimization for improved SOH estimation.

Incremental capacity (IC) analysis is used to extract battery degradation features from real vehicle operation data, where IC curve characteristics help improve SOH estimation despite noise and irregular charging segments [13]. To enhance data quality, similarity-based approaches such as clustering and dynamic time regularization (DTW) are applied to identify reliable charging cycles and remove low-quality samples [14, 15]. These strategies improve feature quality and prediction accuracy, although IC and DTW are discussed only for data quality enhancement and are not directly used in the final BO-CNN-BiLSTM prediction model.

2. Data sources and pre-processing

2.1. Introduction to Data Sources

The battery dataset [16] contains the information needed for both model validation and training. Different battery subsets from the NASA dataset are used for specific tasks in this study. The CS2-series batteries are applied for feature analysis, while B0005 and B0018 are used for model training and validation. Although different battery subsets are utilized for feature extraction and model validation, all datasets originate from the same NASA lithium-ion battery repository and follow consistent charging and discharging protocols. This separation ensures unbiased evaluation and improves model generalization. In batteries with a nominal capacity of 1100m Ah, graphite carbon is employed as the negative electrode material and Li Co O2 as the positive electrode material. The nominal capacity value reported in this study corresponds to the effective operational capacity observed in the selected experimental cycles of the NASA dataset rather than the manufacturer-rated maximum capacity. All batteries were subjected to 800–1000 cycles of constant current and voltage charging at 0.5C and constant current and discharge cycling at 1C, with a 2.7V discharging cutoff voltage and a 4.2V charging cutoff voltage. The battery samples exhibited variability in degradation behaviour, discharge duration, internal resistance progression, and cycle-dependent capacity reduction across different operating stages.

2.2. Battery aging characteristics

To reduce the time needed to create the model and improve prediction accuracy, the correlation between a number of distinctive indicators and SOH is computed using Pearson's correlation coefficient. The following formula can be used to determine the battery SOH as Eq. (1).

$$SOH = \frac{Q_{\text{current}}}{Q_{\text{nominal}}} \times 100\% \tag{1}$$

Where SOH is a way to describe how good a battery is compared and Q_{current} symbolizes the current capacity of the batteryand Q_{nominal} its nominal capacity. The Pearson correlation coefficient ranges from -1 to 1 , where values closer to ± 1 indicate stronger relationships between variables. Battery aging data, including voltage, current, capacity, temperature, and internal resistance, were processed to extract key health indicators such as constant current charging time (CCCT), constant voltage charging time (CVCT), discharge voltage plateau time (DVPT), and internal resistance for SOH estimation [12].

The Pearson's correlation coefficient (r) between these indications and the condition of the battery can be found using Eq. (2). The specific computation findings are shown in Table 1. A Pearson correlation threshold of $|r| > 0.80$ was selected to identify strongly related SOH indicators. CCCT and DVPT consistently showed high correlation values across battery samples CS2-35 to CS2-38. This consistency confirms their reliability as predictive indicators for battery SOH. Pearson correlation analysis showed that CCCT and DVPT had a strong relationship with battery SOH across all battery samples, while CVCT and internal resistance exhibited weaker correlations in some cases. Considering both prediction relevance and ease of data collection in real vehicle conditions, CCCT and DVPT were selected as the primary indicators for SOH prediction.

$$r = \frac{\sum_{i=1}^{n}(X - \bar{X})(Y - \bar{Y})}{\left(\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right)\left(\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}\right)} \tag{2}$$

2.3. Data preprocessing

(1) Reducing Noise in Data Using Wavelet Packet Transform

Data noise in the measured data is inevitable due to sensor faults and other external interference sources, which has a negative impact on the model's training and prediction. An efficient adaptive time-frequency signal processing technique that can extract both high-frequency and low-frequency signals adaptively is wavelet packet transform, and its calculation formula is represented by Eq. (3).

Table 1. Shows the Pearson's correlation value between the health of the battery and distinctive parameters.

Characteristic indicators		CCCT	CVCT	DVPT	Internal resistance	Significance
Pearson correlation coefficient (r)	CS235	0.981	-0.065	0.998	-0.898	$p < 0.05$
	CS236	0.972	-0.565	0.994	-0.965	$p < 0.05$
	CS237	0.966	-0.621	0.984	-0.958	$p < 0.05$
	CS238	0.958	-0.522	0.968	-0.061	$p < 0.05$

$$d_j' = \begin{cases} 0 & \left|d_j\right| < \lambda_j \\ \operatorname{sign}\left(d_j\right)\left(\left|d_j\right| - \lambda_j\right) & \left|d_j\right| \geq \lambda_j \end{cases} \tag{3}$$

Here, λj represents the threshold value used for signal denoising. The heuristic threshold criterion with soft thresholding was applied to all decomposed sub-bands, and a three-level wavelet packet decomposition was selected to balance noise reduction, feature preservation, and computational efficiency. as Eq. (4):

$$\lambda_j = \begin{cases} \sigma\sqrt{2\ln N} & \alpha > \beta \\ \sigma\sqrt{Q_a} & \alpha < \beta \end{cases} \tag{4}$$

where is the sum of squares of the N decomposition coefficients. The correlation coefficients for both CCCT and DVPT are improved in all four sample sets. The selected decomposition level and thresholding strategy directly influence the balance between noise suppression and signal preservation, where insufficient decomposition may retain noise while excessive decomposition may distort useful features. Table 2 offers the Pearson's correlation coefficients for calculating the correlations between the battery SOH and the aging features.

(2) Normalization of data

The normalization process is expressed by Equation 5, where X stands for the normalized variable, X_{scaled} for the normalized value, and $X_{\min}$ for the variable's minimum value. The magnitude of different variables affects the fitting accuracy of the neural network model, so it is necessary to normalize the maximum and minimum values of all variables to the $[0,1]$ interval. Compared with alternative normalization approaches such as z-score standardization, min-max normalization was selected because it preserves the relative distribution of battery feature values within a fixed range.

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{5}$$

Based on the actual values of the publicly accessible dataset, the maximum and minimum values of the SOH characterisation metrics are established, i.e Eq. (6).

$$[\,\text{CCCTDVPTSOHCCCTDVPT}_{\ \min\ _{\min}}\ \text{SOH}_{\min}\,]_{\max\ _{\max}\ _{\max}} \tag{6}$$

3. Bo-cnn-bilstm model construction

3.1. CNN-BiLSTM network construction

In order to adaptively extract data features and apply convolutional operations to the input information, the convolutional layer in CNN makes use of a convolutional kernel. To achieve dimensionality reduction, a pooling window is slid based on the feature matrix produced from the convolutional layer, and the maximum value of the pooling window is chosen as the output for each slide. The features are used for prediction by the output layer and the fully connected layer. The calculation formula is Eqs. (7) to (10)

$$c_i = f\left(w_i \cdot x_i + b_i\right) \tag{7}$$

$$\gamma\left(c_i, c_{i-1}\right) = \max\left(c_i, c_{i-1}\right) \tag{8}$$

$$p_i = \gamma\left(c_i, c_{i-1}\right) + \beta_i \tag{9}$$

$$y_i = f\left(t_i p_i + \delta_i\right) \tag{10}$$

Where: $\boldsymbol{c}_i$ is the ith output vector; $f()$ is the function that activates; $\boldsymbol{w}_i$ is the weight matrix; I is the convolution kernel count; • denotes the dot product; $\boldsymbol{x}_i$ is the convolution layer input; $\boldsymbol{b}_i$ is the bias term; $\gamma()$ is the maximum pooling sub-sampling function; $\boldsymbol{p}_i$ is the output of the maximum pooling layer; β_i is the bias; $\boldsymbol{y}_i$ is the output vector; it is the weight matrix; δ_i is the bias.

The forward LSTM hidden layer ascends to read the information sequence output by the CNN, while the inverse LSTM hidden layer descends to read the information sequence output by the CNN [11]. BiLSTM is made up of two inverse LSTM layers, and its computational method. The model was trained using the Adam optimizer with a learning rate of 0.001 , a batch size of 32 , and a maximum of 100 epochs is as follows Eqs. (11) to (13)

$$h_t = \vec{h}_t * \bar{h}_t \tag{11}$$

$$\vec{h}_t = \text{LSTM}\left(x_t, \vec{h}_{t-1}\right) \tag{12}$$

$$\bar{h}_t = \text{LSTM}\left(x_t, \bar{h}_{t+1}\right) \tag{13}$$

Table 2. Pearson correlation coefficients following the removal of feature noise.

Characteristic indicators		CCCT	DVPT	Significance
Pearson correlation coefficient	CS2-35	0.974	0.988	$p < 0.05$
	CS2-36	0.972	0.994	$p < 0.05$
	CS2-37	0.961	0.987	$p < 0.05$
	CS2-38	0.955	0.974	$p < 0.05$

The CNN-BiLSTM network combines CNN for local feature extraction and BiLSTM for capturing spatio-temporal battery degradation patterns to improve SOH prediction. As shown in Fig. 1, CCCT and DVPT features are processed through CNN layers with max-pooling and then forwarded to the BiLSTM layer for temporal modeling and prediction.

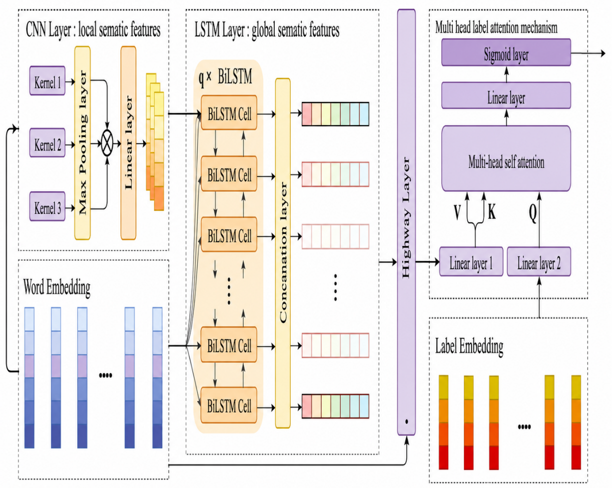

Fig. 1. CNN-BiLSTM model structure.

Fig. 1 includes architectural details such as convolutional layer configuration, kernel sizes, activation functions, pooling operations, and BiLSTM hidden layer settings to improve model reproducibility. The CNN component applies one-dimensional convolutions with ReLU activation followed by max-pooling for feature extraction and dimensionality reduction. The extracted features are then reshaped and passed to the BiLSTM layer for temporal dependency learning and SOH prediction.

3.2. BO-CNN-BiLSTM network construction

Bayesian Optimization (BO) was employed to automatically optimize CNN-BiLSTM hyperparameters, including learning rate, hidden layer nodes, and regularization coefficients, to improve prediction accuracy. Using Bayes theorem, Gaussian process regression, and the Expected Improvement (EI) function, BO iteratively searched for optimal parameters through data preprocessing, hyperparameter tuning, and model prediction, reducing manual effort and improving search efficiency.

4. Test results and analysis

4.1. The BO's outcomes-Error analysis and prediction using CNN-BiLSTM

The training set for the NASA dataset consisted of the first 70% of cycles, while the test set was made up of the remaining 30%. A sequential partitioning strategy was adopted instead of randomized splitting to preserve the temporal dependency of battery degradation data. The train-test split was performed sequentially based on the battery cycling order to preserve the temporal dependency of the time-series degradation data. SOH prediction is carried out based on a variety of health indicators after the CNN-BiLSTM model's three parameters—learning rate, hidden layer nodes, and regularization coefficients—are optimally solved using the BO algorithm. The prediction accuracy of the combined model is assessed using MAE and RMSE, and the findings are compared with those of the CNN, LSTM, BiLSTM, and CNN-BiLSTM models. The hyperparameters of the baseline models were selected within commonly adopted optimal ranges to avoid performance bias during comparison. Table 3 displays the ideal characteristics that the BO algorithm looked for in relation to batteries B0005 and B0018.

Fig. 2 compares the prediction performance of CNN, LSTM, BiLSTM, CNN-BiLSTM, and BO-CNN-BiLSTM models, where the proposed BO-CNN-BiLSTM achieved better fitting accuracy for SOH estimation. While LSTM and BiLSTM improved time-series learning and CNN-BiLSTM enhanced feature extraction, the BO-CNN-BiLSTM model reduced overfitting and provided more accurate prediction performance. This divergence between training and validation performance confirms the presence of overfitting in the CNN-BiLSTM model before optimization.

Small changes in CNN-BiLSTM hyperparameters can significantly affect prediction accuracy; therefore, Bayesian Optimization (BO) was applied to automatically determine optimal parameter settings, improving efficiency and reducing manual tuning. As shown in Tables 4 and 5, the BO-CNN-BiLSTM model achieved the lowest prediction errors, with a maximum MAE of 0.288% and RMSE of 0.365%,

Table 3. Results of the ideal hyperparameter search.

Battery	Initial learning rate	Hidden layer nodes	Regularization coefficient
B0005	8.826×10^{-3}	113	3.878×10^{-5}
B0018	1.288×10^{-2}	35	5.545×10^{-5}

demonstrating improved SOH prediction performance.

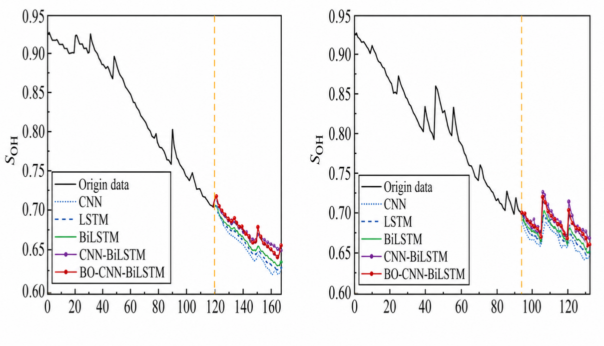

Fig. 2. Prediction results of 5 models.

4.2. Combined model practical performance test

To examine the BO-CNN-BiLSTM model's practical performance, the training set consists of the first 50%, 60%, and 70% cycles of the B0005 and B0018 batteries, whereas the test set consists of the subsequent cycles. The SOH prediction results based on different training sets are shown in Figs. 3 and 4.

Tables 6 and 7 display the SOH prediction errors of the various training sets. As the number of cycles in the training set is decreased, the prediction errors of the combined model progressively rise, with the MAE reaching 0.662% and the RMSE reaching 0.766%. The combined BO-CNN-BiLSTM models are useful and demonstrate good prediction accuracy for various battery aging conditions.

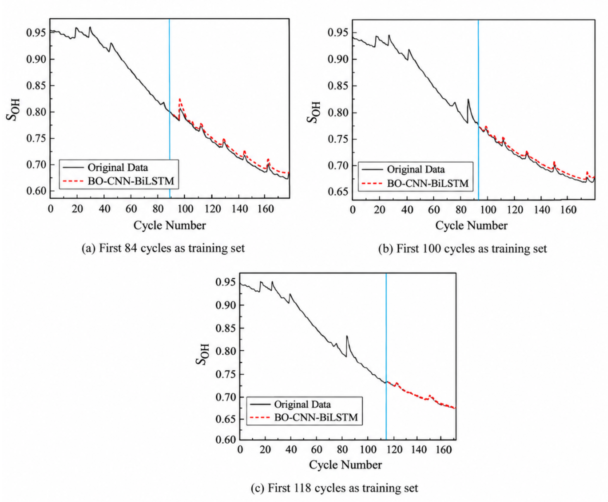

Fig. 3. Results of SOH prediction based on various training sets (B0005).

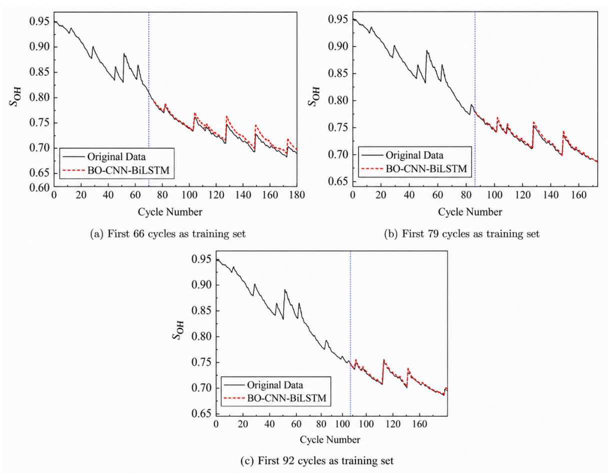

Fig. 4. Results of SOH prediction based on various training sets (B0018).

5. Conclusion

In this work, we present an approach to forecast the SOH of lithium batteries for new energy vehicles based on big data analysis, and we demonstrate experimentally its advantages in terms of accuracy and efficiency. Using the CNN and Bi-LSTM in combination with the BO algorithm, we successfully extracted and analyzed different battery health features from the publicly available NASA dataset. The experimental results show that our approach performs better than single neural network models and conventional techniques in terms of mean absolute error and root mean square error. This study therefore provides new theoretical and technical support for the safe functioning and health management of power batteries in future energy vehicles.

It also provides a practical road map and case study for the future creation of intelligent battery management systems. Future work will focus on extending the CNN-BiLSTM architecture for real-time SOH estimation within onboard battery management systems under dynamic operating conditions. In practical deployment, the Bayesian Optimization algorithm can be integrated offline within the battery management system to periodically update opti-

Table 4. SOH prediction errors of different models (B0005).

Error	CNN	LSTM	BiLSTM	CNN- BiLSTM	BO- CNN-BiLSTM
MAE/%	1.352	1.187	0.745	0.415	0.288
RMSE/%	1.521	1.278	0.856	0.487	0.365

Table 5. SOH prediction errors of different models (B0018).

Error	CNN	LSTM	BiLSTM	CNN- BiLSTM	BO- CNN-BiLSTM
MAE/%	1.526	0.857	0.724	0.556	0.289
RMSE/%	1.787	1.156	0.878	0.621	0.325

Table 6. Estimation errors of SOH for various training sets (B0005).

Number of training set iterations	MAE %	RMAE %
84	0.588	0.645
100	0.398	0.423
118	0.288	0.365

Table 7. Estimation errors of SOH for various training sets (B0005).

Number of training set iterations	MAE %	RMAE %
66	0.665	0.758
78	0.408	0.422
90	0.287	0.325

mal hyperparameters based on newly collected operational data. In practice, it can support BMS for accurate SOH estimation, with consideration of computational cost.

Declarations

Data Availability

Not applicable

Conflicts of Interest

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

Funding Statement

This work was supported by Basic Science (Natural Science) Research Project of Jiangsu Higher Education Institutions of China (21KJD580005), Qinglan Project of Jiangsu Province, The Sixth Jianghai Talents Project of Nantong City.

Author Contribution

Xiaojing Chu and **Chenhong Feng** (School of Automotive Engineering, Tianjin Vocational Institute) contributed to the conceptualization, methodology, investigation, and writing of the original draft.

Yongming Ding (Tianjin Shitong Network Technology Co., Ltd.) contributed to data curation, software development, validation, and review & editing of the manuscript.

All authors have read and approved the final version of the manuscript.

Ethical Approval

This article does not contain any studies involving human participants or animals performed by any of the authors.

Consent to Participate

Not applicable.

Consent to Publication

All authors have provided consent for publication of this manuscript.

References

[1] Z. He, X. Shen, Y. Sun, et al., (2021) *"State-of-health estimation based on real data of electric vehicles concerning user behavior"* **Journal of Energy Storage** ***41***: 102867. DOI: 10.1016/j.est.2021.102867.

[2] S. Hong, T. Yue, and H. Liu, (2022) *"Vehicle energy system active defense: A health assessment of lithium-ion batteries"* **International Journal of Intelligent Systems** ***37***(12): 10081–10099. DOI: 10.1002/int.22309.

[3] J. Zhao and A. F. Burke, (2022) *"Electric vehicle batteries: Status and perspectives of data-driven diagnosis and prognosis"* **Batteries** ***8***(10): 142. DOI: 10.3390/batteries8100142.

[4] Z. Wang, G. Feng, D. Zhen, et al., (2021) *"A review on online state of charge and state of health estimation for lithium-ion batteries in electric vehicles"* **Energy Reports** ***7***: 5141–5161. DOI: 10.1016/j.egyr.2021.08.113.

[5] Z. Xu, J. Wang, P. D. Lund, et al., (2021) *"Estimation and prediction of state of health of electric vehicle batteries using discrete incremental capacity analysis based on real driving data"* **Energy *225***: 120160. DOI: 10.1016/j.energy.2021.120160.

[6] Y. Toughzaoui, S. B. Toosi, H. Chaoui, et al., (2022) *"State of health estimation and remaining useful life assessment of lithium-ion batteries: A comparative study"* **Journal of Energy Storage *51***: 104520. DOI: 10.1016/j.est.2022.104520.

[7] P. Li, Z. Zhang, R. Grosu, et al., (2022) *"An end-to-end neural network framework for state-of-health estimation and remaining useful life prediction of electric vehicle lithium batteries"* **Renewable and Sustainable Energy Reviews *156***: 111843. DOI: 10.1016/j.rser.2021.111843.

[8] J. Hong, Z. Wang, W. Chen, et al., (2021) *"Online accurate state of health estimation for battery systems on real-world electric vehicles with variable driving conditions considered"* **Journal of Cleaner Production *294***: 125814. DOI: 10.1016/j.jclepro.2021.125814.

[9] Y. Zhang and Y. F. Li, (2022) *"Prognostics and health management of Lithium-ion battery using deep learning methods: A review"* **Renewable and Sustainable Energy Reviews *161***: 112282. DOI: 10.1016/j.rser.2022.112282.

[10] J. Tian, X. Liu, S. Li, et al., (2023) *"Lithium-ion battery health estimation with real-world data for electric vehicles"* **Energy *270***: 126855. DOI: 10.1016/j.energy.2023.126855.

[11] S. K. Assayed, C. S. Shieh, et al., (2025) *"Engineering intelligent healthcare systems: Understanding medical queries with AI and NLP"* **Journal of Engineering and Applied Science 72**(1): 212. DOI: 10.1186/s44147-025-00800-y.

[12] X. Li, K. Dai, Z. Wang, et al., (2020) *"Lithium-ion batteries fault diagnostic for electric vehicles using sample entropy analysis method"* **Journal of Energy Storage *27***: 101121. DOI: 10.1016/j.est.2019.101121.

[13] C. Zhang, S. Zhao, Z. Yang, et al., (2022) *"A reliable data-driven state-of-health estimation model for lithium-ion batteries in electric vehicles"* **Frontiers in Energy Research *10***: 1013800. DOI: 10.3389/fenrg.2022.1013800.

[14] X. Hu, Y. Che, X. Lin, et al., (2020) *"Health prognosis for electric vehicle battery packs: A data-driven approach"* **IEEE/ASME Transactions on Mechatronics *25***(6): 2622–2632. DOI: 10.1109/TMECH.2020.2986364.

[15] M. Elmahallawy, T. Elfouly, A. Alouani, et al., (2022) *"A comprehensive review of lithium-ion batteries modeling, and state of health and remaining useful lifetime prediction"* **IEEE Access *10***: 119040–119070. DOI: 10.1109/ACCESS.2022.3221137.

[16] C. Zhang, G. Shan, B. H. Roh, and J. Jiang, (2025) *"FM2 Learning: LLM-based federated multi-task multidomain learning for consumer electronics and IoT enhancement"* **IEEE Transactions on Consumer Electronics *71***(4): 11977–11988. DOI: 10.1109/TCE.2025.3603592.