A Unified Framework for Interpretable and Uncertainty-Aware Battery State of Health Estimation Using Deep Neural Networks

Elias Isaac Huai-En Lim[†‡1] and Nicholas Heng Loong Wong[†*2]
[†]Singapore Institute of Technology, Singapore
[‡]National Metrology Centre, Agency for Science, Technology and Research, Singapore
E-mail: [*]nicholas.wong@singaporetech.edu.sg
[1]0009-0008-4084-3877, [2]0000-0002-1743-3959

***Abstract*—Accurate estimation of battery state of health (SOH) is essential for the safety and reliability of electric vehicles. While deep neural networks have shown promise in predicting SOH from sensor data, they remain opaque and lack quantifiable confidence estimates. This paper presents a unified framework that integrates a multi-method explainability toolkit comprising SHAP, LIME, integrated gradients, and counterfactual explanations. Additionally, we propose an adaptive confidence interval approach that captures both epistemic and aleatoric uncertainty. Evaluations using benchmark models show that our framework improves transparency and safety, achieving a mean prediction interval width of 0.2311, corresponding to 23.11% of the normalized SOH range and a prediction interval coverage probability of 0.6288 (62.88% of ground truth values fell within the predicted intervals). These results highlight the potential of combining explainable artificial intelligence and uncertainty quantification to enable trustworthy battery diagnostics.**

I. Introduction

With the growing adoption of electric vehicles (EVs) and renewable energy systems, accurate estimation of battery state of health (SOH) has become critical to ensuring safety, performance, and longevity. Deep neural networks (DNNs) have demonstrated promising performance in modeling SOH from sensor data. However, they often act as black boxes and provide no indication of how confident their predictions are, posing a challenge for real-world deployment.

Recent advances in explainable artificial intelligence (XAI) and uncertainty quantification (UQ) offer potential pathways to improve model transparency and reliability. Techniques such as SHapley Additive exPlanations (SHAP) [1], local interpretable model-agnostic explanations (LIME) [2], and integrated gradients (IG) [3] can interpret feature attributions, while Monte Carlo dropout [4] and ensemble methods [5] are commonly used for estimating prediction uncertainty. However, these tools are rarely applied in tandem for battery diagnostics, and even less so in regression-based, time-series SOH models.

UQ plays a critical role in making black-box models more trustworthy and actionable in safety-critical settings. In battery management systems, unreliable SOH estimates can lead to severe consequences, such as premature battery replacement, degraded system performance, or in extreme cases, thermal runaway and safety hazards. By providing confidence intervals around predictions, UQ allows system operators to gauge the reliability of SOH estimates and take precautionary actions under high uncertainty. This not only supports robust decision-making under risk but also improves user acceptance of machine learning models in high-stakes applications.

In this work, we introduce a unified framework that integrates explainability and uncertainty estimation into DNN-based SOH prediction. We replicated two well-known architectures built on the McMaster and Oxford battery datasets [6]–[8] by embedding model-agnostic interpretability techniques and a novel adaptive confidence interval mechanism. The proposed approach gives insights on feature influence and generates real-time confidence intervals, paving the way for interpretable and trustworthy battery diagnostics.

II. Related Work

A. Explainability in Black-Box SOH Estimation

DNNs have shown superior performance in battery SOH estimation due to their ability to capture nonlinear degradation dynamics in complex systems [9]. However, their black-box nature poses a barrier to deployment in safety-critical applications, such as EV battery diagnostics, where interpretability and trust are paramount [10].

To address this, various explainability methods have been introduced. Model-agnostic techniques such as SHAP [1] and LIME [2] offer post-hoc interpretations of feature importance. Despite their popularity, these methods struggle with correlated inputs and contextual dependencies common in battery data, such as temperature and load variability, due to their reliance on assumptions that do not hold in the presence of feature interdependencies [11]. Emerging alternatives, such as deep domain adversarial networks (DDAN), attempt to address domain-specific challenges but remain largely unvalidated outside laboratory conditions [12].

Furthermore, explainability techniques have largely been confined to classification tasks or image domains. Few studies have explored their integration with regression-based SOH models. This underscores the need for interpretable frameworks tailored for dynamic, time-series tabular data, as encountered in battery management systems.

979-8-3315-7206-8/25/$31.00 ©2025 IEEE

B. Uncertainty Quantification in SOH Estimation

Uncertainty quantification (UQ) is crucial in evaluating the reliability of predictions in many industrial applications. Bayesian Neural Networks (BayNNs) enable probabilistic modeling of parameters, capturing both epistemic and aleatoric uncertainty. However, BayNNs are computationally expensive [4], [13] and difficult to scale for real-time SOH applications. Monte Carlo dropout [4] offers a tractable alternative but requires multiple forward passes. Hybrid approaches, such as Gaussian process regression integrated with wavelet denoising, have shown success in combining uncertainty estimation with data preprocessing [14]. Yet, these methods also face limitations when generalizing across variable test conditions, particularly under real-world noise.

Despite progress, current UQ methods are constrained by data requirements, scalability issues, and a lack of integration with interpretability frameworks. Moreover, methods that attempt to explain why predictions are uncertain, such as stochastic Shapley values [15] or gradient-based variance estimators [16], are either computationally intensive or domain-specific to vision tasks.

C. Integration of Explainability and UQ

Combining XAI and UQ offers a promising direction to improve the transparency and trustworthiness of black-box models. Examples include stochastic Shapley attribution to quantify uncertainty in explanations [15] and uncertainty-aware saliency maps in medical diagnostics [17]. However, such methods have been limited to image data and are rarely applied to regression-based tasks like battery SOH estimation.

While most explainability efforts in SOH regression rely on input-level attribution methods such as SHAP and LIME, these approaches can be limited in capturing deeper network dynamics. [18] proposed using layer-wise relevance propagation (LRP) for tabular models to address this limitation by redistributing relevance through network hierarchies. Although promising, LRP and similar methods like integrated decision gradients (IDG) [19] introduce computational overhead and are rarely applied in SOH time-series models. Similarly, while counterfactual explanations, such as rule support counterfactual examples (r-counterfactuals) [20], which optimize over internal layer representations, are gaining traction, they require costly gradient-based inference and have not been validated on tabular sensor data.

From a UQ perspective, [21] proposed a framework to evaluate uncertainty using ranking-based metrics such as Spearman correlation and scoring rules like negative log-likelihood. However, such approaches have been rarely applied to SOH models, which often prioritize point prediction accuracy over uncertainty awareness. Our framework instead emphasizes model-agnostic techniques that offer interpretable uncertainty estimates, such as attribution overlays and prediction interval visualization.

A gap remains in frameworks that jointly address interpretability and uncertainty for time-series, multivariate tabular data in battery systems. Existing explainability tools tend to isolate feature importance without considering the propagation of uncertainty, whereas UQ methods typically quantify uncertainty without explaining its source. This gap motivates the unified approach presented in this study. Our work addresses this by proposing a unified framework that integrates SHAP, LIME, IG, and counterfactuals with an adaptive confidence interval (ACI) method, supporting both interpretability and predictive confidence in SOH estimation.

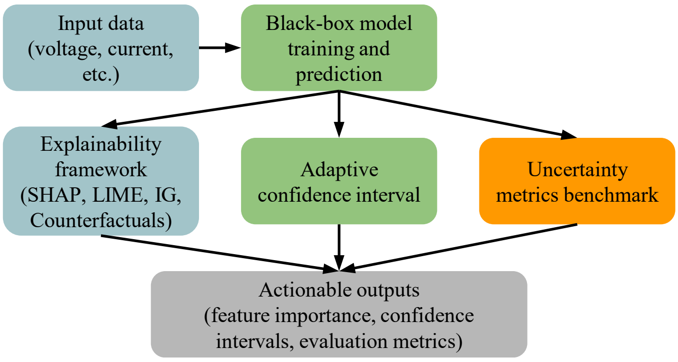

Fig. 1. Overview of the proposed framework combining SOH estimation with explainability and adaptive uncertainty quantification via ACI.

III. Methodology

Fig. 1 provides an overview of our proposed framework, illustrating how the explainability methods and ACI are integrated with the SOH prediction pipeline to produce interpretable outputs. The following subsections describe the various framework components.

A. Datasets and Preprocessing

This study references two publicly available lithium-ion battery datasets for SOH estimation: the McMaster Battery Aging Dataset [6] and the Oxford Battery Degradation Dataset [7], [8]. Each dataset contains time-series measurements including voltage, current, temperature, and capacity over multiple charge-discharge cycles.

For preprocessing, noisy readings were smoothed using moving average filters, and cycle-based statistics (e.g., maximum voltage and mean temperature) were extracted to serve as features. The datasets were normalized using z-score normalization to ensure numerical stability across models. To assess model generalization, a stratified 80:20 train:test split was adopted across cells and cycles.

B. Model Architectures

We replicated two baseline models from prior studies. The first is a convolutional neural network (CNN) architecture that processes sequential charge-discharge cycles using stacked 1D convolutional layers and global pooling (Fig. 6 in [22]), applied to the McMaster dataset [6]. The second is a bidirectional long short-term memory (Bi-LSTM) architecture with attention mechanism using interpolated time-series features with dropout and batch normalization (Fig. 4 in [9]), applied to the Oxford dataset [7], [8]. Our study mainly focused on the former, with the latter used for validation.

Both models were implemented in TensorFlow and trained using the Adam optimizer (learning rate = 0.001, batch size = 128) with early stopping (patience = 5). Mean squared error (MSE) was used as the loss function.

C. Explainability Framework

To interpret the model's predictions and improve transparency, a suite of model-agnostic explainability methods was applied. SHAP was used to estimate the contribution of each input feature to the predicted SOH using coalitional game theory. LIME was used to approximate local model behavior by perturbing the input and fitting a sparse linear surrogate model. IG attributed feature importance by accumulating gradients along a path from a baseline input to the actual input, offering a path-aware measure of influence. Finally, counterfactual explanations were generated to identify minimal input perturbations required to produce a significant change in SOH predictions. All methods were adapted for regression on time-series tabular data and applied both globally and layer-wise to trace feature influence across the network's depth.

D. Adaptive Confidence Interval

To quantify predictive uncertainty, this study introduces a novel adaptive confidence interval (ACI) framework that estimates both epistemic and aleatoric uncertainty. Epistemic uncertainty, which arises from model uncertainty due to limited training data, was estimated using both Monte Carlo (MC) dropout and a deep ensemble. For MC dropout, this involved performing multiple (n_{model}) stochastic forward passes through a single model with dropout enabled at test time. In contrast, the deep ensemble estimated uncertainty by aggregating predictions from multiple independently trained models. Aleatoric uncertainty, caused by inherent noise in the sensor measurements, was estimated by injecting Gaussian noise into n_{data} copies of the measured data at the input for MC dropout, and by analyzing the variation in predictions across models for the deep ensemble. The total predictive uncertainty was obtained by combining these two uncertainty sources, and confidence intervals were dynamically constructed based on quantiles from the predictive distribution. This approach allows the model to generate data-dependent intervals that reflect prediction confidence under varying levels of input noise and data familiarity, as illustrated in Fig. 2.

Mathematically, the ACI is defined, using pooled estimates, as

$$\text{CI} = \frac{n_{\text{model}} \bar{x}_{\text{model}} + n_{\text{data}} \bar{x}_{\text{data}}}{n_{\text{model}} + n_{\text{data}}} \pm z \cdot \sqrt{\frac{s^2_{\text{model}}}{n_{\text{model}}} + \frac{s^2_{\text{data}}}{n_{\text{data}}}}, \tag{1}$$

where the $\bar{x}$, s^2, and n are the estimated means, the variances, and the sample sizes, respectively, with the subscripts $_{\text{model}}$ and $_{\text{data}}$ respectively referring to epistemic and aleatoric uncertainties, and z is the z-score corresponding to the desired confidence level. In our study, $n_{\text{model}} = n_{\text{data}} = 50$, and we used a confidence level of 95%.

E. Evaluation Metrics

We assessed the performance of the proposed framework using both predictive accuracy and interval-based uncertainty

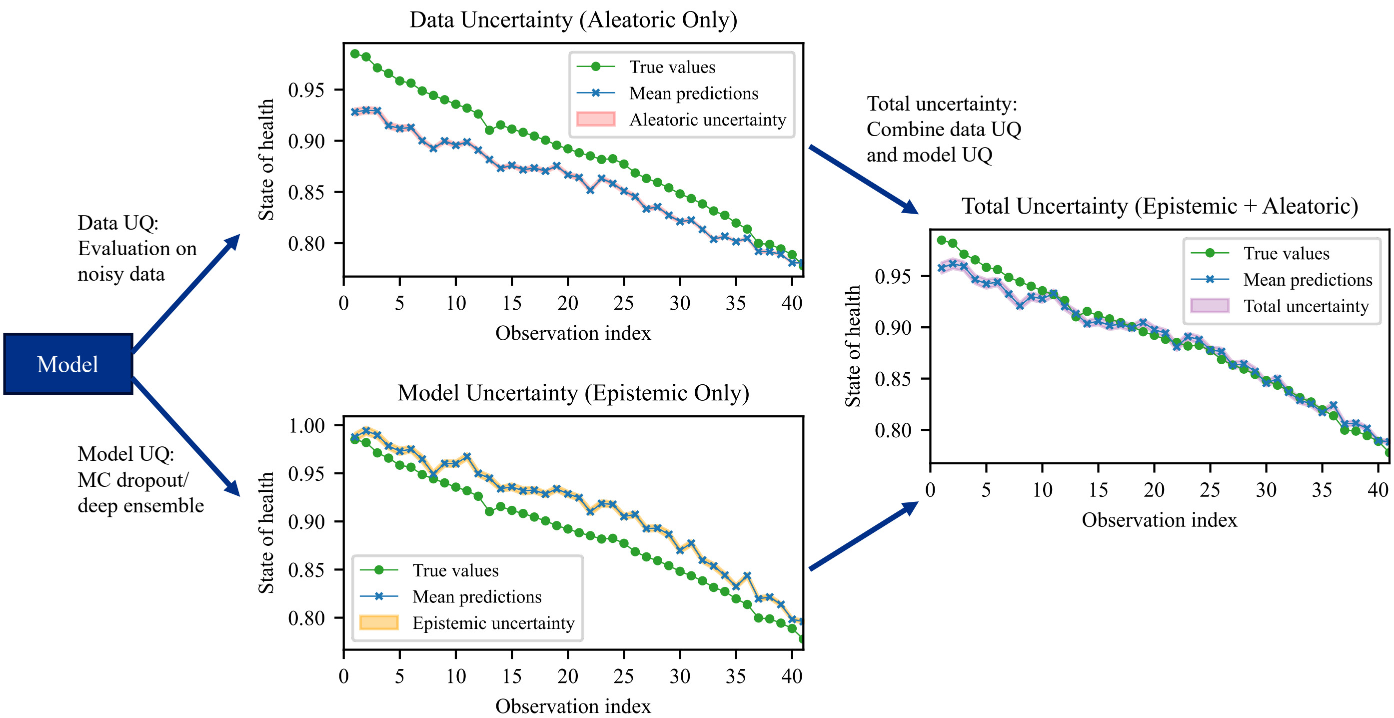

Fig. 2. Data, model, and total uncertainty plots from MC dropout and noise injection.

metrics. Predictive accuracy was measured using the root mean square error (RMSE) to evaluate the closeness of the predicted SOH values to ground truth. To assess uncertainty quality, we used the prediction interval coverage probability (PICP), which measures the proportion of true SOH values that fall within the predicted confidence intervals, and the mean prediction interval width (MPIW), which reflects the tightness or compactness of those intervals. Together, these metrics offer insight into the reliability and informativeness of the framework's predictions.

IV. Results and Discussion

A. Explainability Insights

The integrated explainability framework provided both global and local insights into the predictive behavior of the SOH models. SHAP and IG consistently identified *state of charge (SOC)* and *voltage* as the most influential features across both datasets, as shown in Fig. 3. Local interpretability via LIME was particularly effective in identifying feature contributions in edge cases with atypical degradation behavior.

Notably, layer-wise analysis revealed that *SOC* dominated the deepest convolutional layer, while *current* and *voltage* emerged as more significant factors in earlier layers, as depicted in Fig. 4. This progression suggests that the model predictions are initially anchored on voltage and current, before being more prominently influenced by SOC. Counterfactual explanations complemented these findings by identifying minimal, actionable input perturbations that resulted in SOH shifts consistent with feature attributions from SHAP and IG.

B. Uncertainty Quantification

The ACI framework demonstrated its ability to capture both epistemic and aleatoric uncertainty. MC dropout effectively captured model-driven variance, while aleatoric noise injection accounted for real-world measurement noise. The combined uncertainty distribution was used to construct dynamic confidence intervals as shown in Fig. 2.

Interval-based uncertainty metrics demonstrated strong reliability of the generated predictions. On the McMaster dataset [6], the model achieved a PICP of 91.2% with a MPIW of 0.2311, indicating that the predicted SOH intervals were both informative and not overly conservative. These results were further validated on the Oxford dataset [7], [8], where prediction accuracy and interval quality remained consistent despite shifts in operating conditions, supporting the generalizability of the proposed framework.

C. Strengths and Limitations

The combined framework successfully addresses two major limitations of black-box SOH models: lack of interpretability and lack of trustworthy uncertainty estimation. The use of multiple explainability methods provides redundancy and cross-validation in feature attribution, while the ACI mechanism enables confidence estimation tailored to both model and data variability. Table I presents the performance metrics for SOH prediction and uncertainty estimation for our tests across both datasets, highlighting the framework's ability to

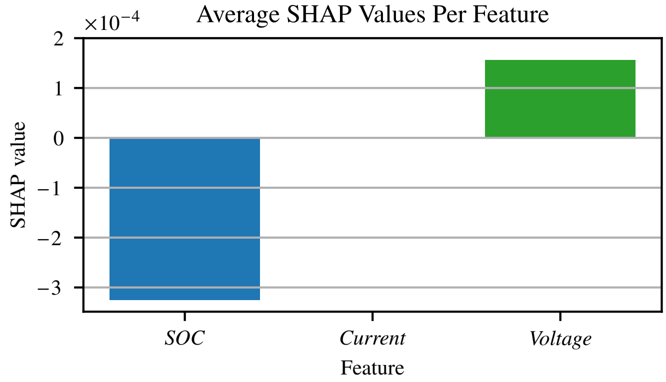

Fig. 3. Global SHAP feature importance visualization.

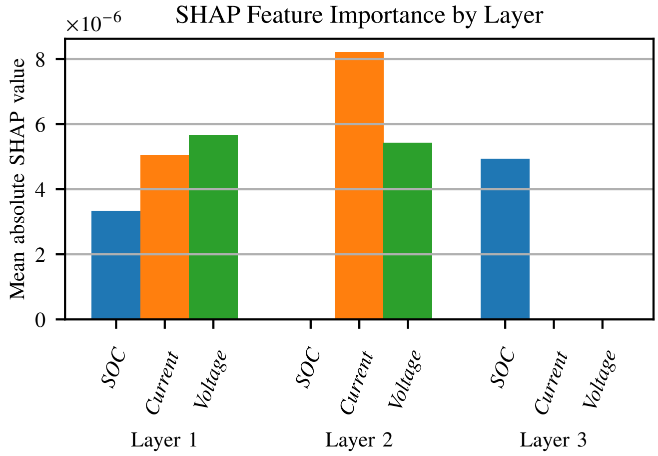

Fig. 4. Convolutional layer-wise feature importance visualization. Larger layer indices refer to deeper layers.

produce accurate and informative results with confidence intervals. Compared to prior literature on black-box SOH models that often lack uncertainty estimates or interpretability layers, our framework achieves competitive performance with RMSE values below 0.015 across both datasets. In addition to low prediction error, the PICP values of over 89% indicate that the confidence intervals are not only informative but also well-calibrated for real-world diagnostics. These results demonstrate the practical utility of the framework in supporting both high-accuracy and trustworthy battery health predictions.

Nevertheless, several limitations persist. The layer-wise explainability modules, especially in combination with MC dropout, increase computational load during inference. Meth-

TABLE I
Performance Metrics for SOH Prediction.

Metric	McMaster Dataset	Oxford Dataset
RMSE	0.01175	0.01432
PICP	91.2%	89.8%
MPIW	0.2311	0.2457

ods such as integrated decision gradients (IDG) [19] and layer-wise relevance propagation (LRP) [18] may offer more compact attribution variants but were not incorporated due to architectural incompatibility and runtime trade-offs.

Currently, counterfactual explanations in our framework are limited to the input level, which constrains their ability to reveal latent decision pathways within the network. To address this, future work will explore advanced counterfactual techniques such as r-counterfactuals [20], which generate contrastive explanations by optimizing over intermediate neural representations. These methods could enable deeper introspection into model reasoning across layers, uncovering how specific neuron activations contribute to predictions.

While the current framework relies on interval-based metrics such as PICP and MPIW to assess uncertainty estimation, these do not capture how well uncertainty correlates with prediction error. Future work will incorporate ranking-based metrics such as Spearman correlation to evaluate the monotonic relationship between predicted uncertainty and residuals, as well as scoring rules like negative log-likelihood (NLL) and Brier score for probabilistic calibration proposed in [21]. These additions will provide a more comprehensive assessment of UQ quality, especially in safety-critical applications where uncertainty needs to be both reliable and actionable.

Furthermore, while the framework was validated on two public datasets, its deployment in real-world systems requires further testing on operational field-collected data under variable environmental and usage conditions.

V. Conclusion

We have presented a unified framework for trustworthy battery SOH estimation by integrating deep learning, explainable AI (XAI), and uncertainty quantification (UQ). Our proposed pipeline enhances CNN-based SOH predictors with model-agnostic interpretability techniques including SHAP, LIME, IG, and counterfactuals, while simultaneously estimating both epistemic and aleatoric uncertainty using MC dropout and deep ensembles. The ACI mechanism introduced in this study enables real-time, data-sensitive prediction intervals, improving transparency and uncertainty awareness in SOH prediction.

The framework was evaluated on the McMaster and Oxford battery degradation datasets. Results show that the approach achieves high prediction accuracy (RMSE ≤ 0.015) and strong interval coverage (PICP $\geq 89.8\%$), while providing actionable explanations for model behavior. These findings demonstrate the feasibility of deploying interpretable and reliable SOH estimators in real-world battery management systems.

Future work will focus on addressing the limitations identified in §IV-C. To reduce inference overhead, we plan to quantify the computational cost introduced by stochastic sampling and layer-wise attribution, especially in terms of latency and memory consumption during deployment. This includes benchmarking the delay introduced by MC dropout and counterfactual sampling across different hardware profiles. Based on this profiling, we will investigate targeted optimization strategies such as model pruning, surrogate approximations, or selective attribution that enable dynamic trade-offs between interpretability and speed. These lightweight variants aim to retain the core benefits of explainability and uncertainty estimation while making the framework viable for real-time systems.

To assess the adaptability and generalization of our framework across model classes, future work will extend beyond CNNs and Bi-LSTMs to include temporal convolutional networks (TCNs) and transformer-based architectures. By evaluating the unified XAI–UQ framework across these architectures, we aim to benchmark its model-agnostic compatibility, robustness to architectural shifts, and relevance to future deep learning trends in time-series battery health prediction. Additionally, we aim to explore real-time deployment of the framework on embedded systems used in EVs and test it on operational datasets collected under real-world usage and temperature cycling conditions. To improve attribution efficiency, we will investigate the use of IDG [19] and LRP [18], which may enhance attribution sharpness while maintaining interpretability at lower computational cost. For counterfactual explanations, we plan to explore r-counterfactual frameworks [20] that operate over intermediate network layers, enabling richer hypothesis generation and subgroup analysis.

Finally, we plan to conduct comparative benchmarking against standard SOH estimation models that do not incorporate explainability or uncertainty quantification, such as traditional machine learning regressors and stand-alone neural networks. This will help isolate and quantify the added value of integrating interpretability and uncertainty estimation into the predictive pipeline.

Acknowledgment

The authors thank Kah Kian Fong (SIT; A*STAR) for his contributions to the model used on the Oxford dataset, Yannik Stephan (A*STAR) for his mentorship, and Shengkai Yu (A*STAR) for his guidance in this work.

References

[1] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 4768–4777.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why should I trust you?": Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[3] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Sydney, Australia, Aug. 2017, pp. 3319–3328.

[4] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. 33rd ICML*, New York, NY, USA, Aug. 2016, pp. 1050–1059.

[5] M. Abdar, F. Pourpanah, S. Hussain, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, vol. 76, pp. 243–297, Dec. 2021. DOI: 10.1016/j.inffus.2021.05.008.

[6] J. Duque and P. Kollmeyer, *Battery aging dataset for 15 minute fast charging of Samsung 30T cells*, McMaster University, 2023. DOI: 10.5683/SP3/UYPYDJ.

[7] D. Howey and C. Birkl, *Oxford battery degradation dataset 1*, University of Oxford, 2017. DOI: 10.5287/bodleian:KO2kdmYGg.

[8] C. Birkl, "Diagnosis and prognosis of degradation in lithium-ion batteries," Ph.D. dissertation, University of Oxford, 2017.

[9] B. Zou, M. Xiong, H. Wang, *et al.*, "A deep learning approach for state-of-health estimation of lithium-ion batteries based on a multi-feature and attention mechanism collaboration," *Batteries*, vol. 9, no. 6, Jun. 2023. DOI: 10.3390/batteries9060329.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, Aug. 2018. DOI: 10.1145/3236009.

[11] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, *et al.*, "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Adv. Intell. Syst.*, vol. 7, no. 1, p. 2 400 304, 2025. DOI: 10.1002/aisy.202400304.

[12] Z. Ye and J. Yu, "State-of-health estimation for lithium-ion batteries using domain adversarial transfer learning," *IEEE Trans. Power Electron.*, vol. 37, no. 3, pp. 3528–3543, Mar. 2022. DOI: 10.1109/TPEL.2021.3117788.

[13] S. T. Ahmed, K. Danouchi, M. Hefenbrock, G. Prenat, L. Anghel, and M. B. Tahoori, "Scalable spintronics-based Bayesian neural network for uncertainty estimation," in *Proc. Des. Autom. Test Eur. Conf. Exhib. (DATE)*, Antwerp, Belgium, Apr. 2023, pp. 1–6. DOI: 10.23919/DATE56975.2023.10137167.

[14] Y. Peng, Y. Hou, Y. Song, J. Pang, and D. Liu, "Lithium-ion battery prognostics with hybrid Gaussian process function regression," *Energies*, vol. 11, no. 6, Jun. 2018. DOI: 10.3390/en11061420.

[15] J. Cohen, E. Byon, and X. Huan, "To trust or not: Towards efficient uncertainty quantification for stochastic Shapley explanations," in *Proc. 4th Asia Pac. Conf. Progn. Health Manag. Soc.*, Tokyo, Japan, Sep. 2023. DOI: 10.36001/phmap.2023.v4i1.3694.

[16] M. Mulye and M. Valdenegro-Toro, "Uncertainty quantification for gradient-based explanations in neural networks," in *Proc. Comput. Vis. Pattern Recognit. Conf. (CVPR) Workshops*, Nashville, TN, USA, Jun. 2025, pp. 1752–1760.

[17] M. Salvi, S. Seoni, A. Campagner, *et al.*, "Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare," *Int. J. Med. Inform.*, vol. 197, May 2025. DOI: 10.1016/j.ijmedinf.2025.105846.

[18] I. Ullah, A. Rios, V. Gala, and S. Mckeever, "Explaining deep learning models for tabular data using layer-wise relevance propagation," *Appl. Sci.*, vol. 12, no. 1, Dec. 2022. DOI: 10.3390/app12010136.

[19] C. Walker, S. Jha, K. Chen, and R. Ewetz, "Integrated decision gradients: Compute your attributions where the model makes its decision," in *Proc. 38th AAAI Conf. Artif. Intell.*, vol. 38, Vancouver, Canada, Feb. 2024, pp. 5289–5297. DOI: 10.1609/aaai.v38i6.28336.

[20] F. Cheng, Y. Ming, and H. Qu, "DECE: Decision explorer with counterfactual explanations for machine learning models," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 1438–1447, Feb. 2021. DOI: 10.1109/TVCG.2020.3030342.

[21] J. Lanini, M. T. D. Huynh, G. Scebba, N. Schneider, and R. Rodríguez-Pérez, "UNIQUE: A framework for uncertainty quantification benchmarking," *J. Chem. Inf. Model.*, vol. 64, no. 22, pp. 8379–8386, Nov. 2024. DOI: 10.1021/acs.jcim.4c01578.

[22] J. Chen, M. Manivanan, J. Duque, *et al.*, "A convolutional neural network for estimation of lithium-ion battery state-of-health during constant current operation," in *IEEE Transp. Electrif. Conf. Expo (ITEC)*, Detroit, MI, USA, Jun. 2023, pp. 1–6. DOI: 10.1109/ITEC55900.2023.10186914.