RESEARCH ARTICLE

Ensemble Learning for Precise State-of-Charge Estimation in Electric Vehicles Lithium-Ion Batteries Considering Uncertainty

AYA HARAZ[ID]**1, KHALID ABUALSAUD**[ID]**2, (Senior Member, IEEE), AND AHMED M. MASSOUD**[ID]**1, (Senior Member, IEEE)**

[1]Department of Electrical Engineering, Qatar University, Doha, Qatar
[2]Department of Computer Science and Engineering, Qatar University, Doha, Qatar

Corresponding author: Aya Haraz (aya.haraz@qu.edu.qa)

ABSTRACT Accurate state-of-charge (SoC) estimation is crucial for enhancing the performance, longevity, safety, and reliability of lithium-ion batteries (LiBs) in electric vehicles (EVs). This study presents a comprehensive machine learning (ML)-based approach for SoC estimation of EV LiBs, addressing the challenges of model reliability, uncertainty, and real-world data variability. To ensure the model's robustness and generalizability, preprocessing techniques, including normalization and scaling, were employed alongside rigorous cross-validation methods. A well-structured ML pipeline was developed to integrate these processes, optimizing the entire model development cycle for efficiency and practical implementation. In the ML pipeline, we utilized Extra Trees Regressor (ETR) and Light Gradient Boosting Machine (LightGBM) and proposed an ensemble model, combining the strengths of ETR and LightGBM, namely ETR-GBM. We benchmarked the model's performance against other ML models, such as CatBoost and Random Forest (RF). Under uncertain conditions, the proposed model emphasized its reliability and robustness, and its conclusions underscored the efficacy of the SoC estimation approach. The ETR-GBM consistently outperforms the individual models (ETR, LightGBM, XGBoost, CatBoost, Support Vector Regression (SVR), Random Forest (RF), and Bayesian) when noise is added to the training dataset. With a noise standard deviation of 0.1, the ETR-GBM demonstrated superior performance, achieving a Root Mean Square Error (RMSE) of 0.41%, surpassing the individual models, which exhibited RMSE values ranging from 0.85% to 0.91%.

INDEX TERMS Machine learning, electric vehicles, state-of-charge, lithium-ion batteries, ensemble model, Extra Tree Regressor, light gradient boosting machine, uncertainty quantification.

NOMENCLATURE

BMS	Battery Management System.
CC	Constant Current.
CV	Constant Voltage.
ETR	Extra Tree Regressor.
EVs	Electric Vehicles.
LightGBM	Light Gradient Boosting Machine.
OCV	Open Circuit Voltage.
SoC	State-of-Charge.

I. INTRODUCTION

Air pollution is undeniably a significant global health issue, with studies indicating it results in a substantial number of premature deaths each year. Research estimates that exposure to fine particulate matter is responsible for around 5.3 to 5.4 million deaths worldwide [1]. To address this, countries have implemented policies to ban petrol and diesel-powered vehicles by 2030, accelerating the shift toward electric vehicles (EVs) [2], [3]. EVs have experienced substantial growth over the past few decades, emerging as a prominent mode of transportation [4]. Unlike gas-powered vehicles, EVs and their battery systems are relatively new and continuously

evolving [5]. A primary obstacle to the widespread adoption of EVs is "range anxiety", i.e., the fear that EV's battery will deplete before reaching a charging station [6]. Lithium-ion batteries (LiBs), the most effective energy storage devices for delivering power, are integral to EVs [7]. LiBs maintain a safe, efficient, and stable EV operational state due to their high energy density and a relatively long-life cycle [8]. Accurate State-of-Charge (SoC) estimation is vital for the reliable operation of LiBs, directly impacting the EV's remaining driving range and the battery management system's (BMS) efficiency [9]. SoC estimation is inherently challenging due to the battery's complex and dynamic behavior under varying operating conditions. This is in addition to the unavailability of required data or uncertainty [10] in the relevant available data.

A. RELATED WORKS FOR UNCERTAINTY IN SOC ESTIMATION

Uncertainty quantification (UQ) methods are widely employed to tackle complex real-world challenges across various fields of science and engineering [11], [12], [13]. There are two main types of uncertainty: aleatoric and epistemic. Aleatoric uncertainty, or data uncertainty, is the inherent unpredictability in the data. This type of uncertainty cannot be reduced because it is a fundamental aspect of data distribution. On the other hand, epistemic uncertainty, or knowledge uncertainty, arises from insufficient knowledge or understanding [11]. The available data is often incomplete, noisy, inconsistent, or has multiple patterns [14].

The complexity of LiBs—affected by factors such as parasitic side reactions and harsh environmental conditions—poses a significant challenge. This complexity underscores the need for diverse datasets that capture varying operating conditions, temperatures, aging states, and material differences. Fast charging under complex loading conditions—such as variable current levels and extreme operating temperatures—introduces multiple sources of uncertainty [15]. Collecting such comprehensive datasets is labor-intensive, time-consuming, and logistically challenging. The limited availability of real-world EV data further complicates the issue, as most studies rely on laboratory-generated data, which may not accurately reflect the variability and unpredictability of real-world conditions. Consequently, machine learning (ML) models trained on lab data may struggle to perform effectively in real-world EV applications, especially under extreme conditions. Several uncertainties should be quantified and addressed in ML at each step of the process, including (i) the selection and collection of training data, (ii) the completeness and accuracy of the training data, (iii) a comprehensive understanding of the deep learning (or traditional ML) model, including its performance boundaries and limitations, and (iv) uncertainties related to the model's performance when applied to real-world operational data [16]. Furthermore, data-driven approaches, particularly those akin to deep learning, encounter at least four interrelated challenges: (i) a lack of theoretical underpinnings, (ii) the absence of causal models, (iii) vulnerability to imperfect data, and (iv) high computational costs [11]. Approaches such as data augmentation, transfer learning, or synthetic data generation could help address these challenges.

Several studies have extensively explored the impact of data uncertainty on batteries model accuracy. They offered limited insights into the dynamic updating of identified parameters or empirical relationships, which hindered their ability to achieve the required accuracy for real-time estimation [17]. Ref. [18] presented a method to enhance the accuracy of SoC estimation in batteries by utilizing a robust Extended Kalman Filter (EKF) approach. The goal is to improve the reliability of SoC estimation, particularly with measurement errors, noise, and model uncertainties. Ref. [19] proposed an approach for battery prognostics, aiming to predict the SoH and remaining useful life (RUL) of batteries. It combines a Relevance Vector Machine (RVM) with wavelet denoising to reduce uncertainty and extract trends and uses a mean entropy-based method to optimize time series reconstruction. Ref. [17] proposed method for online SoH estimation of LiBs using a second-order RC equivalent circuit model, combined with EKF and recursive least squares, to dynamically link SoC with SoH. Ref. [20] presented a method using ML to predict the performance and aging of LiFePO4 batteries in EVs. By analyzing data from 420 cells and 9 battery packs and applying noise reduction and feature engineering, the study achieves highly accurate capacity estimates and battery life estimations. Ref. [21] proposed a data-driven approach for predicting future capacities and RUL of LiBs. It combines Long Short-Term Memory (LSTM) and Gaussian Process Regression (GPR) models to accurately predict battery capacity while simultaneously managing uncertainty, particularly addressing fluctuations due to capacity regenerations. Ref. [22] proposed a Bayesian Neural Network (BNN)-based method for predicting the knee capacity of LiBs, considering uncertainty quantification. Ref. [23] proposed a LiB degradation prediction model with uncertainty quantification. It combines a 1D Convolutional Neural Network (1dCNN) with Bi-LSTM and quantile regression to predict battery capacity at different stages. At the same time, kernel density estimation helps manage prediction uncertainty and optimize maintenance decisions. Ref. [24] proposed a highly accurate RUL prediction model that combines deep learning with nonstationary Gaussian process regression (NSGPR). It tackled key challenges like high-dimensional data, noise, and time-dependency in system degradation while providing reliable uncertainty quantification for more precise predictions.

B. RELATED WORK FOR SOC ESTIMATION

The BMS plays a crucial role in managing EV LiBs by regulating their temperature, monitoring voltage and current, detecting faults, and estimating remaining energy to enhance battery life and performance while ensuring safety [25], [26].

Despite advancements in BMS technology, predicting the remaining mileage range remains challenging due to battery degradation and instability in various conditions [27]. Recent scholarly research has explored various methodologies for estimating the SoC of EV batteries. These methodologies can be categorized into four principal groups: direct methods, model-based methods, learning algorithms, and hybrid methods, as shown in FIGURE 1. Direct methods include techniques such as coulomb counting and open-circuit voltage measurement [28], which are basic yet effective approaches for SoC estimation. Model-based methods [29] utilize Kalman filters, particle filters, and recursive least squares for predictive SoC analysis, relying on mathematical models of the battery's behavior. Learning algorithms employ data-driven approaches, leveraging advanced computational models like neural networks [30] and fuzzy logic [31], [32] to learn from historical data and predict the SoC. Hybrid methods [33] combine these approaches, integrating the estimation accuracy of model-based methods with the adaptive learning capabilities of neural networks to enhance the accuracy and reliability of SoC estimation. Exploring how different models perform under varying noise conditions is key to developing robust and reliable SoC estimation systems for EVs. In EVs, the collected data used for SoC estimation is often affected by various noise sources, making it crucial for models to maintain high accuracy in these conditions.
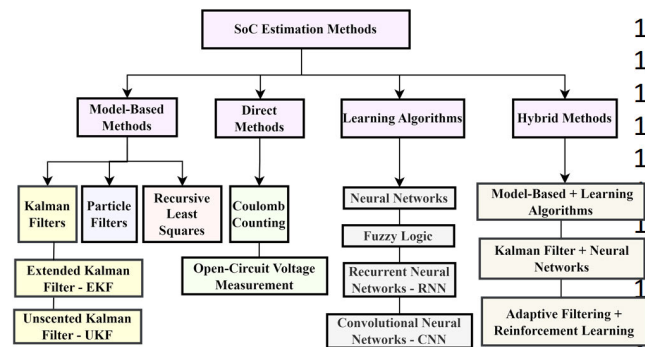

FIGURE 1. State of Charge (SoC) Estimation methods: Direct methods, Model-Based methods, Learning algorithms, and Hybrid methods.

Recent research [34], [35], [36], [37], [38], [39], [40], [41], [42] on SoC estimation methods for EV batteries reveals several gaps and advancements. Learning algorithms often lack testing under diverse conditions and have high computational needs. Validation in real-time conditions and enhanced scalability and adaptability are key focus areas. Model-based methods face integration and real-time processing challenges, resolved by creating real-time pipelines. Hybrid methods improve efficiency with advanced feature engineering and algorithms' parameter tuning. Direct methods enhance robustness through cross-validation, simplifying implementation, improving scalability, and long-term validation. TABLE 1 shows the most related research papers, their gaps, and the differences compared to the proposed work.

TABLE 1. State-of-Charge (SOC) Estimation methods: Existing gaps and main contributions.

Ref.	SoC Category Method	Existing Gap	Contribution
[34]	Learning Algorithms	Limited testing across diverse conditions	Using ML for Uncertainty conditions
[35]	Learning Algorithms	Complexity and high computational resource needs	Develop a real-time SoC estimation pipeline
[36]	Learning Algorithms	Validation in real-time and varying conditions	Use cross-validation and real-time implementation
[37]	Learning Algorithms	Scalability and adaptability	Feature engineering and data preprocessing
[38]	Model-Based Methods	Integration and real-time processing	Real-time SoC estimation pipeline
[39]	Hybrid Methods	Real-world data and computational efficiency	Feature engineering and robust ML techniques
[40]	Direct Methods	Complexity and computational resource needs	Real-time SoC estimation pipeline
[41]	Direct Methods	Scalability and integration	Real-time SoC estimation pipeline
[42]	Direct Methods	Adaptability and long-term validation	Using various types of ML to ensure adaptability

C. POSITIONING THE PROVIDED CONTRIBUTION FOR SOC ESTIMATION

Although previous works introduced several advanced SoC and SoH estimation methods, certain limitations remain. First, a study that utilized the GB-SVR method to capture non-linear dynamics in LiBs, offering improved computational efficiency and generalizability, has been presented in [43]. Another combined Extreme Learning Machine (ELM) and Random Vector Functional Link (RVFL) networks in a hybrid ensemble for more accurate SoH predictions, alongside a Nonlinear Autoregressive model and Bootstrap-based uncertainty management, has been explored in [44]. Their methods often encounter limitations, such as reliance on complex uncertainty management techniques. In this work, we focus on proposing a robust framework to simplify uncertainty management while enhancing the accuracy across various noise levels. Second, methods like the Average Ensemble Method, which aggregated multiple ELM-based SoH estimates, and the Hierarchical Ensemble Method, though improving prediction accuracy, lacked validation under more diverse operating conditions [45]. Another approach combining LightGBM, and neural networks sought to address computational burden by working with sparse data points and incorporating skip connections for improved accuracy [46]. However, these techniques often neglected the impact of data uncertainty or variability in real-world datasets, limiting their generalizability. This paper presents a comprehensive ML-based approach for SoC estimation of EV LiBs, addressing the challenges of model reliability, uncertainty, and real-world data variability. To ensure the model's robustness and generalizability, preprocessing techniques, including normalization and scaling, are employed alongside

rigorous cross-validation methods. A well-structured ML pipeline is developed to integrate these processes, optimizing the entire model development cycle for efficiency and practical implementation. This work addresses generalizability and reliability issues using Extra Trees Regressor (ETR) and Light Gradient Boosting Machine (LightGBM) models. In the proposed ML pipeline, we utilize ETR and LightGBM to introduce an ensemble model, namely ETR-GBM, to improve adaptability to real-world conditions, particularly in the presence of noise and anomalies in the datasets. In this work, the performance of the proposed approach is assessed by benchmarking against other ML models. Cross-validation in dataset generation effectively reduces the risk of model overfitting. The main contributions of this research are summarized below.

1. Examining the ETR and LightGBM model reliability and generalizability.
2. Developing an ensemble ML model, ETR-GBM, which integrates the strengths of ETR and Light-GBM while using preprocessing techniques and cross-validation methods to enhance SoC estimation.
3. Assess the model performance robustness to the effect of the training and testing dataset noise to assist in dealing with unexpected anomalies and better adapt to real-world conditions.

The paper is organized into four main sections. Section I is the Introduction, while Section II provides the Methodology, a detailed explanation of the various ML models utilized in this study. The Results and Discussion section is covered in Section III. Finally, Section IV summarizes the research findings, highlights the significance of the proposed work, and suggests future research directions.

II. METHODOLOGY

The methodology for developing and evaluating the proposed ML model for SoC estimation is shown in FIGURE 2. The process begins with loading the dataset and involves feature-engineering steps such as normalizing SoC values and selecting relevant features. The dataset is split into 80% training and 20% testing sets with standardized features. The ML model is then defined and evaluated using 5-fold cross-validation. The model was first assumed to be trained on clean data, while white Gaussian noise was considered with the test data part to simulate real-world conditions. In the second experiment, white Gaussian noise was added to the training dataset, and the dataset test part was assumed clean to test the model's ability to handle noisy inputs during training. Estimations of the model's performance on the noisy data were evaluated. This approach emphasizes the model's robustness and accuracy in SoC estimation under various noise conditions.

A. BACKGROUND OF LIGHT GRADIENT BOOSTING MACHINE (LIGHTGBM)

Light Gradient Boosting Machine (LightGBM) [47], is a gradient-boosting algorithm that builds a robust model by

combining decision trees and correcting errors in each iteration to improve accuracy. It uses histogram optimization, which discretizes continuous features into bins to reduce computational complexity and speed up training. Unlike traditional methods, LightGBM grows trees leaf-wise, selecting the leaf with the most significant error reduction, allowing for deeper and more accurate trees. The model's output is the sum of individual tree outputs:

$$\phi_T(x) = \sum_{t=1}^{T} \phi_t(x) \rightarrow \phi_T \in \Theta \qquad (1)$$

where $\phi_T(x)$ is the model's output as a function of input x, and $\phi_t(x)$ represents the output from each tree.

The objective function LightGBM aims to minimize is

$$R_{t(x)} = \arg\min_{R \in \phi} \left(F\left(Y, \sum_{t=1}^{T-1} \phi_{t-1}(x) + R_t(x) \right) \right) \quad (2)$$

where Y is the target variable, F is the loss function, and $\phi_{t-1}(x)$ is the cumulative output from previous iterations.

LightGBM is designed for efficiency and scalability, making it well-suited for large datasets. FIGURE 3 shows an overview of LightGBM.
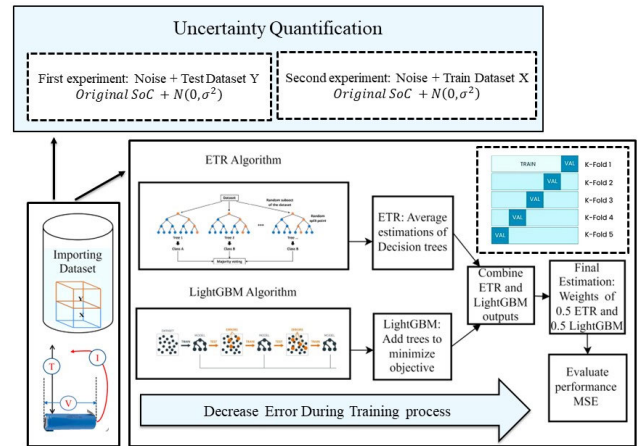

FIGURE 2. The machine learning pipeline for accurate SoC estimation of LiBs.

B. BACKGROUND OF EXTRA TREE REGRESSOR (ETR)

The Extra Tree Regressor (ETR) is used for regression tasks like SoC estimation in EV batteries. It enhances the Random Forest model by using unpruned trees and selecting random split points for features, minimizing error:

$$\text{Split}_{\text{ETR}} = \arg\min_{f,s}[\text{Error}(f, s)] \qquad (3)$$

where f represents a feature, s is a randomly picked split point for that feature, and Error (f, s) computes the decrease in error resulting from the split. The final output of the ETR model is the average of the outputs from all the individual trees in the ensemble, represented as

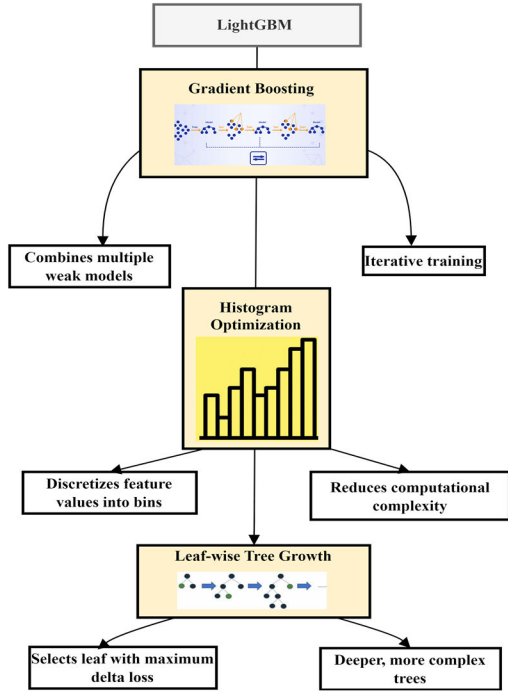$$Y_{\text{ETR}} = \frac{1}{N} \sum_{i=1}^{N} T_i(X) \qquad (4)$$

FIGURE 3. An overview of light gradient boosting machine (LightGBM).

where Y_{ETR} is the anticipated output, N is the number of trees in the ensemble, T_i is the i-th tree, and X is the input feature vector.

Based on their complementary strengths, the combination of ETR and LightGBM has been deployed in this work for SoC estimation in LiBs. The ETR ensures robustness and stability by introducing randomness through feature splitting, which reduces overfitting and variance and efficiently handles high-dimensional data. LightGBM complements this by capturing complex non-linear patterns through its gradient-boosting framework and enhancing accuracy with histogram-based optimization. Together, these models prioritize accuracy, effectively meeting the critical requirements of SoC estimation in LiBs.

C. PROPOSED ENSEMBLE LEARNING MODEL (ETR-GBM)
Ensemble learning is used to improve overall performance, which shows the ability to enhance accuracy and robustness. The work in Ref. [48] introduced an ensemble learning (EL) method using a Gaussian kernel-based distance-weighted assignment to improve prediction accuracy and robustness. In this approach, models are weighted based on their performance: those with smaller errors receive higher weights, while models with larger errors are downweighed to reduce the impact of outliers. The key steps include calculating the Euclidean error distance $\|d_j\|$, assigning weights with the Gaussian kernel $\omega_j = e^{-\frac{\|d_j\|^2}{2\sigma^2}}$, normalizing the weights $W_j = \frac{\omega_j}{\sum_{j=1}^{3} \omega_j}$, and computing the final ensemble prediction as $\hat{y}_{EL} \sum_{j=1}^{3} W_j \hat{y}_j$. In contrast, our contribution is the simple ensemble averaging method, where all model predictions are

combined equally. The final prediction is simply the average of all outputs $\hat{y}_{avg} = \frac{1}{n} \sum_{j=1}^{n} \hat{y}_j$. Both the simple averaging method and the distance-based weighted assignment are valid ensemble techniques, but they differ in approach. While the previous work's method provides refined, performance-based weighting, our contribution offers a straightforward and efficient alternative by treating all models equally.

In this study, we propose an ensemble learning model, ETR-GBM, for estimating the SoC in LiBs. This model combines ETR and LightGBM algorithms. We will detail the model architecture of both components and their integration. ETR builds an ensemble of unpruned decision trees using the following process:

1. For each node, randomly select K features without replacement.
2. For each selected feature, randomly choose a split point.
3. Among these K splits, select the best split according to a normalized gain criterion:

$$\text{Score}(s, t) = \frac{I(s, t)}{\text{H}(t)} \qquad (5)$$

where $I(s, t)$ is the mutual information of split s at node t and $\text{H}(t)$ is the entropy of node t.

4. Repeat until a stopping criterion is met (e.g., minimum samples per leaf).

The prediction of the ETR for a new sample x is the average of predictions of all trees:

$$\hat{y}_{\text{ETR}}(x) = \frac{1}{M} \sum_{m=1}^{M} T_m(x) \qquad (6)$$

where M is the number of trees and $T_m(x)$ is the prediction of the m-th tree.

LightGBM builds an additive model of decision trees. The objective function to be minimized is:

$$\text{Obj} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i\right) + \sum_{j=1}^{t} \Omega\left(f_j\right) \qquad (7)$$

where L is the loss function, Ω is the regularization term, and f_j is the j-th tree.

At each iteration k, Light GBM adds a new tree f_k to minimize:

$$\text{Obj}^{(k)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)\right) + \Omega(f_k) \qquad (8)$$

Using Taylor expansion, this can be approximated as:

$$\text{Obj}^{(k)} \approx \sum_{i=1}^{n} \left[L\left(y_i, \hat{y}_i^{(k-1)}\right) + g_i f_k(x_i) + \frac{1}{2} h_i f_k^2(x_i) \right] + \Omega(f_k) \qquad (9)$$

where g_i and h_i are first and second-order gradients of the loss function.

LightGBM uses a leaf-wise growth strategy, splitting the leaf with maximum delta loss:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{F'_\downarrow + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \qquad (10)$$

where G_{Lr}, G_R, H_L, H_R are the sums of gradients and hessians in left and right nodes, λ is L 2 regularization, and γ is the minimum gain for a split.

Ensemble Integration

In our approach, we used a simple average method for model combination [49]. A straightforward approach is to assign equal weights to all models, assuming they contribute equally, which is given by:

$$w_i = \frac{1}{N} \tag{11}$$

where N is the number of models. This method is easy to implement, avoiding optimizing weights, the risk of overfitting, and does not require additional tuning. In our proposed method, the ETR-GBM model combines ETR and Light-GBM using:

$$\hat{y}(x) = \frac{1}{2}\left(\hat{y}_{ETR}(x) + \hat{y}_{LightGBM}(x)\right) \tag{12}$$

where $\hat{y}(x)$ is the final estimation, $\hat{y}_{ETR}(x)$ is the ETR estimation, and $\hat{y}_{LightGBM}(x)$ is the LightGBM estimation. The Mean Squared Error (MSE) is:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}(x_i)\right)^2 \tag{13}$$

This ensemble balances the strengths of both models, aiming for more robust estimations across different battery conditions by assuming both contribute equally to the result. This simple averaging is effective when the models perform comparably or complement each other well.

This ensemble approach leverages the strengths of both ETR (randomness and robustness) and LightGBM (efficiency and handling of complex patterns) to provide accurate and reliable SoC estimations for LiBs. Because ETR randomness helps prevent overfitting and reduces variance, making it robust against noise. On the other hand, LightGBM excels at capturing complex patterns in large datasets, thereby reducing bias. By combining these models, their strengths balance out, leading to more accurate and stable predictions.

III. RESULTS

This section provides an in-depth look at the experimental setup and results. It begins by describing the battery data source and an overview of the system setup for implementation. It will also highlight the feature engineering techniques used to improve model accuracy, followed by implementing various ML models. The results of two essential experiments are presented, showing the performance of the models under different data uncertainty levels. Finally, the section compares the findings for both experiments.

A. BATTERY DATASET

The dataset in [50] is part of the comprehensive battery dataset curated by the NASA Ames Research Centre of Excellence. This dataset encompasses the aging data for four LiBs with serial numbers B0005, B0006, B0007, and

---

Algorithm 1 Algorithm of Model Ensemble with Extra Trees Regressor and LightGBM Regressor

Input: Dataset: {X, y}
Features: {voltage measured, current measured, temperature measured, current load, voltage load} Target: SoC
Test size: 0.2
Random state: 42
Models: {ETR, LightGBM}
Algorithm:

1	**Data Preprocessing:** Split dataset into training and test sets: X train, X test, y train, y test = train test split (features, target)
2	**Model Initialization:** ETR model = Extra Trees Regressor (estimators = 100) LGBM model = LightGBM Regressor (estimators = 100)
3	**Training the Models:** Train models on the training set: ETR model. Fit (X train, y train) LGBM model. Fit (X train, y train)
4	**Ensemble the Models:** Define the Ensemble function: ensemble models estimation(X): ETR_est = ETR model. Estimation(X) LGBM_est = LGBM model. Estimation (X) ensemble Est = ETR_est + LGBM_est
5	**Estimation:** Estimate on the test set using the ensemble model: ensemble Est = ensemble models estimation (X test)
6	**Evaluation:** Define the evaluation metric (RMSE) evaluate model (y_{true}, y_{est}) $RMSE = \sqrt{mean\_squared\_error\,(y_{true}, y_{est})}$
7	**Calculate RMSE:** RMSE = evaluate model (y test, ensemble Est)

Output: Ensembled Model RMSE

B0018, featuring LiCoO2/LiNiCoAlO2 as the cathode material and graphite as the anode material [51] as shown in TABLE 2. The battery specifications include a rated capacity of 2.0 Ah. Testing conditions involved a temperature of 24°C, with the charging process comprising charging at 1.5A constant current (CC) until the voltage reached 4.2V, followed by charging at 4.2V constant voltage (CV) until the charge current dropped below 20mA. The failure Threshold is 70% of the rated capacity. The discharge process involved discharging at 2A CC until the voltage dropped to 2.7V over 168 cycles [52]. The dataset contains 50,285 entries and 10 columns, capturing various aspects of the battery's performance over time. The columns in TABLE 2 include the main features of the table: cycle (cycle number), ambient temperature (°C), datetime (date and time of measurement), capacity (Ah), voltage measured (V), current measured (A), temperature measured (°C), current load (A), voltage load (V), and time (s).

As described in the dataset summary (TABLE 2), the B0005 dataset's constant discharging process was conducted at a constant current (CC) of 2A until a cutoff voltage of 2.7V. This choice aligns with the study's primary objective of simulating uncertainty and validating it through ML techniques. Constant discharging current provides a stable, controlled environment essential for minimizing external variability,

allowing focused analysis of battery behavior under uncertainty, developing robust ML models, and ensuring reliable and repeatable results. It also simplifies the analysis by reducing modeling complexity, enabling a clear battery performance evaluation.

In contrast, while reflective of real-world usage, dynamic discharging currents introduce variability that can obscure uncertainty patterns and complicate ML model training and validation. Other datasets, such as the public CALCE battery dataset, and studies using dynamic discharging currents, like those employing Dynamic Equivalent Electric Circuit Models (DEECM) [53], analyze battery behavior under variable loads and nonlinear conditions. While these approaches offer valuable insights into dynamic performance, their variability, and complexity are less suited to this study's focus on controlled uncertainty simulation and effective ML validation.

The proposed method was implemented exclusively on the B0005 dataset, focusing on adding noise (std) at different levels. Noise was added to the test dataset in the first experiment and the training dataset in the second. This approach was designed to systematically evaluate the model's robustness to varying noise levels, providing insights into its performance reliability and generalizability in real-world conditions.

TABLE 2. NASA battery dataset summary.

Attribute	Description
Dataset	B0005 (part of NASA battery dataset).
Materials	Cathode: LiCoO2/LiNiCoAlO2; Anode: Graphite
Capacity	2.0 Ah
Temperature	24°C
Charge Process	CC at 1.5A until 4.2V; CV at 4.2V until current < 20mA
Discharge Process	CC at 2A until 2.7V
Cycles	168 cycles
Entries	50,285
Columns	Cycle, Ambient Temp (°C), Datetime, Capacity (Ah), Voltage Measured (V), Current Measured (A), Temp Measured (°C), Current Load (A), Voltage Load (V), Time (s)

B. ADDRESSING UNCERTAINTY FOR THE B0005 DATASET

The Additive White Gaussian Noise (AWGN) [54] used in the proposed work is one of the most extensively studied types of noise. It is characterized by a normal distribution and constant power across all frequencies, making it ideal for simulations and theoretical studies due to its mathematical simplicity. The uncertainty quantification involved introducing AWGN to the data's magnitude and SoC to produce the uncertainty as the value per unit is (p.u.), as shown in FIGURE 4. The AWGN added to the dataset can be represented by the following equation [55]:

$$Noisy\ SoC = Original\ SoC + N\left(0, \sigma^2\right) \quad (14)$$

where $N\left(0, \sigma^2\right)$ represents Gaussian noise with a mean of 0 and variance σ^2, where σ is the standard deviation. Otherwise, Additive White Laplacian Noise [56] shares similarities

with AWGN [9], [57] but follows a Laplace distribution with heavier tails, providing greater robustness to outliers. Moreover, Additive White Uniform Noise has a uniform distribution, where all values within a specific range are equally likely. FIGURE 5 and FIGURE 6 compare the distribution of SoC values. FIGURE 5 presents the original distribution without noise, characterized by a clear structure with distinct peaks and relatively consistent patterns across the range of SoC values. This reflects the raw characteristics of the dataset, which is free from any external interference. In contrast, FIGURE 6 illustrates the impact of adding noise with a standard deviation of 0.1. Peaks observed in the noisy dataset are less pronounced, and the frequency histogram appears more randomized than the original dataset. This shift indicates that the noise introduces variability, which can mimic uncertainties. Such modifications are crucial for testing ML models and simulating scenarios with inherent measurement fluctuations.

C. FEATURE ENGINEERING TECHNIQUES

Several feature engineering techniques, such as normalizations, scaling, and cross-validation, are employed to enhance the performance and robustness of the proposed ML models. These methods are fundamental to preparing the data for accurate and consistent model performance. The following subsections demonstrate the essential steps for the proposed feature engineering process to be utilized in the proposed model.
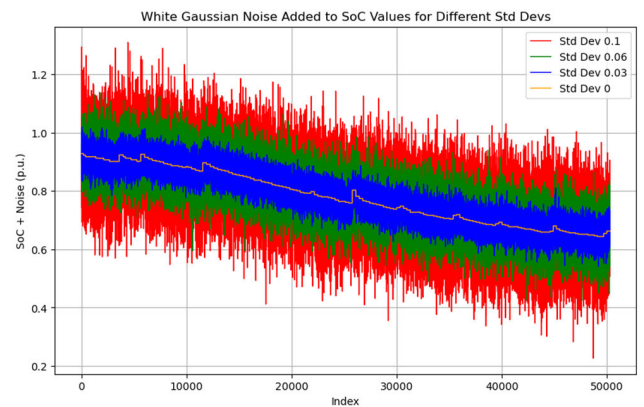

FIGURE 4. Impact of white gaussian noise on SoC estimation while adding different standard deviations for noise.

1) NORMALIZATION

Normalization [58] plays a crucial role in calculating the SoC of the battery, a parameter that is indirectly derived from correlated factors such as voltage, current, temperature, and capacity. Capacity values were normalized using the battery's rated capacity of 2.0 Ah. The normalization can be mathematically expressed as:

$$SoC = \frac{C}{C_{rated}} \quad (15)$$

where C is the observed capacity and C_{rated} is the rated capacity of the battery (2.0 Ah).

2) SCALING

Feature scaling was applied to standardize input data, a crucial step for models sensitive to feature magnitudes [59]. This was achieved using the formula:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \qquad (16)$$

where x is the original feature value, μ is the mean, and σ is the standard deviation of the feature in the training set. After scaling, the transformed feature x_{scaled} has a mean of zero $\mu = 0$ and a standard deviation of one $\sigma = 1$. As a result, all features are on the same scale, ensuring they contribute equally to the model, preventing any single feature from disproportionately influencing the learning process of ML models. After computing μ and σ, both the training and test datasets are standardized for consistency.

D. CROSS-VALIDATION AND MODEL EVALUATION

To validate the performance of the proposed model, we employed 5-fold cross-validation, using the MSE as the scoring metric. 5-fold cross-validation is favored for its balanced approach, efficiently balancing computational resources with accurate model performance estimates [60]. It strikes a middle ground between reducing variance and managing computational costs compared to methods like 10-fold cross-validation or leave-one-out cross-validation [61]. It offers a reliable compromise between minimizing bias and variance in model evaluation by training the model on 80% of the data and testing it on the remaining 20% in each fold. The dataset generated through cross-validation helps minimize the risk of model overfitting.
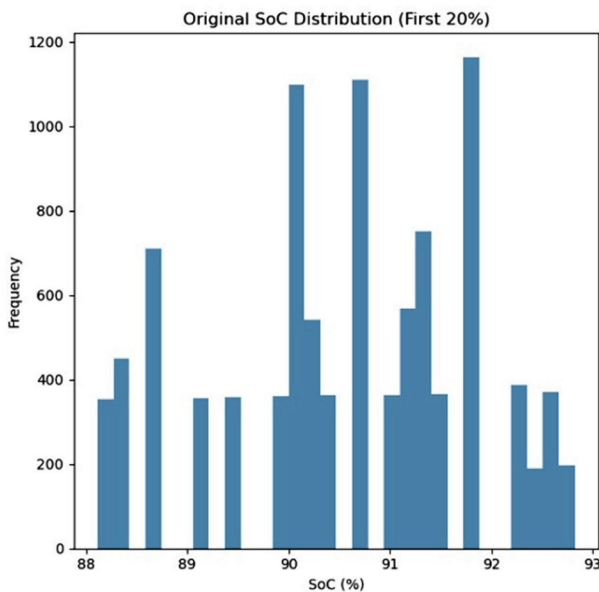

FIGURE 5. Original SoC distribution for the first 20% of the dataset.
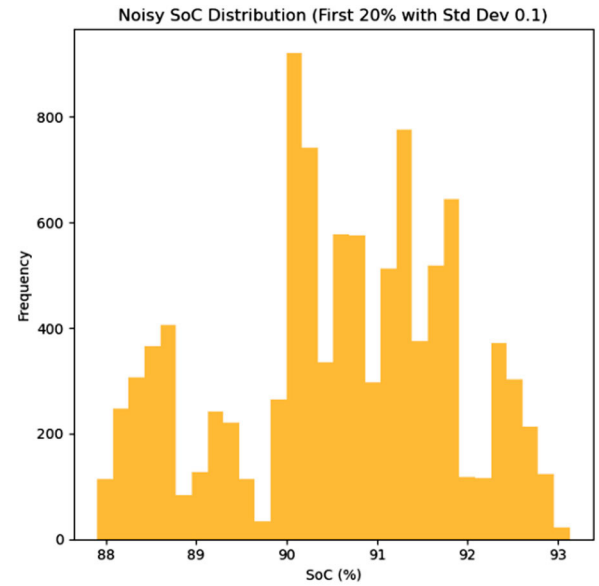

FIGURE 6. Noisy SoC distribution for the first 20% of the dataset (Std Dev 0.1).

We calculated the root mean squared error (RMSE) metric to evaluate the model's performance on the test set. RMSE is considered in this work to assess the performance of the proposed ML model.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i-1}^{n} \left(y_i - \hat{y}_i \right)^2} \qquad (17)$$

where $\frac{1}{n}\sum_{i-1}^{n} y_i$ is the mean of the actual values and $\hat{y}$ estimated value at index i. We used RMSE as the sole metric for SoC estimation due to its ability to penalize larger errors more heavily. This is critical in battery management systems, where large deviations in SoC estimations can impact safety and performance. RMSE's focus on larger errors aligns with the need for precise and reliable SoC estimates in practical applications.

E. MODEL IMPLEMENTATION

The methodology involves two distinct trials. In the primary trial, Gaussian noise is introduced into the test dataset to replicate the inherent uncertainties found in real-world data. In addition to the ETR-GBM, other ML algorithms—including ETR, LightGBM, Stacking Ensemble, Gated Recurrent Unit (GRU), SVR, CatBoost, and Bayesian Ridge Regression—are employed to assess the proposed concept. In the second trial, noise is introduced into the training dataset. Cross-validation ensured robustness, and the pipeline was trained and evaluated using the RMSE metric. FIGURE 7 shows a heatmap of correlation analysis among the main six features: measured voltage, measured current, temperature, load current, load voltage, and capacity. It highlights key relationships, such as the strong negative correlation between measured temperature and voltage and a moderate positive

correlation between measured voltage and load voltage, providing insights into feature interactions for SoC estimation.

F. FIRST EXPERIMENT RESULTS

We studied the effect of adding white Gaussian noise (AWGN) on different ML models for SoC estimation. Noise was introduced to the test dataset only to evaluate the robustness and reliability of the models under uncertain conditions. The findings indicate that the ETR-GBM model outperforms other ML approaches at varying noise levels, as shown in FIGURE 8.
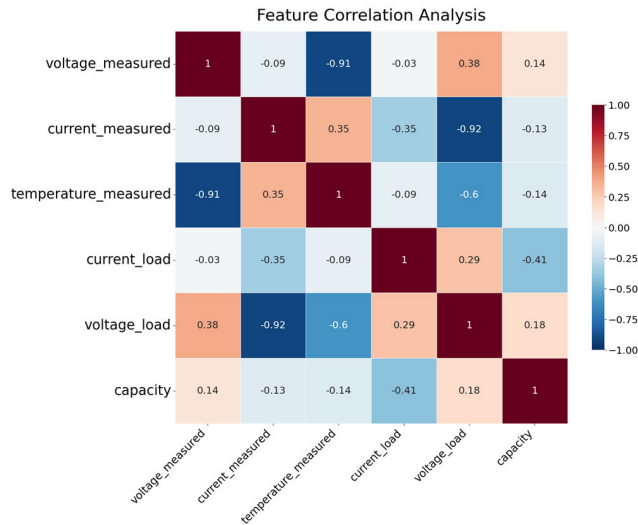
FIGURE 7. The heat map shows the correlation analysis, illustrating the relative significance of various features in an estimation model.

FIGURE 8. Impact of white gaussian noise in testing dataset on RMSE for different machine learning algorithms.

FIGURE 9. Impact of white gaussian noise in training dataset on RMSE for different machine learning algorithms.

G. SECOND EXPERIMENT RESULTS

In this section, a comprehensive analysis of the model performances was presented, focusing specifically on the impact of AWGN on only the training dataset to show the accuracy of SoC estimation. The AWGN was introduced with varying standard deviation values (ranging from 0.01 to 0.1), and the models were evaluated using the RMSE.

FIGURE 9 illustrates ML models exhibit increasing RMSE as the standard deviation of the noise increases, indicating a decline in estimation accuracy with higher noise levels. In this range (0.01 to 0.1), the ETR-GBM model achieves a lower RMSE when compared to other ML algorithms, as shown in FIGURE 9. indicating higher robustness and better accuracy under noisy conditions.

In the testing dataset scenario (FIGURE 8) where additive white noise is introduced exclusively to the test dataset while the training dataset remains free from noise, ETR-GBM and the Bayesian method exhibit comparable levels of anti-noise performance. This outcome can be attributed to the shared training condition on a clean dataset, which minimizes the differences in model behavior under these testing circumstances.
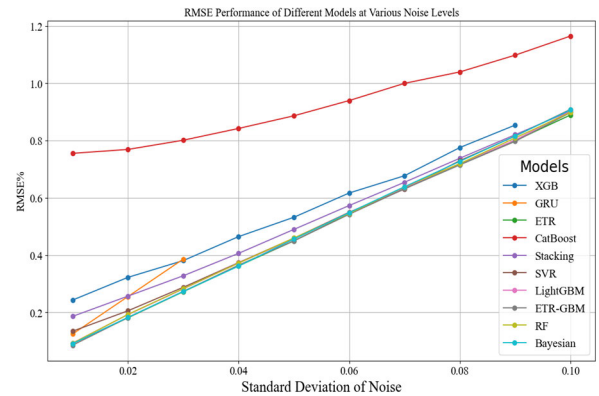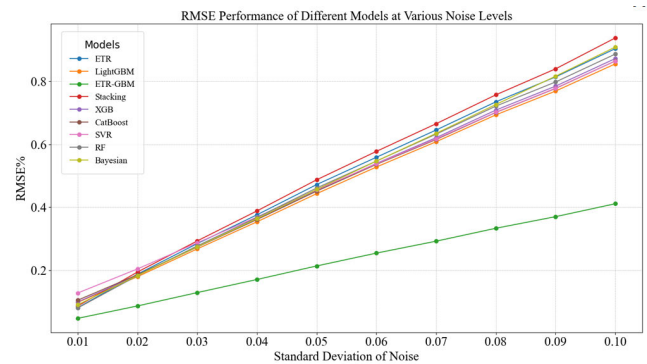
Conversely, FIGURE 9 examines the scenario where additive white noise is introduced exclusively to the training dataset while the test dataset remains entirely noise-free. This setup highlights the models' distinct capabilities in handling noisy training data. ETR-GBM demonstrates a marked advantage, achieving significantly lower RMSE values across all noise levels. In contrast, the Bayesian method struggles to adapt to high-noise conditions, resulting in higher RMSE values. This highlights the importance of robust anti-noising mechanisms, exemplified by ETR-GBM's high adaptability.

ETR and LightGBM ensemble balances the bias-variance trade-off by combining ETR's robustness against noise with LightGBM accuracy in capturing complex patterns. This combination enhances accuracy and improves resilience to noisy data.

H. ENSEMBLE METHODS COMPARISON

In this study, the effectiveness of the proposed ETR-GBM ensemble for SoC estimation in LiBs was evaluated and compared to three established methods under varying noise levels, demonstrating its potential for improved accuracy and robustness. The results show that the ETR-GBM model

generally has the lowest RMSE while adding noise to the training dataset, indicating better performance in handling noise. In this range (0.01 to 0.1), noise levels increase, and all models experience higher RMSE values, with the ETR-GBM model remaining the most resilient, as shown in FIGURE 10. The results show improvement over the ensemble models used in previous work. By incorporating advanced techniques, our approach outperforms prior methods. The ELM-RVFL [44] and Hierarchical ELM [45] methods show similar trends with higher RMSE, while the Ensemble ELM-RF [62] model performs moderately. These results suggest that the ETR-GBM model could offer more reliable SoC estimation than the other methods, particularly in noisy conditions.
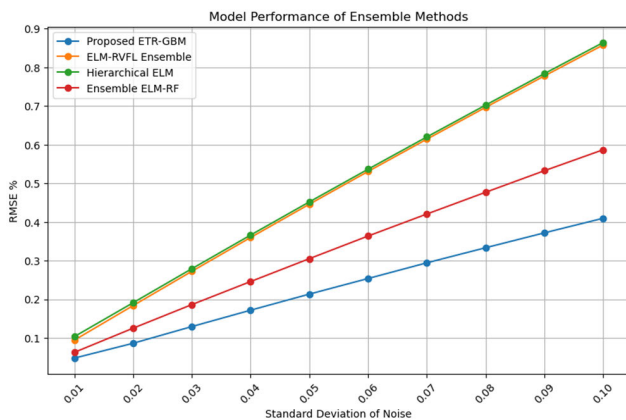

FIGURE 10. Impact of white gaussian noise on RMSE for different ensemble methods compared to the proposed ETR-GBM.

IV. CONCLUSION

The paper implemented the SoC estimation model for LiBs, demonstrating significant advancements in accuracy and reliability through an ensemble ML approach. The process included data preprocessing and utilizing various ML techniques. The work focuses on investigating the impact of AWGN on both the training dataset and the testing dataset. The proposed method estimates the SoC of LiBs using ETR-GBM. The process involves loading and preprocessing the dataset, normalizing SoC, and splitting the data into training and testing sets.

In the conclusion of the proposed work, the AWGN was added to the test and training dataset, revealing that different models performed optimally at several noise levels.

Future work could delve into integrating noise reduction techniques not only with the ETR-GBM model but also with other ML models. Additionally, investigating the impact of several types of noise or data anomalies on the performance of these models would provide a more comprehensive understanding of the challenges in SoC estimation. Moreover, Future work will focus on optimizing the ETR-LightGBM model to enhance its computational efficiency while maintaining its high estimative accuracy. Furthermore, improving uncertainty quantification in RUL

prediction and SoC estimation will be key, as they are closely linked. SoC estimation provides real-time battery charge status, while RUL prediction forecasts the remaining operational lifespan. Accurate SoC estimation is essential for precise RUL predictions, enabling better battery management and informed decision-making for both short-term use and long-term maintenance.

REFERENCES

[1] U. Im, S. E. Bauer, L. M. Frohn, C. Geels, K. Tsigaridis, and J. Brandt, "Present-day and future $PM_{2.5}$ and O_3-related global and regional premature mortality in the EVAv6.0 health impact assessment model," *Environ. Res.*, vol. 216, Jan. 2023, Art. no. 114702.

[2] A. Emadi, *Advanced Electric Drive Vehicles*. New York, NY, USA: CRC Press, 2015.

[3] E. Chemali, M. Preindl, P. Malysz, and A. Emadi, "Electrochemical and electrostatic energy storage and management systems for electric drive vehicles: State-of-the-art review and future trends," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 4, no. 3, pp. 1117–1134, Sep. 2016.

[4] C. Xu, Q. Dai, L. Gaines, M. Hu, A. Tukker, and B. Steubing, "Future material demand for automotive lithium-based batteries," *Commun. Mater.*, vol. 1, no. 1, pp. 1–10, Dec. 2020.

[5] D. Pevec, J. Babic, A. Carvalho, Y. Ghiassi-Farrokhfal, W. Ketter, and V. Podobnik, "A survey-based assessment of how existing and potential electric vehicle owners perceive range anxiety," *J. Cleaner Prod.*, vol. 276, Dec. 2020, Art. no. 122779.

[6] S. Sautermeister, M. Falk, B. Bäker, F. Gauterin, and M. Vaillant, "Influence of measurement and prediction uncertainties on range estimation for electric vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 8, pp. 2615–2626, Aug. 2018.

[7] M. U. Ali, A. Zafar, S. H. Nengroo, S. Hussain, M. Junaid Alvi, and H.-J. Kim, "Towards a smarter battery management system for electric vehicle applications: A critical review of lithium-ion battery state of charge estimation," *Energies*, vol. 12, no. 3, p. 446, Jan. 2019.

[8] Y. Yang, Y. Ye, Z. Cheng, G. Ruan, Q. Lu, X. Wang, and H. Zhong, "Life cycle economic viability analysis of battery storage in electricity market," *J. Energy Storage*, vol. 70, Oct. 2023, Art. no. 107800, doi: 10.1016/j.est.2023.107800.

[9] A. Haraz, K. Abualsaud, and A. Massoud, "State-of-health and state-of-charge estimation in electric vehicles batteries: A survey on machine learning approaches," *IEEE Access*, vol. 12, pp. 158110–158139, 2024, doi: 10.1109/ACCESS.2024.3486989.

[10] Y. Jiang, T. Ortmeyer, and M. Fan, "Data-driven fast uncertainty assessment of distribution systems with correlated EV charging demand and renewable generation," *IEEE Trans. Sustain. Energy*, vol. 14, no. 3, pp. 1446–1456, Jul. 2023, doi: 10.1109/TSTE.2023.3236446.

[11] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "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.

[12] B. T. Phan. (Sep. 17, 2019). *Bayesian Deep Learning and Uncertainty in Computer Vision*. Accessed: Aug. 14, 2024. [Online]. Available: http://hdl.handle.net/10012/15056

[13] Y. Gal, "Uncertainty in deep learning," Ph.D. thesis, Univ. Cambridge, Nanjing Univ. Aeronaut. Astronaut. (NUAA), Nanjing, China, 2016. [Online]. Available: https://www.cs.ox.ac.uk/people/yarin.gal/website/thesis/thesis.pdf

[14] A. Malinin and M. J. F. Gales, "Uncertainty estimation in deep learning with application to spoken language assessment," Doctoral dissertation, Dept. Eng., Univ. Cambridge, Cambridge, U.K., Oct. 2019, doi: 10.17863/CAM.45912.

[15] A. Tomaszewska, Z. Chu, X. Feng, S. O'Kane, X. Liu, J. Chen, C. Ji, E. Endler, R. Li, L. Liu, Y. Li, S. Zheng, S. Vetterlein, M. Gao, J. Du, M. A. Parkes, M. Ouyang, M. Marinescu, G. J. Offer, and B. Wu, "Lithium-ion battery fast charging: A review," *ETransportation*, vol. 1, Aug. 2019, Art. no. 100011.

[16] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Mach. Intell.*, vol. 1, no. 1, pp. 20–23, Jan. 2019, doi: 10.1038/s42256-018-0004-1.

[17] X. Tan, D. Zhan, P. Lyu, J. Rao, and Y. Fan, "Online state-of-health estimation of lithium-ion battery based on dynamic parameter identification at multi timescale and support vector regression," *J. Power Sources*, vol. 484, Feb. 2021, Art. no. 229233, doi: 10.1016/j.jpowsour.2020.229233.

[18] P. Pillai, K. R. Pattipati, and B. Balasingam, "A robust extended Kalman filtering approach for state of charge estimation in batteries," *IEEE J. Emerg. Sel. Topics Ind. Electron.*, vol. 5, no. 3, pp. 1154–1170, Jul. 2024, doi: 10.1109/JESTIE.2023.3339429.

[19] H. Li, D. Pan, and C. L. P. Chen, "Intelligent prognostics for battery health monitoring using the mean entropy and relevance vector machine," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 7, pp. 851–862, Jul. 2014, doi: 10.1109/TSMC.2013.2296276.

[20] J. Zhao, H. Ling, J. Liu, J. Wang, A. F. Burke, and Y. Lian, "Machine learning for predicting battery capacity for electric vehicles," *eTransportation*, vol. 15, Jan. 2023, Art. no. 100214, doi: 10.1016/j.etran.2022.100214.

[21] K. Liu, Y. Shang, Q. Ouyang, and W. D. Widanage, "A data-driven approach with uncertainty quantification for predicting future capacities and remaining useful life of lithium-ion battery," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3170–3180, Apr. 2021, doi: 10.1109/TIE.2020.2973876.

[22] Y. Ke, Y. Jiang, R. Zhu, W. Peng, and X. Tan, "Early prediction of knee point and knee capacity for fast-charging lithium-ion battery with uncertainty quantification and calibration," *IEEE Trans. Transport. Electrific.*, vol. 10, no. 2, pp. 2873–2885, Jun. 2024, doi: 10.1109/TTE.2023.3304670.

[23] C. Chen, G. Tao, J. Shi, M. Shen, and Z. H. Zhu, "A lithium-ion battery degradation prediction model with uncertainty quantification for its predictive maintenance," *IEEE Trans. Ind. Electron.*, vol. 71, no. 4, pp. 3650–3659, Apr. 2024, doi: 10.1109/TIE.2023.3274874.

[24] Z. Xu, Y. Guo, and J. H. Saleh, "Accurate remaining useful life prediction with uncertainty quantification: A deep learning and nonstationary Gaussian process approach," *IEEE Trans. Rel.*, vol. 71, no. 1, pp. 443–456, Mar. 2022, doi: 10.1109/TR.2021.3124944.

[25] P. Malysz, R. Gu, J. Ye, H. Yang, and A. Emadi, "State-of-charge and state-of-health estimation with state constraints and current sensor bias correction for electrified powertrain vehicle batteries," *IET Electr. Syst. Transp.*, vol. 6, no. 2, pp. 136–144, Jun. 2016.

[26] L. McCurlie, M. Preindl, and A. Emadi, "Fast model predictive control for redistributive lithium-ion battery balancing," *IEEE Trans. Ind. Electron.*, vol. 64, no. 2, pp. 1350–1357, Feb. 2017.

[27] H. Gabbar, A. Othman, and M. Abdussami, "Review of battery management systems (BMS) development and industrial standards," *Technologies*, vol. 9, no. 2, p. 28, Apr. 2021.

[28] S. Zhang, X. Guo, X. Dou, and X. Zhang, "A data-driven Coulomb counting method for state of charge calibration and estimation of lithium-ion battery," *Sustain. Energy Technol. Assessments*, vol. 40, Aug. 2020, Art. no. 100752, doi: 10.1016/j.seta.2020.100752.

[29] D. N. T. How, M. A. Hannan, M. S. Hossain Lipu, and P. J. Ker, "State of charge estimation for lithium-ion batteries using model-based and data-driven methods: A review," *IEEE Access*, vol. 7, pp. 136116–136136, 2019, doi: 10.1109/ACCESS.2019.2942213.

[30] W. Zine, Z. Makni, E. Monmasson, L. Idkhajine, and B. Condamin, "Interests and limits of machine learning-based neural networks for rotor position estimation in EV traction drives," *IEEE Trans. Ind. Informat.*, vol. 14, no. 5, pp. 1942–1951, May 2018, doi: 10.1109/TII.2017.2765398.

[31] H. S. Salama, S. M. Said, I. Vokony, and B. Hartmann, "Adaptive coordination strategy based on fuzzy control for electric vehicles and superconducting magnetic energy storage—Towards reliably operating utility grids," *IEEE Access*, vol. 9, pp. 61662–61670, 2021, doi: 10.1109/ACCESS.2021.3074578.

[32] H. Maghfiroh, O. Wahyunggoro, and A. I. Cahyadi, "Energy management in hybrid electric and hybrid energy storage system vehicles: A fuzzy logic controller review," *IEEE Access*, vol. 12, pp. 56097–56109, 2024, doi: 10.1109/ACCESS.2024.3390436.

[33] B. Moulik, A. K. Dubey, and A. M. Ali, "A battery modeling technique based on fusion of hybrid and adaptive algorithms for real-time applications in pure EVs," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 2760–2771, Mar. 2023, doi: 10.1109/TITS.2022.3224561.

[34] M. S. H. Lipu, M. A. Hannan, A. Hussain, M. H. M. Saad, A. Ayob, and F. Blaabjerg, "State of charge estimation for lithium-ion battery using recurrent NARX neural network model based lighting search algorithm," *IEEE Access*, vol. 6, pp. 28150–28161, 2018, doi: 10.1109/ACCESS.2018.2837156.

[35] M. Savargaonkar, A. Chehade, and A. A. Hussein, "A novel neural network with Gaussian process feedback for modeling the state-of-charge of battery cells," *IEEE Trans. Ind. Appl.*, vol. 58, no. 4, pp. 5340–5352, Jul. 2022, doi: 10.1109/TIA.2022.3170842.

[36] A. Khalid and A. I. Sarwat, "Unified univariate-neural network models for lithium-ion battery state-of-charge forecasting using minimized Akaike information criterion algorithm," *IEEE Access*, vol. 9, pp. 39154–39170, 2021, doi: 10.1109/ACCESS.2021.3061478.

[37] Y. Choi, S. Ryu, K. Park, and H. Kim, "Machine learning-based lithium-ion battery capacity estimation exploiting multi-channel charging profiles," *IEEE Access*, vol. 7, pp. 75143–75152, 2019, doi: 10.1109/ACCESS.2019.2920932.

[38] P. Qin and L. Zhao, "An online transfer learning framework for cell SOC online estimation of battery pack in complex application conditions," *IEEE Trans. Transport. Electrific.*, vol. 10, no. 3, pp. 5974–5986, Sep. 2024, doi: 10.1109/TTE.2023.3324822.

[39] J. P. Ortiz, G. P. Ayabaca, A. R. Cardenas, D. Cabrera, and J. D. Valladolid, "Continual reforcement learning using real-world data for intelligent prediction of SOC consumption in electric vehicles," *IEEE Latin Amer. Trans.*, vol. 20, no. 4, pp. 624–633, Apr. 2022, doi: 10.1109/TLA.2022.9675468.

[40] M. Savargaonkar, I. Oyewole, A. Chehade, and A. A. Hussein, "Uncorrelated sparse autoencoder with long short-term memory for state-of-charge estimations in lithium-ion battery cells," *IEEE Trans. Autom. Sci. Eng.*, vol. 21, no. 1, pp. 15–26, Jan. 2024, doi: 10.1109/TASE.3222759.

[41] K. Sarrafan, K. M. Muttaqi, and D. Sutanto, "Real-time state-of-charge tracking embedded in the advanced driver assistance system of electric vehicles," *IEEE Trans. Intell. Vehicles*, vol. 5, no. 3, pp. 497–507, Sep. 2020, doi: 10.1109/TIV.2020.2973551.

[42] L. Zhang, K. Li, D. Du, Y. Guo, M. Fei, and Z. Yang, "A sparse learning machine for real-time SOC estimation of Li-ion batteries," *IEEE Access*, vol. 8, pp. 156165–156176, 2020, doi: 10.1109/ACCESS.2020.3017774.

[43] S. Sharma, A. Garg, and B. K. Panigrahi, "Predicting state-of-charge using gradient-boosted SVR ensemble technique for lithium ion battery used in EVs," *IEEE Trans. Transport. Electrific.*, vol. 10, no. 2, pp. 4441–4454, Jun. 2024, doi: 10.1109/TTE.2023.3310159.

[44] B. Gou, Y. Xu, and X. Feng, "State-of-health estimation and remaining-useful-life prediction for lithium-ion battery using a hybrid data-driven method," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 10854–10867, Oct. 2020, doi: 10.1109/TVT.2020.3014932.

[45] W. Liu, Y. Xu, and X. Feng, "A hierarchical and flexible data-driven method for online state-of-health estimation of Li-ion battery," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14739–14748, Dec. 2020, doi: 10.1109/TVT.2020.3037088.

[46] Y. Wang, P. Kou, J. Fan, and N. Zhou, "A novel capacity estimation method for Li-ion battery cell by applying ensemble learning to extremely sparse significant points," *IEEE Access*, vol. 10, pp. 96427–96441, 2022.

[47] S. Jafari and Y.-C. Byun, "Efficient state of charge estimation in electric vehicles batteries based on the extra tree regressor: A data-driven approach," *Heliyon*, vol. 10, no. 4, Feb. 2024, Art. no. e25949, doi: 10.1016/j.heliyon.2024.e25949.

[48] X. Zhao, S. Liu, E. Li, Z. Wang, H. Miao, F. Gu, and A. D. Ball, "A transfer learning-based ensemble learning model for electric vehicles lithium-ion battery capacity estimation using electrochemical impedance spectroscopy," *Sustain. Energy Technol. Assessments*, vol. 68, Aug. 2024, Art. no. 103886, doi: 10.1016/j.seta.2024.103886.

[49] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2012, p. 222.

[50] *SOC and SOH Estimation*. Accessed: Jun. 30, 2024. [Online]. Available: https://www.kaggle.com/datasets/priyar2111/soc-and-soh-estimation

[51] J. Wu, Z. Liu, Y. Zhang, D. Lei, B. Zhang, and W. Cao, "Data-driven state of health estimation for lithium-ion battery based on voltage variation curves," *J. Energy Storage*, vol. 73, Dec. 2023, Art. no. 109191, doi: 10.1016/j.est.2023.109191.

[52] M. Wu, C. Yue, F. Zhang, R. Sun, J. Tang, S. Hu, N. Zhao, and J. Wang, "State of health estimation and remaining useful life prediction of lithium-ion batteries by charging feature extraction and ridge regression," *Appl. Sci.*, vol. 14, no. 8, p. 3153, Apr. 2024, doi: 10.3390/app14083153.

[53] S. Lavety, R. K. Keshri, and M. A. Chaudhari, "A dynamic battery model and parameter extraction for discharge behavior of a valve regulated lead-acid battery," *J. Energy Storage*, vol. 33, Jan. 2021, Art. no. 102031, doi: 10.1016/j.est.2020.102031.

[54] P. Kokil and T. Pratap, "Additive white Gaussian noise level estimation for natural images using linear scale-space features," *Circuits Syst. Signal Process.*, vol. 40, no. 1, pp. 353–374, Jun. 2020, doi: 10.1007/s00034-020-01475-x.

[55] G. Ozcan, M. Pajovic, Z. Sahinoglu, Y. Wang, P. V. Orlik, and T. Wada, "Online battery state-of-charge estimation based on sparse Gaussian process regression," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2016, pp. 1–5.

[56] N. Liu, L. Hong, F. Pan, H. Yang, and J. Wang, "Performance analysis for visible light communications with Gaussian-plus-Laplacian additive noise," *Telecommun. Syst.*, vol. 85, no. 3, pp. 373–387, Dec. 2023, doi: 10.1007/s11235-023-01081-7.

[57] D. Zhang, C. Zhong, P. Xu, and Y. Tian, "Deep learning in the state of charge estimation for Li-ion batteries of electric vehicles: A review," *Machines*, vol. 10, no. 10, p. 912, Oct. 2022, doi: 10.3390/machines10100912.

[58] M. A. Siddiqi and W. Pak, "An agile approach to identify single and hybrid normalization for enhancing machine learning-based network intrusion detection," *IEEE Access*, vol. 9, pp. 137494–137513, 2021, doi: 10.1109/ACCESS.2021.3118361.

[59] M. Wang, W. Fu, X. He, S. Hao, and X. Wu, "A survey on large-scale machine learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 6, pp. 2574–2594, Jun. 2022, doi: 10.1109/TKDE.2020.3015777.

[60] Z. Barut and V. Altuntaş, "Comparison of performance of different K values with K-fold cross validation in a graph-based learning model for incrna-disease prediction," *Kirklareli Üniversitesi Mühendislik ve Fen Bilimleri Dergisi*, vol. 9, no. 1, pp. 63–82, Jun. 2023, doi: 10.34186/klu-jes.1248062.

[61] A. Geroldinger, L. Lusa, M. Nold, and G. Heinze, "Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—A simulation study," *Diagnostic Prognostic Res.*, vol. 7, no. 1, p. 9, May 2023, doi: 10.1186/s41512-023-00146-0.

[62] X. Shu, Z. Chen, J. Shen, S. Shen, F. Guo, Y. Zhang, and Y. Liu, "Ensemble learning and voltage reconstruction based state of health estimation for lithium-ion batteries with twenty random samplings," *IEEE Trans. Power Electron.*, vol. 38, no. 4, pp. 5538–5548, Apr. 2023, doi: 10.1109/TPEL.2023.3235872.
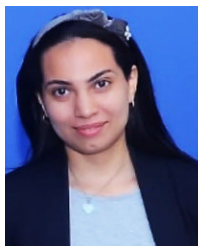
AYA HARAZ received the B.Sc. degree in biomedical engineering from Mansoura University, Mansoura, Egypt, in 2019, the Executive Master of Business Administration (M.B.A.) degree from Infrastructure University Kuala Lumpur, Malaysia, in 2021, and the master's degree in biomedical engineering specializing in integrating computer-aided diagnosis (CAD) systems with artificial intelligence (AI) from Mansoura University, in 2023. She was a Teaching Assistant with the Mathematics Department, Mansoura University, from 2020 to 2023. She is currently a Research Assistant in electrical engineering from Qatar University, Qatar, focusing on project QUHI-CENG-24/25-414. Her research interests include integrating electrical engineering with AI, specializing in image processing, machine learning, and deep learning to enhance system performance for battery management systems (BMSs).

KHALID ABUALSAUD (Senior Member, IEEE) received the B.Sc. degree in computer engineering from Al-Azhar University, Egypt, in 1988, and the M.S. and Ph.D. degrees in computer science from the School of Computing, University Utara Malaysia, in 2009 and 2015, respectively. He is currently a Lecturer of CE with the College of Engineering (CoE), Qatar University. He has authored more than 105 publications in conferences and journals. He is actively involved in many international research and industry-funded projects (over 1.5 million USD) encompassing contemporary issues related to his research interests. His research interests include health systems, wireless sensors for IoT applications, cybersecurity, cloud computing, and computer networks. He received several awards and four best paper awards from IEEE conferences, from 2019 to 2023. He has served as a TPC member and the chair for various reputable IEEE conferences. He served as a Guest Editor for special issue of *IEEE Network Magazine*. He is currently an Associate Editor of *IET Quantum Communication*.

AHMED M. MASSOUD (Senior Member, IEEE) received the B.Sc. (Hons.) and M.Sc. degrees in electrical engineering from Alexandria University, Egypt, in 1997 and 2000, respectively, and the Ph.D. degree in electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 2004. He is currently the Associate Dean for research and graduate studies with the College of Engineering, Qatar University, Qatar, and a Professor with the Department of Electrical Engineering, College of Engineering, Qatar University. He holds 12 U.S. patents. He supervised several M.Sc. and Ph.D. students with Qatar University. He received several research grants addressing research areas, such as energy storage systems, renewable energy sources, HVDC systems, electric vehicles, pulsed power applications, and power electronics for aerospace applications. He has published over 130 journal articles on power electronics, energy conversion, and power quality. His research interests include power electronics, energy conversion, renewable energy, and power quality.

• • •