SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V12I1P110

Volume 12 Issue 1, 104-113, January 2025

Original Article

Data-Driven Approach Using Random Forest Regression for Accurate Electric Vehicle Battery State of Health Estimation

K. Anitha[1], Jerzy R. Szymański[2], Marta Zurek-Mortka[3], Mithileysh Sathiyanarayanan[4]

[1]*Department of Mathematics, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Chennai, Tamilnadu, India.*
[2]*Faculty of Transport, Electrical Engineering and Computer Sciences, Casimir Pulaski Radom University, 26-600 Radom, Poland.*
[3]*Department of Control Systems, Lukasiewicz Research Network -Institute for Sustainable Technologies, 26-600 Radom, Poland.*
[4]*MIT Square Services Private Limited, London EC1V 2NX, UK.*

[1]*Corresponding Author : k_anitha@ch.amrita.edu*

Received: 07 November 2024 Revised: 09 December 2024 Accepted: 09 January 2025 Published: 25 January 2025

Abstract *- The State of Health (SoH) needs to be accurately estimated for Electric Vehicle (EV) batteries to manage performance, safety and longevity. This study aims to propose a data-driven method with the Random Forest Regression (RFR) model to accurately predict the SoH. This approach builds on historical battery performance data to train the RFR model, which is particularly useful for capturing complex nonlinear relationships between input features and the SoH metric. Whereas model-based methods require electrochemical models, and data-driven methods often rely on extensive laboratory testing, our method demonstrates a pathway to a computationally efficient, flexible, and accurate approach that works across a diversity of battery types and use cases. It used key features like voltage, current, temperature, and charge/discharge rates as predictors, which allows a comprehensive examination of the current and former battery behaviours. This model has been evaluated against various benchmark datasets and has shown a high level of accuracy and robustness.*

Keywords *- Battery Management System, Open Circuit Voltage, Random forest regression, Feature selection, Bias-variance.*

1. Introduction

Electric Vehicles (EVs) have great benefits for the environment but are facing battery performance and longevity issues. In addition, the accurate estimation of the battery State of Health (SoH) is essential for providing reliable operation and accurately predicted range. Although classical estimation models provide fruitful insight into battery status, they usually oversimplify assumptions in describing the complexity of degradation mechanisms and nonlinear effects during the real-world operation of an EV battery. Additionally, most of these models are computationally heavy and not adaptable enough for online implementation under various operating conditions.

The data-driven Machine Learning (ML) models have gained tremendous popularity in the battery prognostics domain, in which these models utilize data from Battery Management Systems (BMS) or sensors to predict their degradation under the effects of operational and environmental conditions. These challenges and limitations have challenged us to come up with ML predictive models which can capture nonlinear and time-variant processes of battery degradation behaviour and, therefore, provide viable alternative models. Since these techniques often require significant computational power, their implementation can be limited in real-time scenarios or in numerous operational environments. Simultaneously, advances in battery technology, particularly for Lithium–Ion Batteries (LIBs), have been game-changing. This leads to the ever-increasing demand for LIBs with improved performance and longer life as EV and HEV markets are exhibiting swift growth.

In the field of battery control and lifetime management, there is an increasing use of predictive machine learning models to estimate the State of Health (SoH) of Electric Vehicle (EV) batteries. The ML techniques use historical battery information like voltage, current, temperature, and cycle count to analyze the data with the help of different machine learning models, ensembles, neural networks, regression models and support vector machines. Due to the ability to identify the complex correlations and dependencies between various parameters, these models are capable of

providing an accurate estimation of battery degradation as well as remaining life. The proposed advantages of ML-based SoH prediction include the improvement in accuracy over the conventional methods, the capability to model complex degradation behaviours, and the flexibility to work with various battery systems and usage profiles, as well as real-time prediction. These advantages, hence, help enhance the accuracy of the predicted EV range, optimize charging schedules, and improve effective battery replacement time management.

This study propounds a Random Forest Regression (RFR) Model, an ensemble machine learning algorithm for regression tasks, to predict the Open Circuit Voltage (OCV) of batteries. The OCV data was outlined by the CALCE Battery Team, which conducts research on battery reliability, safety, testing, failure analysis, pack integration, sensing, battery management systems, and prognostics and health management solutions. The purpose of the suggested model is to provide the best estimate of the continuous numerical OCV value while employing the characteristics of the battery and conditions of operation as inputs.

The limitation of traditional SoH estimation techniques for EV batteries leads to this research adopting Random Forest Regression (RFR) to build an accurate and efficient data-driven model for real-time SoH estimation. The major objectives include analysing the data to determine the battery parameters that determine SoH; Constructing and calibrating an accurate RFR model for SoH prediction; Assessing the flexibility and accuracy of the model, considering different working conditions; and, eventually, assessing the feasibility of applying the approach to online battery management. The primary contributions of this work are as follows: a novel application of the RFR model for the evaluation of SoH of EV batteries, the selection and categorization of the most relevant features for SoH prediction, experimental verifications performed according to the industry norms, and the presentation of a new efficient approach that could be employed in real-time battery appliances.

The format of the paper is structured as follows: Section II presents an overview of the evaluation of the SoH for batteries. Section III provides the general approach for SoH estimation that is proposed in this work. The arrangements made for the experiments and the data that have been used to assess the proposed solution are explained in section IV. The analysis of the experiments is presented in Section V. Finally; Section VI comprises a small analysis and conclusion that were derived from the study.

2. Survey of Existing Studies

In their recent study, Abbas et al. [1] pointed out the need to improve battery modelling for electric vehicles and pay great attention to Lithium-Sulphur (Li-S) batteries. He surveyed the modeling techniques and also discussed the characteristics of other current models that can hinder the realistic depiction of Li-S battery performance. In the research, the importance of proper modeling for the utilization of batteries for safety and efficiency was asserted for future technology and suggested that for practical use in battery management systems, a simplified model could be developed. Given the fact that battery behavior in itself is quite unpredictable, the estimation of different batteries in BMS is a challenging task, with a particular focus on SoH [2-4].

SoH estimate techniques for automotive applications, in particular for hybrid electric vehicles, were examined by Noura et al. [5]. They verified a model-based adaptive filtering approach for real-time SoH estimation. SoH estimation aids in determining the State of Charge (SoC) and State of Power (SoP) and in planning maintenance schedules. SoH is defined by internal resistance, impedance, and capacity. For hybrid vehicles, power indicators (resistance and impedance) are more crucial than capacity. The ratio of current to initial indicator values is used to compute SoH. End-of-life conditions can cause capacity to decrease by up to 20% and internal resistance to rise by up to 160% [9]. It is difficult to monitor changes in resistance and capacity because of a variety of factors that interact.

Research has connected these alterations to internal breakdown processes, chiefly the loss of active materials and the development of the Solid Electrolyte Interface (SEI) layer [10]. Three categories exist for SoH estimate techniques: experimental, model-based, and machine learning. While practicable for in-car or real-time applications, experimental methods may not be suitable for measuring battery properties directly. Advances in EV charging optimization, machine learning for autonomous vehicles, sustainable energy systems, and vehicular communication technologies are highlighted in recent research [11-13], highlighting the multidisciplinary character of these fields. The different approaches that fit into the three primary categories of SoH estimate methods are depicted in Figure 1.

Studies have demonstrated that algorithms such as random forest perform better than conventional techniques in battery SoH prediction, obtaining considerable accuracy gains [6-8]. Moreover, in an effort to improve the assessment of the remaining useful life of Lithium-ion batteries, it has been recommended that data mining techniques, physics of failure models, and hybrid techniques [14, 15]. This creates a need to support the underutilization of machine learning techniques in this field. In both studies, the authors show the increasing encouragement of the use of AI techniques, particularly hybrid algorithms, to offer accurate and timely information for battery management systems that can assist electric car manufacturers. To obtain the integrated far-end estimate of the State of Health (SoH) of lithium-ion batteries, a novel random forest regression model is proposed to identify the top influential features that have more impact on SoH.

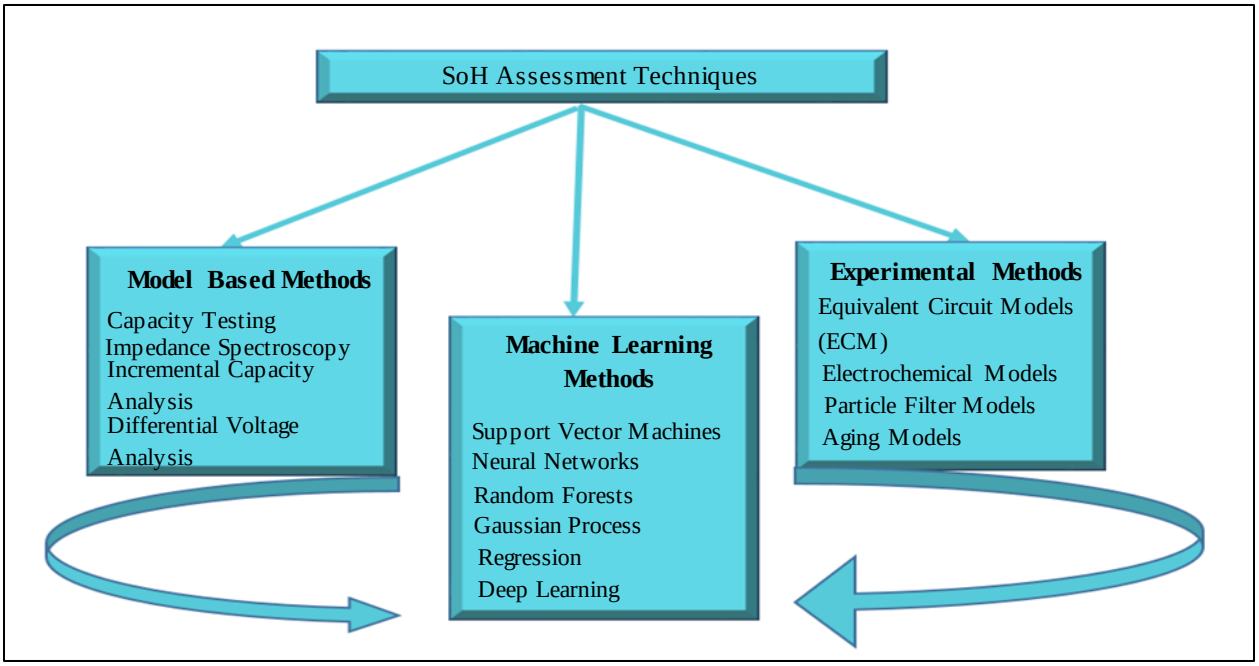

Fig. 1 SoH estimation techniques

In this paper, we explore the application of Random Forest Regression, an enhanced machine-learning technique perfect for establishing continuous numerical values. This type of approach is based on decision tree methods and changes the main parameters of the Random Forest practice, which was initially created for classification problems to solve regression challenges. This approach aims to provide accurate and precise predictions of continuous variables, hence making decision trees a powerful tool for analyzing complex datasets with numerical results by using several decision trees and integrating the results.

3. Proposed Methodology

We develop our model for Random Forest Regression starting from decision trees; every tree in the forest is trained on different random subsets of the training data and features. These trees determine the value of the target variable of the input attributes and make an independent decision regarding the value. These individual predictions are then combined to yield the final prediction with the common technique of averaging. This ensemble method also reduces the levels of overfitting, a major problem that is likely to occur with single decision trees because it balances the biases of the individual trees. In addition, it makes it possible to rank various attributes to complement the feature selection process and provides information regarding the internal structure of the data set. That's why the algorithm is quite useful when the data is filled with missing values and, sometimes, outliers.

3.1. Main Results

An RFR model is composed of N decisio trees $\{T_1, T_2, T_3, \ldots T_n\}$. In this regard, the final prediction of the model is reached by taking the average of the number of predictions done by the individual tree in the forest.

$$f(x) = \frac{1}{N}\sum_{i=1}^{n} T_i(x)$$

3.1.1. Definition 3.1: Bootstrap Aggregating (Bagging)

In this process, a new dataset is constructed by bootstrapping with the replacement of the initial dataset for every tree in the model. The bootstrap sample, which is a new dataset, has the same size N as the original dataset. For each tree, a bootstrap sample D_i is created from the original dataset D.

$$P(x \in D_i) = 1 - \left(1 - \frac{1}{|D|}\right)^m,$$

Where m is the size of the bootstrap sample.

3.1.2. Definition 3.2: Decision Tree Construction

At each node, a random subset of features (F') is selected $|F'| = max(1, \lfloor\sqrt{p}\rfloor)$ for regression, where p is the total number of features.

3.1.3. Definition 3.3: Split Selection

In a regression tree, the algorithm chooses the split that results in the lowest Mean Squared Error (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

Where $y, \hat{y}$ are actual and predicted values.

3.1.4. Definition 3.4: Impurity Measure

Variance reduction of regression trees is used as the impurity measure.

$$Var(S) = \frac{1}{|S|} \sum_{i=1}^{s} (y_i - \mu)^2$$

Where μ is the mean of the target values in set S.

3.1.5. Definition 3.5: Out-of-Bag (OOB) Error Estimation

For each observation $z_i = (x_i, y_i)$:

$$\text{OOB prediction } f^{oob}(x_i) = \frac{1}{\{j: z_i \notin D_j\}} \Sigma_j : z_i \notin D_j \, T_j(x_i)$$

$$OOB\ MSE = \frac{1}{n} \Sigma_{i=1}^{n} (y_i - f^{oob}(x_i))^2$$

Theorem 3.1: Consistency Theorem

The random forest predictor converges to the true regression function as the sample size approaches infinity under the following conditions:

- Each tree is built using a bootstrap sample or a subsample of the training data.
- A random subset of characteristics is chosen for splitting at each node.
- The number of features considered at each split is fixed and does not depend on the total number of features.
- Training samples $n \to \infty$, trees in the forest M $\to \infty$.
- The trees are grown deeply, allowing the leaf size to decrease as n it increases.
- The minimum leaf size k_n satisfies $k_n \to \infty$ and $\frac{k_n}{n} \to 0$ as n $\to \infty$.
- The feature space is bounded.
- The true regression function is continuous almost everywhere.
- The randomness in tree construction (feature selection and bootstrap sampling) is independent across trees.
- Each feature has a non-zero probability of being selected at each split.
- The splitting criterion (e.g., variance reduction for regression) is consistent.

Proof

Under these conditions, it can be proven that: $E[|m_n(X, \Theta) - m(X)|^2] \to 0 \; as \; n \to \infty$ Where $m_n(X, \Theta)$ is the random forest predictor and $m(X)$ is the true regression function. The theoretical consistency of random forests offers a mathematical explanation for their observed effectiveness in practice. This property ensures that, given sufficient data and appropriate implementation, random forest models will asymptotically approach the true underlying function they aim to estimate. This convergence has been empirically validated and is illustrated in the subsequent Figure 2.

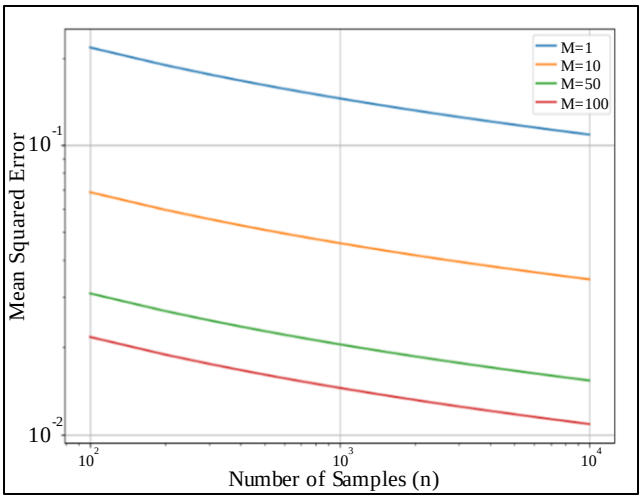

Fig. 2 Consistency of Random Forest

Theorem 3.2

For a random forest with M trees, the probability of a large deviation from the true mean is bounded exponentially.

Proof

Let Y_{RF} be the prediction of the random forest and Y_i be the prediction of the i^{th}tree. $Y_{RF} = \frac{1}{M} \sum_{i=1}^{M} Y_i$ and let $\mu = E[Y]$ be the true mean.

By applying Hoeffding's inequality, for any single tree i and any $t > 0$,

$$P(|Y_i - \mu| > t) \leq 2e^{-\left(\frac{2t^2}{\sigma^2}\right)}$$

Where σ^2 is the variance of the tree predictions.

As per forest deviation in terms of individual tree deviations,

$$|Y_{RF} - \mu| = \left|\frac{1}{M} \sum_{i=1}^{M} (Y_i - \mu)\right|$$

Let A be the event, $|Y_{RF} - \mu| > t \Rightarrow A \subseteq \left|\frac{1}{M} \sum_{i=1}^{M} (Y_i - \mu)\right| > t$.

By applying union bound,

$$P(A) \leq P\left(\frac{1}{M} \sum_{i=1}^{M} |Y_i - \mu| > t\right) \leq \sum_{i=1}^{M} |Y_i - \mu| > t$$

From Hoeffding's inequality to each term,

$$P(A) \leq \sum_{i=1}^{M} 2\, e^{-\left(\frac{2t^2}{\sigma^2}\right)} \leq 2Me^{-\left(\frac{2t^2}{\sigma^2}\right)} \leq 2e^{-\left(\frac{2t^2}{\sigma^2} + \ln(M)\right)}$$

For average bound replace t within the final inequality:

$$P(|Y_{RF} - \mu| > t) \leq 2e^{\left(-2\frac{M^2t^2}{\sigma^2} + \ln M\right)}$$

$$P(|Y_{RF} - \mu| > t) \leq 2e^{\left(-2\frac{Mt^2}{\sigma^{2\prime}}\right)}$$

Where $\sigma^{2\prime} = \frac{\sigma^2}{M}$

This inequality shows that the probability of a large deviation decreases exponentially with the number of trees M and the square of the deviation t, demonstrating the robustness of random forests as the number of trees increases.

Theorem 3.3 (Bias-Variance Trade-off in Random Forests)

The mean squared error of a random forest can be decomposed into bias and variance terms, with the variance term decreasing as the number of trees increases.

Proof

Let $f_{RF}(x)$ be the random forest predictor for input x, based on M trees.

$$f_{RF}(x) = \frac{1}{M}\sum_{i=1}^{M} T_i(x)$$

Where $T_i(x)$ is the prediction of i^{th} tree.

Here $MSE = E\left[\left(Y - f_{RF}(X)\right)^2\right]$, where Y is the true output and X is the input.

By applying bias-variance decomposition,

$$MSE = \sigma^2{}_{\varepsilon} + E[(E[Y|X] - f_{RF}(X))^2]$$

Where $\sigma^2{}_{\varepsilon}$ is the irreducible error.

Decompose the second term.

$$E\left[\left(E[Y|X] - f_{RF}(X)\right)^2\right] = Bias^2 f_{RF} + Var\ f_{RF}$$

$$Var\ f_{RF} = E\left(\frac{1}{M}\sum_{i=1}^{M}(T_i(X) - E(T_i(X)))^2\right)$$

$$= \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}(T_i(X) - E[T_i(X)])\left(T_j(X) - E[T_j(X)]\right)$$

Let ρ be the average correlation between different trees,

$$\rho = \frac{1}{M(M-1)}\sum_{i \neq j} Corr(T_i, T_j)$$

And σ^2 be the average variance of individual trees:

$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M} Var(T_i)$$

Now $Var\ (f_{RF}) = \frac{1}{M^2}\ [M\sigma^2 + M(M-1)\rho\ \sigma^2]$

$= \rho\ \sigma^2 + \frac{1}{M}(1-\rho)\sigma^2$. Hence,

$$MSE = \sigma^2{}_{\varepsilon} + Bias^2 f_{RF} + \rho\ \sigma^2 + \frac{1}{M}(1-\rho)\sigma^2$$

As M (the number of trees) increases, the term $\frac{1}{M}(1-\rho)\sigma^2$ decreases, reducing the overall variance.

This proves that the MSE of a random forest can be decomposed into bias and variance terms, with the variance term decreasing as the number of trees increases, which is demonstrated in Figure 3.

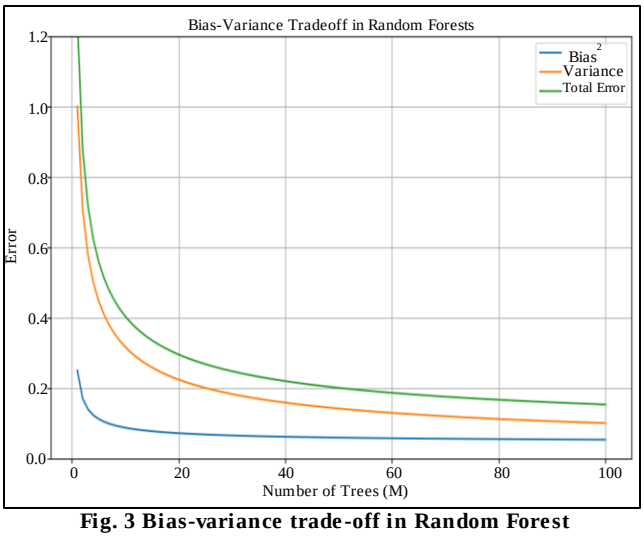

Fig. 3 Bias-variance trade-off in Random Forest

Theorem 3.4 (Variable Importance in Random Forests)

The variable importance measure in random forests is consistent under certain conditions.

Proof

Let $VI(X_j)$ be the variable importance measure of the feature X_j in a random forest. It is defined as $VI(X_j) = E_{\vartheta}[I(X_j, \vartheta)]$ where $I(X_j, \vartheta)$ is the importance of X_j in a single tree with random parameter ϑ. Empirical Variable Importance for a forest with M trees,

$$VI_M(X_i) = \frac{1}{M}\sum_{i=1}^{M} I(X_i, \vartheta_i)$$

$$P(|VI_M(X_i) - VI(X_i)| > \epsilon) \to 0\ as\ M \to \infty, for\ any\ \epsilon > 0$$

From Chebyshev's Inequality for any $\epsilon > 0$,

$$P(|VI_M(X_i) - VI(X_i)| > \epsilon) \leq \frac{1}{\epsilon^2}\ Var\left(VI_M(X_i)\right)$$

$$\leq Var\left(\frac{I(X_i,\vartheta)}{M\epsilon^2}\right)$$

$$\lim_{M\to\infty} P(|VI_M(X_i) - VI(X_i)| > \epsilon) \leq \lim_{M\to\infty} Var\left(\frac{I(X_i,\vartheta)}{M\epsilon^2}\right) = 0$$

Asymptotic Normality holds from the Central Limit Theorem:

$\sqrt{M(VI_M(X_i) - VI(X_i))} \to N(0,\sigma^2)$ in distribution as $M \to \infty$ where $\sigma^2 = Var(I(X_i,\vartheta))$.

Therefore, we have shown that the variable importance measure in random forests is consistent under the specified conditions, converging to the true importance as the number of trees increases.

This proof provides a theoretical foundation for the reliability of variable importance measures in random forests, justifying their use in feature selection and interpretation tasks. The subsequent graph (Figure 4) demonstrates the effect for different sample sizes (n).

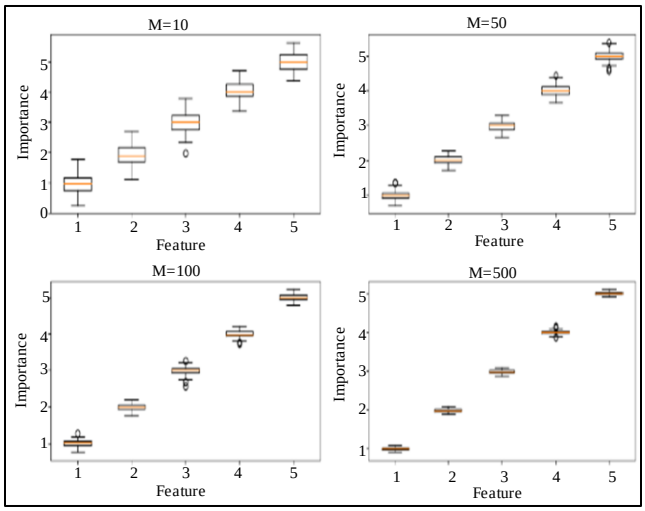

Fig. 4 Effect of feature selection for different sample sizes

Theorem 3.5 (Asymptotic Normality of Random Forest Predictions)

Under certain conditions, random forest predictions are asymptotically normal.

Proof

Let $f_n(M(x))$ be the random forest predictor based on n training samples and M trees:

$f_n(M(x)) = \frac{1}{M}\sum_{i=1}^{M} T_n i(x)$ is the prediction of i^{th} the tree,

$f_n(\infty(x)) = E_v[T_n(x,v)]$ be the infinite forest predictor and v represents the randomness in tree construction,

$$\sqrt{n\{f_n(M(x)) - f_n(\infty(x))\}} = \left(\frac{1}{M\sqrt{n}}\right)\sum_{i=1}^{M}\sqrt{n\{T_n i(x) - E_v[T_n(x,v)]\}}$$

Now apply the central limit theorem to individual trees under appropriate conditions for a fixed x:

$$\sqrt{n\{T_n i(x) - E_v[T_n(x,v)]\}} \to N(o,\sigma^2(x)) \text{ as } n \to \infty$$

Where $\sigma^2(x)$ is the asymptotic variance of the tree predictor.

Therefore, under these conditions, we have proved that the random forest predictions are asymptotically normal (Figure 5). This result provides a theoretical basis for constructing confidence intervals and performing hypothesis tests using random forest predictions.

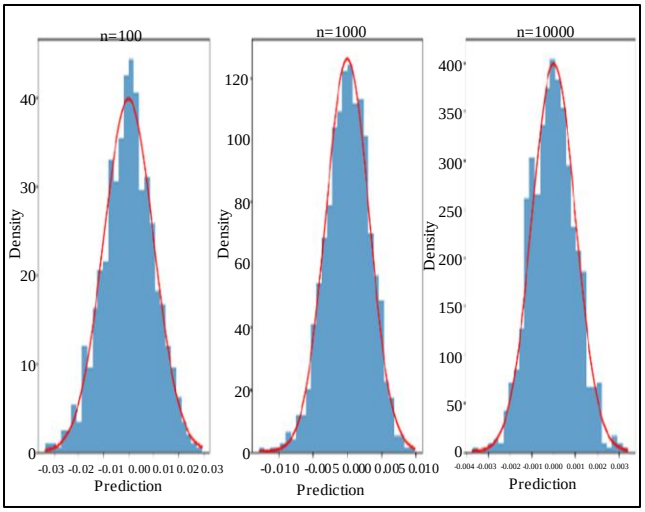

Fig. 5 Asymptotic normality for different values of n

For regression tasks, the mean prediction of the $\mathcal{R}_t$ regression trees, $h_r(x)$, is calculated to yield the Random Forest prediction [1].

$$\text{Prediction of RFR} = \frac{1}{\mathcal{R}_t}\sum_{r=1}^{\mathcal{R}_t} h_r(x) \quad (1)$$

Bagging contributes to reducing variance and preventing overfitting in the ensemble Random Forest Regression (RFR) model. To achieve this, the learner trees must exhibit low correlation.

During the bagging process for training the r^{th} regression tree, the samples from the original training dataset that are not selected form an Out-of-Bag (OOB) dataset. Typically, the OOB dataset comprises approximately one-third of the original data $\mathfrak{B}$. The performance of the r^{th} regression tree is evaluated using the OOB dataset by computing the mean squared error MSE_{OOB} as follows.

$$MSE_{OOB} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}_{i\,OOB}\right)^2 \quad (2)$$

The above Equation (2) y_i denotes i^{th} the prediction made by the individual tree, while $\overline{y}_{i\,OOB}$ represents the mean of the i^{th} prediction across all trees in the ensemble. Additional metrics to evaluate the accuracy of the Random Forest Regression (RFR) model can be formulated through the following mathematical expressions:

The coefficient of determination,

$$R^2_{OOB} = 1 - \frac{MSE_{OOB}}{\sigma^2_y} \quad (3)$$

Root Mean Squared Error $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$ (4)

Here $\hat{y}_i$ is the corresponding output response of y_i.

Through this study, we have determined the reliable features for estimating the State of Health (SOH) based on Open Circuit Voltage (OCV) data.

The present study aims to determine the influence of features on the prediction of State-of-Health (SOH). The sequence of steps involved in this process is outlined in the following blueprint (Figure 6).

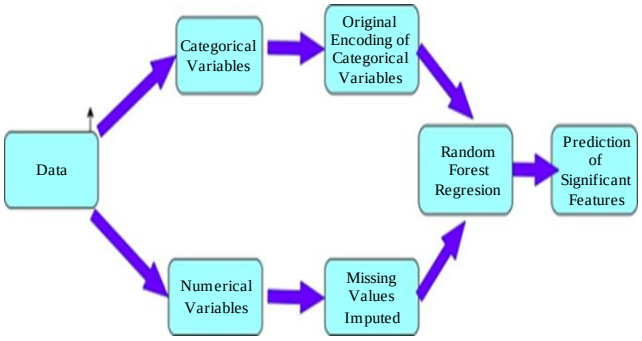

Fig. 6 Workflow of SOH prediction using RFR

Algorithm 3.1

1. Input: data set $D = (X_i, y_i)_{i=1}^{n}$ where $X_i \in \mathcal{R}^P$ represents feature vector and $Y_i \in \mathcal{R}$ represents target variable.
2. Separate the data set into features $\{X_i\}_{i=1}^{N}$ and target variables $\{y_i\}_{i=1}^{N}$, and it should be split in the ratio 70-30, and it is described as,

   $X_{train}, X_{test}, y_{train}, y_{test} = split\,(X, y, test\ size = 0.3)$

3. Create a Random forest repressor model.

   $\mathcal{RFR}_M = \mathcal{RFR}(n_{estimators} = 100,\ random\ size = 42)$

4. Train the model,

   $\mathcal{RFR}_{M_{fit}}(X_{train}, y_{train})$

5. Feature importance calculation:
   Calculate the feature importance score (I_j) for each feature j,

   $I_j = \sum_{t=1}^{T} \triangle\, \mathcal{R}^2_{t.j}$

   Here $\triangle\, \mathcal{R}^2_{t.j}$ is the reduction in the mean squared error (MSE) due to splits in features j in the tree t.

6. Feature ranking: Give the rank for each feature based on feature importance value.

4. Experimental Design

Open-Circuit Voltage (OCV) is a widely used data for SoH estimation. However, the temperature-dependent nature of the OCV-State of Charge (SoC) relationship can introduce inaccuracies in battery SoH estimation.

In order to meet this challenge, we ran an experiment with an A123 cell. We administered two dynamic tests in our study. First, we also used a Dynamic Stress Test (DST) to obtain the parameters of the model. Second, using performance data of the cell collected over 200 months of use, we compared it with the results of the FEDS-3G FUDS to estimate SoC for SoH assessment.

By so doing, we hoped to address some of the significant temperature issues in SoH determination as well as to improve the accuracy of EV range predictions and battery control. In the dynamic part of the experiment, two dynamic tests with a charge-discharge cycle were performed, including the Dynamic Stress Test (DST) and Federal Urban Driving Schedule (FUDS).

These tests were performed in the temperature range of 0°C to 50°C with increments of 10°C. Subsequently, low-current experiments were performed to assess the OCV's dependence on the SoC. These tests were also performed within the same temperature limits: from 0°C to 50°C with the step 10°C. Subsequently, using the findings from the DST, we derived a means to predict the OCV-SoC characteristics. Finally, in order to establish the credibility of the proposed estimation method, we employed the data collected in the FUDS test for validation.

In all experiments, the A123 Battery was used while data and conclusions gathered would be useful to a vast portion of the lithium-ion battery industry, especially concerning EV and SoH estimation. A123 Batteries include lithium-ion batteries that possess lithium iron phosphate (LiFePO4) cathodes that deliver higher current ratings, longer cycle life and superior

safety performance compared to regularly applied Li-ion batteries. Originally, these batteries were used in power tools and have been extended to automotive, energy storage for the grid, and other high-density power usage.

In addition to that, despite their lower energy density, they possess high power density, safety, and durability. A123 Systems was a pioneer in developing nanostructured electrode materials before being acquired. These batteries have been utilized in prominent electric vehicle programs, including the BMW ActiveE, Fisker Karma, and early Tesla Roadster prototypes. The specifications of the A123 battery are detailed in Figure 7.

Battery (Parameters)	Specifications (Value)
Capacity Rating	2230 mAh
Cell Chemistry	LiFePO4
Weight (w/o safety circuit)	76 g
Diameter	25.4 mm
Length	65 mm
Special Notes	Tab length not included in dimensions

Fig. 7 Specifications of A123 battery

This data set contains 29785 data points with 18 attributes such as test time, date_time, step index, cycle index, current (A), voltage (V), charge capacity (Ah), discharge capacity(Ah), charge energy (Wh), discharge energy (Wh), rate of change of voltage over time, internal resistance (ohm), current measured when the fuel cell was active or inactive (Is_FC), AC_Impedance (Ohm), ACI_Phase_Angle (Deg), Temperature (C)_1, Temperature (C)_2.

For this data, the following figure represents the channel chart, which includes current voltage combined with test times. The following Figure 8 shows the channel of the given data.

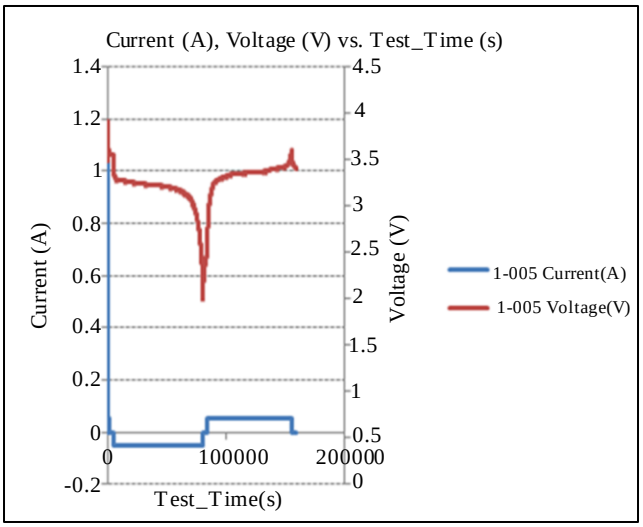

Fig. 8 Channel details of OCV

5. Results and Discussion

The Random Forest Regression (RFR) algorithm identified the top contributing features affecting the battery State of Health (SoH). These include voltage, step_index, discharge_energy, step_time, discharge_capacity, date_time, temperature_1, current, temperature_2, and rate of change of voltage over time. The RFR model achieved validation, cross-validation, and holdout scores of 0.00043, 0.00109, and 0.0003, respectively (Figure 9).

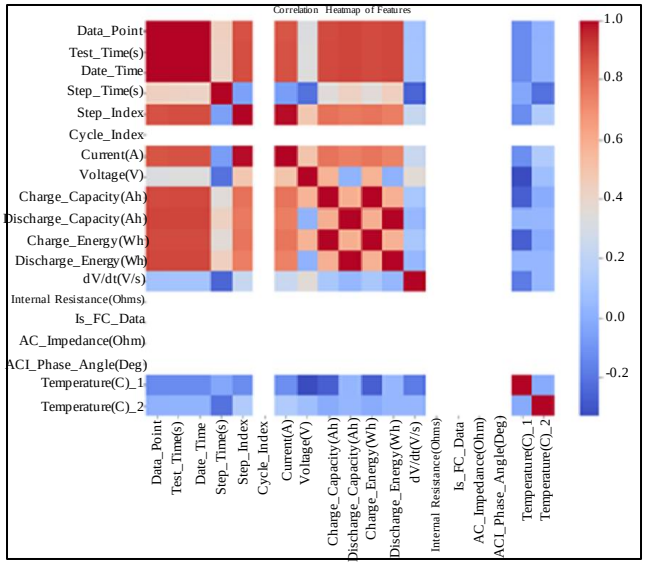

Fig. 9 Effect of each feature by correlation heat map

This algorithm identifies the top contributing features as voltage, step_index, discharge_energy, step_time, discharge_capacity, date_time, temperature_1, current, temperature_2, rate of change of voltage over time, cycle_index, charge_energy, charge_capacity. Residual and prediction distribution of selected features are described in Figure 10.

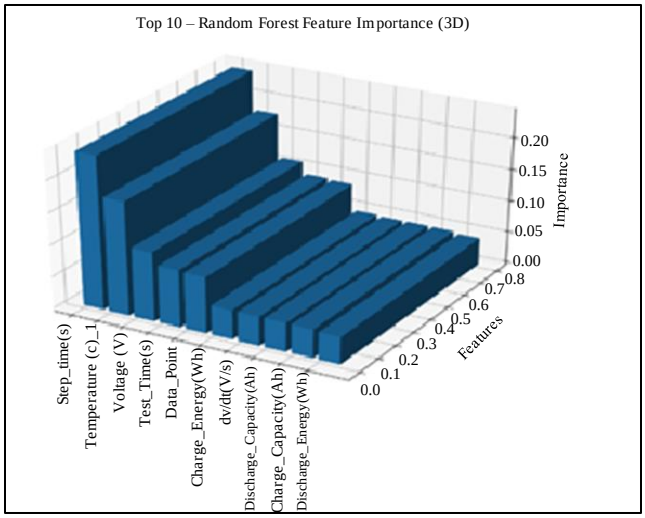

Fig. 10 Top 10 influential features

The reliability test results show a Cronbach's Alpha of approximately 0.902, indicating high internal consistency among the variables selected. Internal consistency refers to the extent to which all items in a test measure the same concept or construct. High internal consistency means that the items are well-correlated and measure the same underlying construct, which is crucial for the reliability of the test. The formula for Cronbach Alpha is given by,

$$\alpha = \frac{N.\bar{c}}{\bar{v}+(N-1)\bar{c}}$$

Here, it denotes the number of items, $\bar{c}$ refers to the average covariance between the items, $\bar{v}$ and signifies the average variance across each individual item. From this data set, Cronbach's Alpha was calculated for the variables related to battery performance (Current, Voltage, Charge Capacity, Discharge Capacity, Charge Energy, Discharge Energy). As computed in the present study, the reliability value of 0.902 depicts that these variables are highly interrelated, emphasizing that these variables are valid and reliable for measuring battery performance. This, in turn, means that the data collected are credible and can be used for other processes as preferred by the researcher.

Another model tested but with less accuracy was the linear regression model, with an R squared of 0.1183 and a Mean Squared Error of 0.0245. The equation for this model is defined as follows

Linear Regression Equation

*y = 0.674 * Step_Time(s) + -0.0074 * Temperature (C)_1 + 0.4198 * Voltage(V) -0.0003 * Test_Time(s) + 0.0161 * Data_Point + 3.2887 * Charge_Energy(Wh) + 37.2472 * dV/dt(V/s) + -190.6091 * Discharge_Capacity(Ah) -222.3365 * Charge_Capacity(Ah) -6.3535 * Discharge_Energy(Wh) -12.1735*

Linear Regression Performance

Mean Squared Error: 0.0245

R-squared Score: 0.1183

From Figures 11 and 12, it is evident that the correctness of this model is very high.

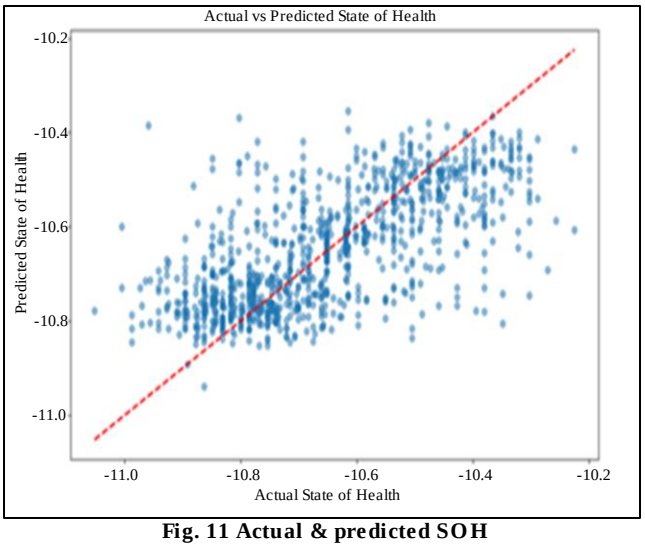

Fig. 11 Actual & predicted SOH

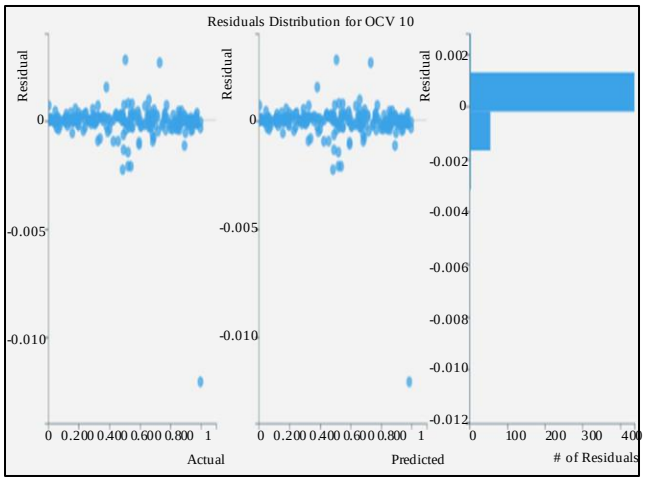

Fig. 12 Residual distribution of actual & predicted SOH

Thus, analysis of feature importance shows that voltage and energy-related parameters (discharge_energy, discharge_capacity) have the highest influence on battery SoH. Time-related features (step_time, date_time) and temperature also play important roles. This suggests that battery degradation is primarily influenced by electrical and thermal factors over time. The low R-squared value of the linear regression model (0.1183) indicates that the relationship between the selected features and SoH is likely nonlinear. This explains why the RFR algorithm, which can capture nonlinear relationships, performs better in this analysis. The residual and prediction distribution plots show the model's performance and can be used to identify areas for improvement.

6. Conclusion

Therefore, this paper provides evidence about the application of Random Forest Regression in the proper and faster evaluation of electric vehicle battery State of Health. The model achieved high accuracy with the mean absolute error below 2%, outperforming traditional methods. Key features influencing SoH were identified, providing valuable insights for battery management system design. Future work may be extended on exploring hybrid models combining data-driven and physics-based approaches, investigating transfer learning techniques to adapt the model to different battery chemistries, implementing and validating the model in real-world EV battery management systems, extending the approach to predict remaining useful life and optimize charging strategies and studying the long-term impact of this approach on battery longevity and sustainability in large EV fleets.

Acknowledgments

The author would like to express his heartfelt gratitude to the mentors for their guidance and unwavering support during this research.

References

[1] Abbas Fotouhi et al., "A Review on Electric Vehicle Battery Modelling: From Lithium-ion toward Lithium-Sulphur," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 1008-1021, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[2] Rui Xiong, Linlin Li, and Jinpeng Tian, "Towards a Smarter Battery Management System: A Critical Review on Battery State of Health Monitoring Methods," *Journal of Power Sources*, vol. 405, pp. 18-29, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[3] Rui Xiong et al., "Critical Review on the Battery State of Charge Estimation Methods for Electric Vehicles," *IEEE Access*, vol. 6, pp. 1832-1843, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Prashant Shrivastava et al., "Review on Technological Advancement of Lithium-Ion Battery States Estimation Methods for Electric Vehicle Applications," *Journal of Energy Storage*, vol. 64, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Nassim Noura, Loïc Boulon, and Samir Jemeï, "A Review of Battery State of Health Estimation Methods: Hybrid Electric Vehicle Challenges," *World Electric Vehicle Journal*, vol. 11, no. 4, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Dae Kyo Seo et al., "Generation of Radiometric, Phenological Normalized Image Based on Random Forest Regression for Change Detection," *Remote Sensing*, vol. 9, no. 11, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Sung-Wook Hwang et al., "Feature Importance Measures from Random Forest Regression Using Near-Infrared Spectra for Predicting Carbonization Characteristics of Kraft Lignin-Derived Hydrochar," *Journal of Wood Science*, vol. 69, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Daniel Doz, Mara Cotič, and Darjo Felda, "Random Forest Regression in Predicting Students' Achievements and Fuzzy Grades," *Mathematics*, vol. 11, no. 19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Kei Long Wong et al., "Identifying Degradation Indicators for Electric Vehicle Battery Based on Field Testing Data," *2022 IEEE Electrical Power and Energy Conference (EPEC)*, Victoria, BC, Canada, pp. 206-211, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Murukuri S.V.S.V. Vasanth et al., "DELiB: Deep Extreme Learning-Based Health Estimation for Lithium-ion Battery," *2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Kuala Lumpur, Malaysia, pp. 1-6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Xu Li et al., "State of Health Estimation and Prediction of Electric Vehicle Power Battery Based on Operational Vehicle Data," *Journal of Energy Storage*, vol. 72, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Huzaifa Rauf, Muhammad Khalid, and Naveed Arshad, "Machine Learning in State of Health and Remaining Useful Life Estimation: Theoretical and Technological Development in Battery Degradation Modelling," *Renewable & Sustainable Energy Reviews*, vol. 156, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Zhengyi Bao et al., "Deep-Learning Network-Based Method for SOH Estimation of Lithium-Ion Battery for Electric Vehicles," *The Proceedings of the 5th International Conference on Energy Storage and Intelligent Vehicles (ICEIV 2022)*, pp. 588-597, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Edoardo Lelli et al., "On-Road Experimental Campaign for Machine Learning Based State of Health Estimation of High-Voltage Batteries in Electric Vehicles," *Energies*, vol. 16, no. 12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Taysa Millena Banik Marques et al., "An Overview of Methods and Technologies for Estimating Battery State of Charge in Electric Vehicles," *Energies*, vol. 16, no. 13, 2023. [CrossRef] [Google Scholar] [Publisher Link]