Green Energy and Intelligent Transportation 3 (2024) 100177

Contents lists available at ScienceDirect

Green Energy and Intelligent Transportation

journal homepage: www.journals.elsevier.com/green-energy-and-intelligent-transportation

Full length article

State of charge estimation for electric vehicles using random forest

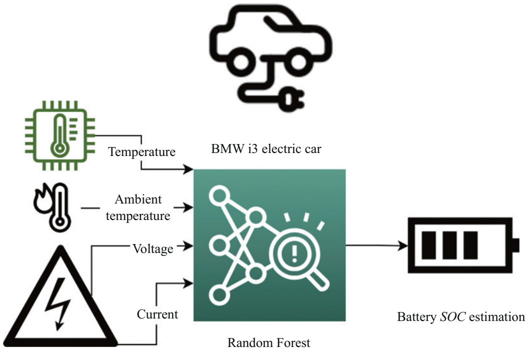
Mohd Herwan Sulaiman [a,*], Zuriani Mustaffa [b]

[a] Faculty of Electrical & Electronics Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), 26600 Pekan Pahang, Malaysia
[b] Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), 26600 Pekan Pahang, Malaysia

HIGHLIGHTS

- An approach utilizing Random Forest for accurate state of charge estimation for electric vehicle.
- The model leverages real driving trips from a BMW i3 EV as training and testing data.
- Comparative analysis demonstrates the superior performance of RF model in addressing SOC estimation.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Electric vehicles
Extreme learning machine
Machine learning
Random Forest
State of charge of battery

ABSTRACT

This paper introduces an innovative approach to addressing a critical challenge in the electric vehicle (EV) industry—the accurate estimation of the state of charge (SOC) of EV batteries under real-world operating conditions. The electric mobility landscape is rapidly evolving, demanding more precise SOC estimation methods to improve range prediction accuracy and battery management. This study applies a Random Forest (RF) machine learning algorithm to improve SOC estimation. Traditionally, SOC estimation has posed a formidable challenge, particularly in capturing the complex dependencies between various parameters and SOC values during dynamic driving conditions. Previous methods, including the Extreme Learning Machine (ELM), have exhibited limitations in providing the accuracy and robustness required for practical EV applications. In contrast, this research introduces the RF model, for SOC estimation approach that excels in real-world scenarios. By leveraging decision trees and ensemble learning, the RF model forms resilient relationships between input parameters, such as voltage, current, ambient temperature, and battery temperatures, and SOC values. This unique approach empowers the model to deliver precise and consistent SOC estimates across diverse driving conditions. Comprehensive comparative analyses showcase the superiority of the RF over ELM. The RF model not only outperforms in accuracy but also demonstrates exceptional robustness and reliability, addressing the pressing needs of the EV industry. The results of this study not only underscore the potential of RF in advancing electric mobility but also suggest a promising integration of the SOC estimation approach into the battery management system of BMW i3. This integration holds the key to more efficient and dependable electric vehicle operations, marking a significant milestone in the ongoing evolution of EV technology. Importantly, the RF model demonstrates a lower Root Mean Squared Error (RMSE) of 5.902,8% compared to 6.312,7% for ELM, and a lower Mean Absolute Error (MAE) of

* Corresponding author.
E-mail address: herwan@umpsa.edu.my (M.H. Sulaiman).

https://doi.org/10.1016/j.geits.2024.100177
Received 2 October 2023; Received in revised form 5 January 2024; Accepted 13 January 2024
Available online 19 January 2024

4.432,1% versus 5.111,2% for ELM across rigorous k-fold cross-validation testing, reaffirming its superiority in quantitative SOC estimation.

1. Introduction

The accurate estimation of the State of Charge (SOC) is crucial for the efficient and reliable operation of electric vehicles (EVs). Notably, the increased utilization of electrified vehicles can be attributed to several contributing factors. These encompass government policies, emissions-reduction legislation, rising fuel costs, heightened environmental awareness, and tax credits for electric vehicle manufacturers and users. As these factors synergistically shape the landscape of electrified transportation, the necessity of precise state of charge estimation in EV battery systems becomes paramount [1]. SOC estimation refers to determining the remaining energy in the battery, which is essential for optimizing battery usage, range estimation, and ensuring the longevity of the battery. However, SOC estimation is a challenging task due to the complex and nonlinear nature of battery behavior [2]. Specifically, the electrochemical reactions within lithium-ion batteries exhibit hysteresis, meaning the voltage-SOC relationship depends on the direction of current flow, i.e. charging vs discharging [3]. The operating temperature also significantly impacts the open circuit voltage, further complicating the correlation between measured voltage and SOC [4]. Additionally, lithium-ion batteries possess a flat open circuit voltage plateau region from 10% to 90% SOC where the voltage remains nearly constant, making it difficult to precisely distinguish SOC in this range [5]. Moreover, as the battery ages, its capacity degrades which must be tracked to maintain accuracy [6]. These multifaceted factors have hindered the development of robust and reliable SOC estimation techniques that can deliver high accuracy across the wide array of EV operating conditions.

Ref. [7] introduces advanced modeling methods such as an improved anti-noise adaptive long short-term memory (LSTM) neural network to predict the remaining useful life of lithium-ion batteries. An improved robust multi-time scale singular filtering-Gaussian process regression-long short-term memory (SF-GPR-LSTM) remaining capacity estimation has been proposed in Ref. [8]. These methods employ innovative techniques to enhance the accuracy and robustness of the rapid battery performance evaluation, especially for lithium-ions batteries. While these innovative techniques have enhanced battery modeling, their ability to provide highly accurate SOC estimates under dynamic real-world driving conditions has yet to be determined.

In recent years, machine learning approaches have garnered substantial attention for SOC as well as in state of health (SOH) estimation in EVs [2,9]. Data-driven models have proven powerful for extracting intricate relationships between battery parameters critical to SOC estimation. Meanwhile, techniques like gradual decreasing current, double correlation analysis, and gated recurrent units (GRU) show promise for enhancing SOH estimation performance. The complementary nature of these approaches highlights the potential of machine learning to uncover insights needed to advance both SOC and SOH estimation for electric vehicle batteries. The insights gained from Ref. [10], which focuses on the importance of sensing systems in accurately monitoring parameters in new energy storage devices, and Ref. [11], which delves into the relationship between electrochemical impedance spectroscopy and the mechanism of capacity decline in lithium-ion batteries, further support the exploration of advanced methodologies and technologies. These will enhance the understanding of energy storage devices and their applications in the context of electric vehicle systems. Within the realm of machine learning algorithms, deep learning techniques [12–18] and Random Forest (RF) [19,20] have emerged as promising candidates, effectively addressing numerous real-world challenges [21–26].

Deep learning algorithms, such as deep neural networks, have been prominently featured in SOC estimation. For instance, Zhang et al. [27] proposed a deep neural network-based approach for SOC estimation in Li-ion batteries, showcasing its prowess in capturing intricate battery dynamics and achieving remarkable accuracy. SOC estimation also has been solved by using the well-known Kalman Filter (KF) approaches with various adaptive and variants of KF that have been presented in literature such as Adaptive Extended Kalman Filter (AEKF) [28], square root unscented KF [29], modified extended KF (MEKF) [30], and Affine Iterative Adaptive Extended Kalman Filter (AIAEKF) [31]. However, KF and its variants excel in pattern recognition for estimation and prediction by relying on any particular model definition of a process and measurement model that fail to capture complex relationships within the data. In contrast to Kalman filter (KF) approaches, deep learning algorithms, such as artificial neural networks (ANNs) and LSTM networks, have demonstrated exceptional promise in SOC estimation for EV batteries. Their ability to discern intricate relationships between input parameters and SOC values positions them as invaluable tools in tackling the non-linear and dynamic nature of SOC estimation, particularly in complex driving conditions [32,33]. While deep learning algorithms often surpass KF methods in terms of accuracy, it is important to note that they may require more computational resources and training data [34]. A compelling avenue for improvement lies in hybrid approaches that combine the strengths of both deep learning algorithms and Kalman filters to enhance SOC estimation accuracy and robustness [35]. The choice between these techniques depends on the specific application and available resources.

On the other hand, RF, classified as an ensemble learning algorithm that combines multiple decision trees to make predictions [36], has earned recognition for its versatility across various domains, including battery SOC estimation. Researchers have extensively explored the implementation of RF in SOC estimation for EV batteries. A notable example is found in Ref. [37], where a random forest regression technique was proposed for real-time capacity estimation of Li-ion batteries. This work underscored the potential of RF in concurrent SOC and battery capacity estimation.

The use of RF in SOC estimation is motivated by its ability to handle large datasets, robustness to noise, and feature importance analysis. RF can effectively capture the nonlinear relationships between battery parameters and SOC, leading to accurate estimation results. Moreover, RF offers interpretability, allowing researchers to analyze the importance of different features in the estimation process. It is important to emphasize that the utilization of RF in various applications makes it a promising approach to consider for addressing the SOC estimation problem. RF has been utilized in energy consumption prediction [38], electricity theft detection [39], season-based occupancy prediction problems [40], geochemical anomalies [41], internet of things (IoT) [42], hydrogeochemical and sediment parameters prediction [43], determination of key factors affecting the substructure of ballast-less railway track under moving load [44], predicting the utilization factor of blasthole in rock roadways [45], CO_2 emission forecasting [46], milling chatter identification [47], demand forecasting of spare parts [48] and many more.

In this paper, the application of RF in the estimation of SOC for BMW i3 electric vehicles is presented. This paper aims to address the challenges associated with SOC estimation and leverage the capabilities of RF to achieve accurate and reliable results. By utilizing real-world data from the measurement of 70 trips of BMW i3 EV [49,50], the performance of RF in SOC estimation will be investigated. Additionally, in order to show the effectiveness of the developed RF model, the comparison with other machine learning approach, namely Extreme Learning Machines (ELM) also will be performed. Overall, this paper aims to contribute to the field of SOC estimation in EVs by exploring the application of RF and comparing it with other machine learning approach. The findings will provide insights into the performance and suitability of RF for SOC estimation in BMW i3 electric vehicles, contributing to the development of efficient and reliable battery management systems.

The remaining sections of the paper are structured as follows: Section 2 provides a concise overview of RF, while Section 3 explores the application of RF for the SOC estimation model. Section 4 presents the results and subsequent discussion, and lastly, Section 5 presents the paper's conclusion.

2. Random Forest (RF)

RF represents a versatile machine learning algorithm widely employed for classification and regression tasks. In this study, the exceptional capability of RF model for regression is leveraged to estimate the SOC of electric vehicle batteries accurately. RF, introduced by Leo Breiman in 2001 [36], stands as an ensemble method that combines multiple decision trees to yield robust and precise predictions. Each decision tree is trained on a distinct subset of the training data, and the final prediction results from aggregating the outputs of all individual trees.

RF employs the bagging method, which entails training each decision tree on a random subset of the training data with replacement. This strategy effectively reduces overfitting and enhances model generalization [51]. RF boasts several advantages, making it well-suited for the SOC estimation task at hand. It excels with large, high-dimensional datasets, handles noisy data, and exhibits robustness to outliers and missing values. The algorithm's ability to rank feature importance aids in identifying crucial parameters for prediction.

One of the critical hyperparameters of RF is the number of trees in the forest, which significantly impacts model performance. For the SOC estimation task, hyperparameter tuning was conducted to select the optimal configuration. While there is no one-size-fits-all answer for choosing the number of trees, it is recommended to consider factors such as training data size and feature dimensions when making this choice.

3. Application of RF for SOC estimation problem

This paper introduces an approach to estimate the SOC of an electric vehicle battery in real-world conditions, utilizing the ensembled machine learning method known as RF. The quality of data used to train the RF model is crucial in assessing its effectiveness. The dataset should contain relevant domain information and avoid noise or irrelevant data, as RF is a data-driven technique. However, measurement noise and errors are often unavoidable and should be considered during the training and testing of the RF model. It is important to acknowledge that these factors can affect the accuracy of the model and should be taken into account.

In this research, the simulation experiments will make use of an actual dataset comprising 70 journeys made by a BMW i3 electric vehicle [50], which is equipped with a 60 Ah battery pack. This data has been gathered using electric vehicle sensors installed on the car and is sampled at a 1Hz rate through the OBD port. It is important to mention that this dataset might contain missing values (*NaN*) due to measurement errors or other factors. Consequently, preliminary processing steps are imperative to purify the original data. Within the dataset, two distinct State of Charge (SOC) attributes are present: one estimated by the electric vehicle manufacturer and the other directly displayed to the end user. For the purposes of this research, the SOC estimated by the electric vehicle manufacturer is selected as the target variable for training and evaluating the RF model. This choice aligns with the commitment to closely replicate the manufacturer's insights, effectively mirroring their estimations. It is essential to highlight that the manufacturer's estimated SOC, employed as the target variable, corresponds to the real SOC recorded for all trips, serving as a validated and trusted reference point within the electric vehicle domain. On the other hand, the measured voltage, current, temperature of the battery pack, and ambient temperature are used as input for the RF.

To ensure accurate and reliable SOC estimation, thorough data preprocessing will be conducted to handle missing values (*NaN*) and eliminate any potential noise or inconsistencies in the dataset. The utilization of real-world data from the BMW i3 EV, coupled with the ability of RF to adapt the varying data characteristics, is expected to yield robust and meaningful SOC predictions, contributing to the advancement of battery state estimation in electric vehicles.

In the development of the SOC estimation model, a deliberate choice was made to incorporate ambient temperature as one of the input variables. This decision was motivated by the recognition of the critical role that temperature plays in the behavior and performance of lithium-ion batteries, which are widely used in electric vehicles. Ambient temperature is a pivotal factor that influences battery capacity, charge/discharge rates, and overall health. In real-world electric vehicle applications, temperature variations are common, and vehicles are exposed to diverse environmental conditions. By including ambient temperature in the model, the aim is to account for these temperature-related variations and create a robust SOC estimation framework that can adapt to different climates and scenarios. Furthermore, this choice aligns with the practicality of electric vehicle operations, as ambient temperature is a parameter readily available for measurement or estimation. In essence, the decision to consider ambient temperature as an input variable stems from its substantial impact on battery performance and its importance in enhancing the accuracy and reliability of SOC estimation in the context of electric mobility. Fig. 1 illustrates the SOC estimation using RF, and the specific processes of the RF application in SOC estimation are as follows.

Step 1. Data collection and preprocessing

Gather real-world driving trip profiles and record relevant data: Voltage (V), current (A), battery temperature (ºC), and ambient temperature (ºC).

Combine the data into a single dataset, ensuring proper alignment of corresponding readings.

Step 2. Data splitting

Divide the dataset into training and testing sets. For instance, trips #1 to #60 can be used for training, consisting of 945,027 instances. Trips #61 to #70 serve as the testing dataset with 118,974 instances.

Step 3. Model configuration

Configure the Random Forest model, specifying hyperparameters like the number of trees in the forest, maximum tree depth, and the number of features considered at each split. These hyperparameters can be tuned to optimize performance.

Step 4. *k*-fold cross-validation

Implement 5-fold cross-validation ($k = 5$) on the training dataset. This process involves splitting the training data into five subsets or "folds".

Train the RF model on four of these folds and validate it on the remaining one in a rotating fashion, resulting in five sets of model evaluations.

Step 5. Hyperparameter tuning - number of trees

Conduct a hyperparameter tuning experiment to determine the optimal number of trees for the Random Forest. Trial different values such as 25 trees, 50 trees, 75 trees, and 100 trees.

For each configuration, perform 5-fold cross-validation on the training data and evaluate the model's performance using metrics.

Step 6. Model training

Based on the hyperparameter tuning results, select the optimal number of trees that yields the best performance on the training dataset.

Step 7. Model evaluation

Employ the trained RF model with the chosen number of trees to predict SOC values for the testing dataset.

Calculate evaluation metrics such as RMSE, MAE, MAX, and standard deviations to assess the model's performance.

Step 8. Results and analysis

Analyze the results of the RF-based SOC estimation, emphasizing the accuracy, robustness, and consistency achieved with the chosen number of trees.

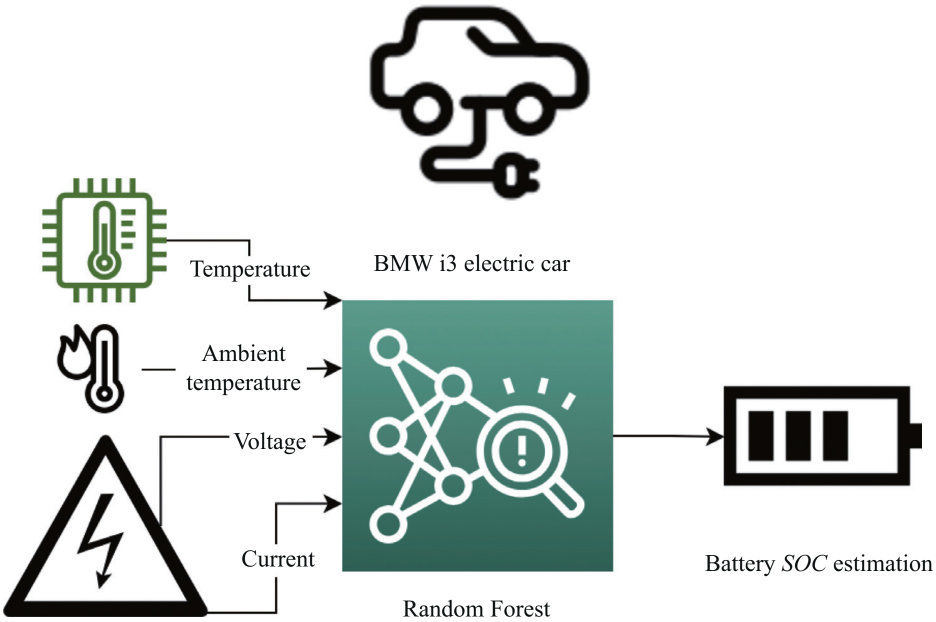


Fig. 1. SOC estimation using Random Forest (RF).

Compare the performance of the RF model with ELM.

Table 1 provides detailed information on the configuration of the data for training, validation, and testing.

To assess the effectiveness of the RF model, various metrics were utilized, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Maximum Error (MAX ERROR), and Standard Deviation (STD DEV). The definitions of these metrics are as follow:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\widehat{y}_i - y_i)^2}{n}} \tag{1}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y}_i| \tag{2}$$

where

$\widehat{y}_i -$ predicted,

$y_i -$ actual,

n-number of observations.

RMSE quantifies the dispersion of prediction errors (residuals) relative to the actual values, whereas MAE signifies the average absolute error within a set of predictions, regardless of their direction. In contrast, STD DEV and MAX ERROR are employed to assess the robustness of the proposed RF model and pinpoint the maximum error occurring at a specific instance, respectively. In this study, the selection of hyperparameter values for the RF model was determined through meticulous experimentation and optimization. The aim was to strike a balance between model complexity and prediction accuracy, taking into account the computational resources required. Specifically, a range of tree numbers, including '25 trees, 50 trees, 75 trees, and 100 trees,' was evaluated to understand their impact on the model's performance, which has been highlighted in Step 5. Through extensive experimentation, it was found that '25 trees' demonstrated the best prediction accuracy among the evaluated hyperparameter settings. This choice was made to ensure an optimal trade-off between predictive power and computational efficiency, making the RF model well-suited for real-world electric vehicle applications.

Table 1
Training and testing data division for SOC estimation using Random Forest.

Battery	Lithium-ion battery pack (60 Ah)
Profiles used	Real driving trips
Training process	Trip #1 to trip #60 that is consists of 945, 027 instances
k-fold cross-validation	$k = 5$
Testing process	Trip #61 to trip #70 that is 118, 974 instances
Input	Voltage in volts (V), current in ampere (A), battery temperature in celsius (°C) & ambient temperature in celsius (°C)
Output (%)	SOC

4. Results and discussion

All simulations for this study were conducted using MATLAB on a MacBook Pro with a 2.40 GHz Quad-Core Intel Core i5 processor and 8 GB RAM. To evaluate the performance of RF in achieving the lowest RMSE, the determination of the number of trees was experimentally executed. In this paper, different numbers of trees, specifically 25 trees, 50 trees, 75 trees, and 100 trees, were selected for the training-testing process, and the best results were recorded for comparison. Additionally, (ELM) proposed by Ref. [52], will be employed to compare performance with RF.

In order to determine number of trees in RF model, simulations for training, cross validation and testing are executed for five times. Running the RF model multiple times offers several valuable advantages. Firstly, conducting multiple runs allows for the averaging of results, which helps in reducing the variance inherent in the RF algorithm. The randomness introduced through features selection and bootstrap sampling can lead to varying predictions between runs. By averaging the results, a more stable and reliable estimation of the target variable can be achieved. Secondly, evaluating the variability of the model across different runs provides valuable insights into the consistency of its performance. This analysis helps identify potential areas of improvement and reveals any instability issues. Thirdly, in ensemble learning scenarios like bagging, conducting multiple RF runs enables the creation of an ensemble of diverse RF models. By aggregating their predictions, the overall performance is often improved. Lastly, considering random initialization's impact on results, running RF multiple times with consistent experimental setups ensures fair comparisons and meaningful evaluations. This practice helps mitigate any bias introduced by the initial random state, leading to more robust and generalizable RF models.

In this study, which has been mentioned previously, various configurations of the RF model were investigated for the SOC estimation problem, using 25 trees, 50 trees, 75 trees, and 100 trees in the ensemble. The outcomes of these simulations are visualized in Fig. 2, revealing a remarkable level of consistency in the performance results across the different tree settings. However, a more in-depth analysis of the results presented in Table 2 highlights that the RF model with 25 trees demonstrated the most favorable performance compared to the other tree configurations. As a result, the RF model utilizing 25 trees was selected as the optimal choice for further development and application in the SOC estimation problem.

In Table 2, the performance metrics of the RF model with ELM for SOC estimation are presented, encompassing different numbers of trees in the ensemble (25 trees, 50 trees, 75 trees, and 100 trees). The "Best" column showcases the optimal performance achieved for each evaluation metric among the tested configurations, while the "Average" column represents the average results obtained from five-time simulations for each metric. Among the evaluated configurations, the RF model with 25 trees emerges as the most favorable choice for SOC estimation. It achieved the lowest RMSE of 5.902,8%, indicating superior accuracy in predicting SOC compared to the other tree configurations. Additionally, the MAE yielded the lowest value of 4.432,1% for the RF model with 25 trees, signifying better precision in its predictions. Moreover, the RF model with 25 trees exhibited the smallest MAX error of 24.217,5%, suggesting less deviation in predicting extreme SOC values. Furthermore, the STD_DEV of the RF model's performance showed minimal variation across different tree configurations, highlighting its consistent performance.

The 25-tree RF architecture clearly emerges as the optimal configuration, outperforming the larger RF models and ELM method. Specifically, the 25-tree RF reduces the RMSE by 0.093,4 (5.902,8 vs 5.996,2) compared to the 50-tree model and 0.311 (5.902,8 vs 6.312,7) over the ELM. This demonstrates superior accuracy with the 25-tree configuration. Additionally, the 25-tree RF yields a 0.574,4 lower MAE than ELM (4.432,1 vs 5.111,2), highlighting greater precision. In terms of maximal deviation, the 25-tree RF exhibits the smallest MAX error at 24.217,5, which is 0.476,3 and 3.572,2 lower than the 75-tree RF and ELM respectively. Moreover, while computation time logically increases with more trees, the improvements in most metrics from 50 trees to 100 trees are marginal compared to the gains from 25 trees to 50 trees. This suggests the additional complexity and training time of larger RFs above 50 trees are unwarranted. Based on the comprehensive comparative analysis, the 25-tree RF architecture optimally balances accuracy, precision, deviation, and efficiency for SOC estimation.

In contrast, the RF model with ELM, although having fastest computation time, failed to demonstrate competitive performance. The “Best” and "Average" column for ELM consistently showed higher RMSE, MAE, MAX, and STD_DEV values compared to the RF models. The remarkable difference in computation time between ELM and RF can be attributed to several key factors. Firstly, ELM's inherent simplicity and feedforward learning approach, involving only one hidden layer in its neural network architecture, leads to faster computations due to fewer parameters to optimize during training. Additionally, random weight initialization strategy in ELM enables rapid convergence during the learning process [53]. On the other hand, RF employs an ensemble of decision trees, which requires building multiple trees, each with a varying depth, to capture complex non-linear relationships present in SOC estimation. The implementation of *k*-fold cross-validation, often used to assess the generalization performance of RF and mitigate overfitting, adds further computational overhead by repeatedly training and validating the model on multiple subsets of the data. As a result, the computation time increases, especially when using a larger number of trees and/or a higher *k* value for cross-validation. The efficiency of ELM is favored in the trade-off between computation time and predictive performance. In contrast, RF, which employs *k*-fold cross-validation, exhibits a slower computation time. This characteristic, however, allows RF to achieve heightened accuracy and robustness in tasks related to SOC estimation. Consequently, RF emerges as a valuable option for datasets necessitating rigorous evaluation and model tuning. Based on its outstanding accuracy and efficiency, the RF model with 25 trees is recommended as the optimal choice for SOC estimation in real-world applications, offering reliable predictions with notable consistency derived from the average results of five-time simulations.

Figs. 3 and 4 illustrate the assessment outcomes for testing data regarding SOC estimation achieved through RF and ELM, respectively. The results unmistakably reveal that RF surpasses ELM, showcasing superior accuracy in capturing the test data's underlying pattern. More precisely, for RF, the highest recorded error is less than 23.73%, occurring at instance #1,785, as depicted in Fig. 3. Conversely, ELM exhibits a higher error rate, with the maximum error reaching 27.86% observed at instance #104,800, as presented in Fig. 4. These results highlight the superior predictive capability of RF over ELM in the SOC estimation task,

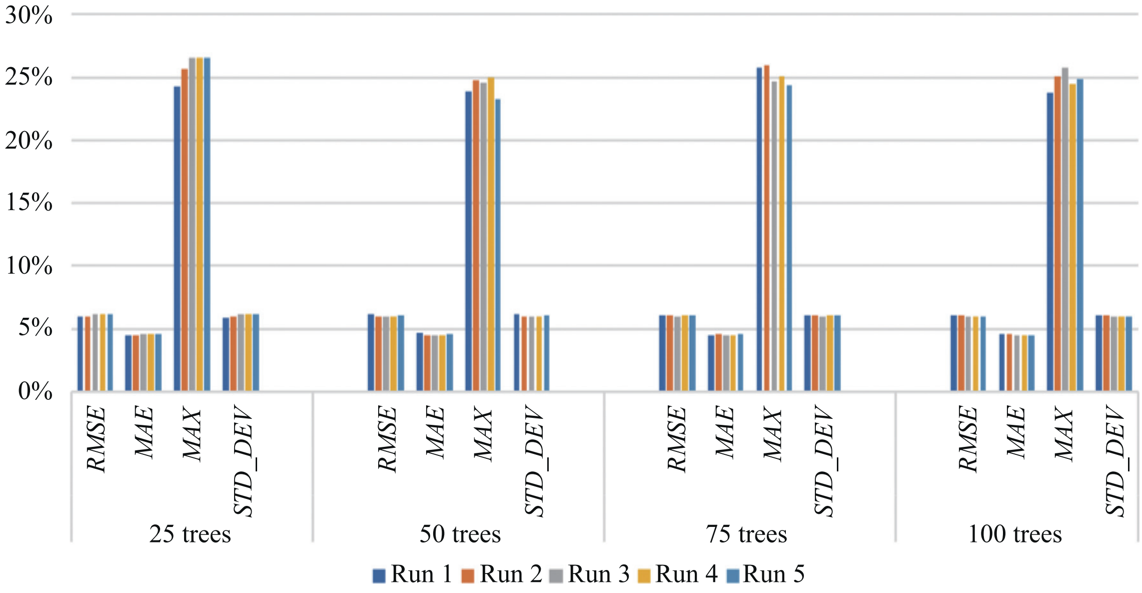


Fig. 2. Performance metrics for various configuration of trees in RF model for SOC estimation.

Table 2
Performances of RF with ELM.

Performance	Best					Average				
Evaluation (%)	RF (25 trees)	RF (50 trees)	RF (75 trees)	RF (100 trees)	ELM	RF (25 trees)	RF (50 trees)	RF (75 trees)	RF (100 trees)	ELM
RMSE	5.902,8	5.925,6	5.994,9	5.954,7	6.312,7	6.062,2	5.999,2	6.022,6	5.989,5	6.564,7
MAE	4.432,1	4.415,2	4.478,8	4.476,7	5.111,2	4.512,9	4.510,6	4.509,0	4.489,4	4.986,6
MAX	24.217,5	24.556,8	24.693,8	25.795,8	27.859,7	25.924,5	24.299,2	25.159,6	24.786,9	28.190,7
STD. DEV.	5.899,9	5.923,1	5.994,8	5.954,5	5.673,9	6.061,6	5.998,2	6.021,6	5.989,2	6.105,9
Computation time (s)	165.28	347.86	526.50	656.11	1.42					

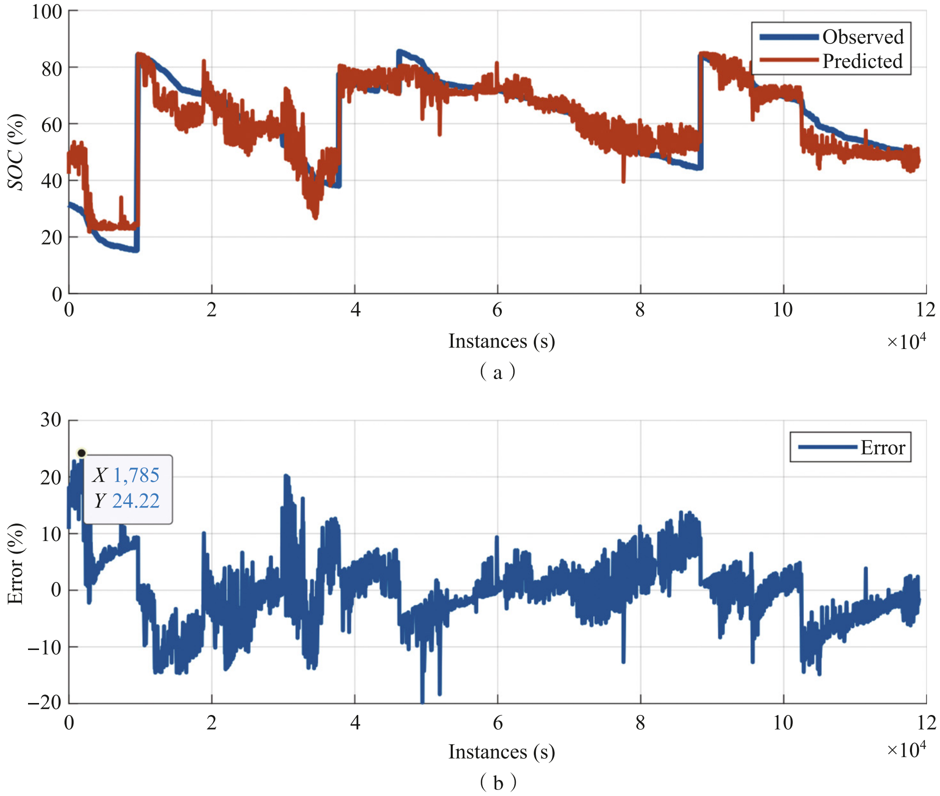


Fig. 3. Results of the (a) SOC estimation by RF for the testing data, (b) error between actual and predicted.

making RF the more effective and precise approach in this study.

It is also worth to mention that in Figs. 3 and 4, the observed large SOC estimation errors at the initial instances can be attributed to the inherent challenges associated with using real-world data captured during EV trips. Unlike synthetic or controlled data, real-world data can exhibit variations and anomalies that may not be present in idealized scenarios. These variations can stem from factors such as abrupt changes in driving conditions, sensor noise, and initial measurement inaccuracies. It is important to note that the SOC estimation model, based on RF, aims to adapt to and learn from these real-world variations. As the model processes more data instances, it progressively refines its predictions, leading to improved accuracy in estimating SOC values. This learning process is a key characteristic of machine learning models, allowing them to capture complex dependencies and enhance their performance as more data becomes available. The capacity of mitigating and diminishing these errors while incorporating additional data illustrates the adaptability and effectiveness of the approach in managing the challenges presented by real-world EV driving conditions.

A comprehensive analysis comparing the SOC curves for RF and ELM, with a focus on their performance, is presented side by side in Fig. 5. The analysis confirms that RF performs better than ELM in predicting SOC, which aligns with the observations presented in Figs. 3 and 4. The SOC curve generated by RF closely matches the actual SOC values, indicating its superior predictive ability. On the other hand, ELM shows less favorable performance, with noticeable deviations from the actual SOC values across all instances, indicating lower accuracy in SOC prediction.

Fig. 6 illustrates the error comparison between RF and ELM for SOC estimation. As been pointed out previously, it is noteworthy that the RF method exhibits a higher error magnitude at the beginning of the instances, particularly at instance #1,785. This phenomenon is consistent with previous observations in SOC estimation for EV, due to factors such as limited data availability and variations in driving behavior during the initial stages of a trip. In contrast, the ELM method displays its maximum error at a significantly later instance, specifically at #104,800. This indicates that at this particular instance, ELM failed to adapt and achieve accuracy comparable to RF. Nevertheless, it is essential to emphasize that overall, the error patterns between RF and ELM are comparable. This observation suggests that RF-based SOC estimation approach, despite exhibiting higher initial errors, converges to a level of accuracy similar to that of ELM as the analysis progresses.

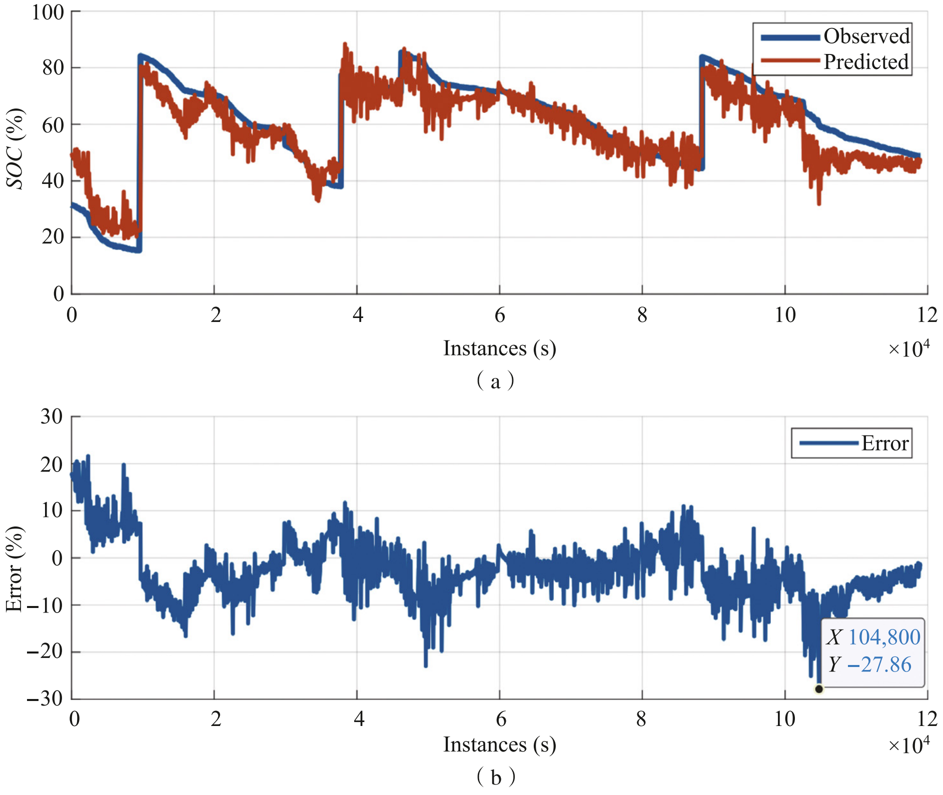


Fig. 4. Results of the (a) SOC estimation by ELM for testing data, (b) error between actual and predicted.

Simulation results indicate that both approaches successfully tracked the SOC output from real data testing. Nevertheless, there is potential for further enhancing the performance of both RF and ELM by incorporating feature selection techniques. The data collected for the BMW i3 EV includes a wide range of parameters, such as elevation, speed throttle, regenerative braking charge, traffic conditions, distance, and duration, in addition to the parameters already utilized in the current study. Through the application of feature selection methods, researchers can discern the

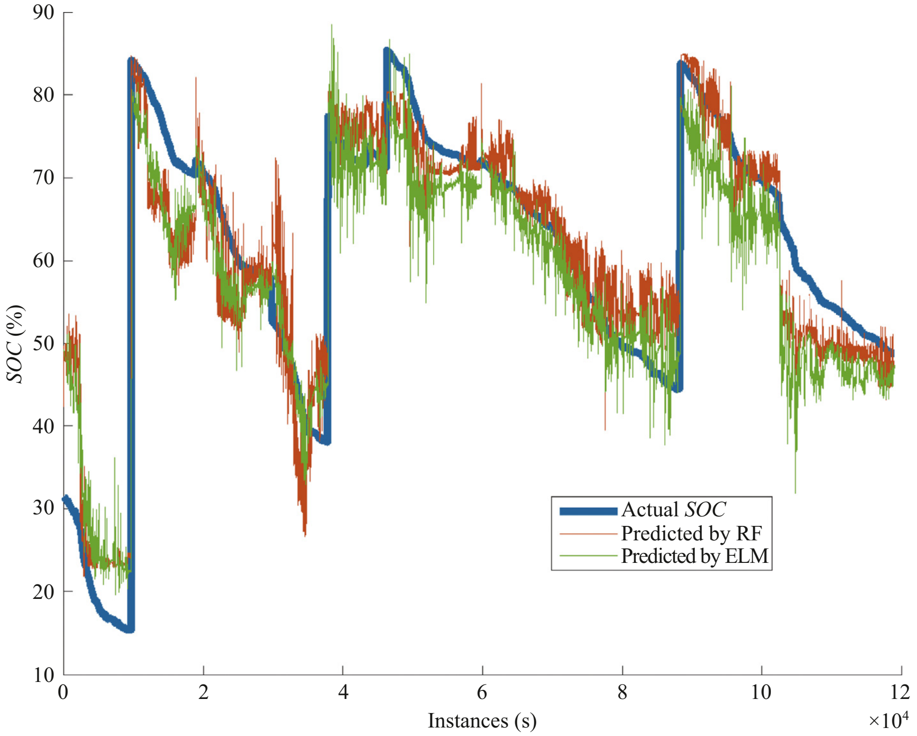


Fig. 5. Comparison performances of RF and ELM for SOC estimation.

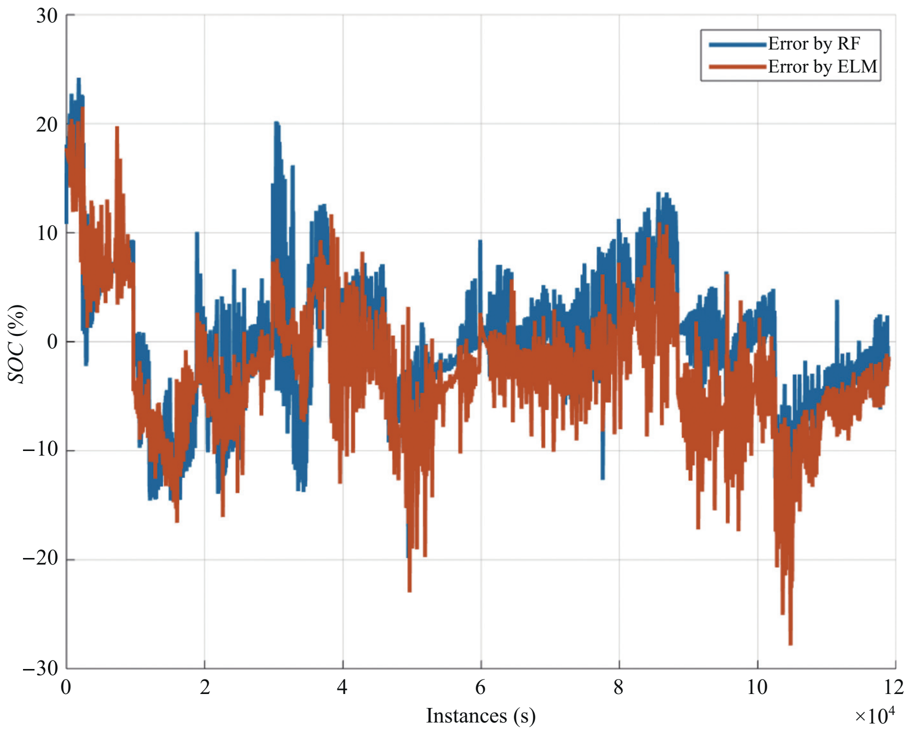


Fig. 6. Error comparison between RF and ELM for SOC estimation.

most relevant and influential parameters that impact SOC and battery health, consequently contributing to improved estimation accuracy. By focusing on the most significant features, the models can be optimized to better capture the underlying relationships and intricacies of SOC estimation in electric vehicles. This approach enables a more focused and efficient modeling process, as it selects the most informative features while reducing noise and irrelevant data. Moreover, opportunities abound for future research to explore diverse directions aimed at refining SOC estimation and gaining better insights into the battery pack's state of health in electric vehicles. By integrating feature selection into the estimation process, researchers can potentially achieve more robust and accurate predictions, benefiting the overall performance and applicability of SOC estimation techniques in EVs.

5. Conclusion

In this study, a RF model was introduced for the precise estimation of SOC in EV batteries, utilizing real-world data from a BMW i3 EV. The design of the RF model was fine-tuned, incorporating optimizations such as the selection of vital hyperparameters, including the number of trees, and configuring input–output relationships. These adjustments aimed to improve the accuracy of SOC estimation. The RF model achieved superior performance over the Extreme Learning Machine (ELM) method, with lower RMSE of 5.902,8% compared to 6.312,7% for ELM, and lower MAE of 4.432,1% versus 5.111,2% for ELM across rigorous *k*-fold cross-validation testing. This demonstrates the higher accuracy and precision of the proposed RF approach. Additionally, the MAX error was reduced from 27.859,7% with ELM down to 24.217,5% with the optimized 25-tree RF configuration, highlighting decreased deviation. The RF model's design was optimized, including the selection of crucial hyperparameters such as the number of trees, and the configuration of input–output relationships to enhance SOC estimation accuracy.

The practical significance of this SOC estimation approach extends to the electric vehicle industry as a whole. It offers the potential to revolutionize battery management, improving EV range prediction accuracy and overall battery health. The robustness and accuracy of the RF model carry significant implications for extending battery lifespan and optimizing battery usage in practical electric vehicle (EV) applications. This advancement aligns with the industry's goals of enhancing electric mobility and sustainability.

Looking ahead, future research can delve into expanding the scope of input parameters, exploring diverse input–output configurations tailored to specific driving conditions, and incorporating feature selection techniques. These endeavors promise to further enhance the accuracy and applicability of the deep learning approach in real-world EV applications. In summary, the proposed RF-based SOC estimation model stands as a compelling and accurate solution, addressing critical challenges in EV battery management. Ongoing research opportunities include the exploration of additional parameters, the customization of input–output relationships for varying conditions, and the integration of feature selection methods. These avenues of exploration reinforce the commitment to advancing SOC estimation in electric vehicles, contributing to the ongoing evolution of electric mobility.

CRediT authorship contribution statement

Mohd Herwan Sulaiman.: Data curation, Writing- Original draft preparation, Investigation, Supervision. **Zuriani Mustaffa.:** Conceptualization, Methodology, Software, Writing- Reviewing.

Data availability statement

The data and materials used to support the findings of this study are available from the corresponding author upon reasonable request. Additionally, the dataset used in this study is openly accessible on IEEE Dataport at the following link: https://ieee-dataport.org/open-access/battery-and-heating-data-real-driving-cycles.

Declaration of competing interest

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

Acknowledgements

This work was supported by the Ministry of Higher Education (MoHE) Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2022/ICT04/UMP/02/1) and Universiti Malaysia Pahang Al–Sultan Abdullah (UMPSA) under Distinguished Research Grant (# RDU223003).

References

[1] Adedeji BP. Electric vehicles survey and a multifunctional artificial neural network for predicting energy consumption in all-electric vehicles. Results Eng 2023/09/01/;19:101283. https://doi.org/10.1016/j.rineng.2023.101283. 2023.

[2] Korkmaz M. SoC estimation of lithium-ion batteries based on machine learning techniques: a filtered approach. J Energy Storage 2023/11/15/;72:108268. https://doi.org/10.1016/j.est.2023.108268. 2023.

[3] Xiong R, Cao J, Yu Q, He H, Sun F. Critical review on the battery state of charge estimation methods for electric vehicles. IEEE Access 2018;6:1832–43. https://doi.org/10.1109/ACCESS.2017.2780258.

[4] Plett GL. Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs: Part 3. State and parameter estimation. J Power Sources 2004/08/12/;134(2):277–92. https://doi.org/10.1016/j.jpowsour.2004.02.033.

[5] He H, Xiong R, Fan J. Evaluation of lithium-ion battery equivalent circuit models for state of charge estimation by an experimental approach. Energies 2011;4(4):582–98 [Online]. Available: https://www.mdpi.com/1996-1073/4/4/582.

[6] Wang J, Liu P, Hicks-Garner J, Sherman E, Soukiazian S, Verbrugge M, et al. Cycle-life model for graphite-LiFePO4 cells. J Power Sources 2011/04/15/2011;196(8):3942–8. https://doi.org/10.1016/j.jpowsour.2010.11.134.

[7] Wang S, Fan Y, Jin S, Takyi-Aninakwa P, Fernandez C. Improved anti-noise adaptive long short-term memory neural network modeling for the robust remaining useful life prediction of lithium-ion batteries. Reliab Eng Syst Saf 2023;230:108920. https://doi.org/10.1016/j.ress.2022.108920.

[8] Wang S, Wu F, Takyi-Aninakwa P, Fernandez C, Stroe D-I, Huang Q. Improved singular filtering-Gaussian process regression-long short-term memory model for whole-life-cycle remaining capacity estimation of lithium-ion batteries adaptive to fast aging and multi-current variations. Energy 2023;284:128677. https://doi.org/10.1016/j.energy.2023.128677.

[9] Zhang C, Luo L, Yang Z, Zhao S, He Y, Wang X, et al. Battery SOH estimation method based on gradual decreasing current, double correlation analysis and GRU. Green Energy Intel Trans 2023/10/01/;2(5):100108. https://doi.org/10.1016/j.geits.2023.100108. 2023.

[10] Yi Z, Chen Z, Yin K, Wang L, Wang K. Sensing as the key to the safety and sustainability of new energy storage devices. Protect Control Modern Pow Sys 2023/06/12;8(1):27. https://doi.org/10.1186/s41601-023-00300-2.

[11] Liu Y, Wang L, Li D, Wang K. State-of-health estimation of lithium-ion batteries based on electrochemical impedance spectroscopy: a review. Protect Control Modern Pow Sys 2023/08/31;8(1):41. https://doi.org/10.1186/s41601-023-00314-w.

[12] Kumari P, Singh AK, Kumar N. Electric vehicle battery state-of-charge estimation based on optimized deep learning strategy with varying temperature at different C Rate. J Eng Res 2023/06/10:100113. https://doi.org/10.1016/j.jer.2023.100113.

[13] Ma L, Zhang T. Deep learning-based battery state of charge estimation: enhancing estimation performance with unlabelled training samples. J Energy Chem 2023/05/01/2023;80:48–57. https://doi.org/10.1016/j.jechem.2023.01.036.

[14] Tian J, Chen C, Shen W, Sun F, Xiong R. Deep learning framework for lithium-ion battery state of charge estimation: recent advances and future perspectives. Energy Storage Mater 2023/08/01;61:102883. https://doi.org/10.1016/j.ensm.2023.102883.

[15] Yu H, Zhang L, Wang W, Li S, Chen S, Yang S, et al. State of charge estimation method by using a simplified electrochemical model in deep learning framework for lithium-ion batteries. Energy 2023/09/01;278:127846. https://doi.org/10.1016/j.energy.2023.127846.

[16] Zafar MH, Mansoor M, Abou Houran M, Khan NM, Khan K, Raza Moosavi SK, et al. Hybrid deep learning model for efficient state of charge estimation of Li-ion batteries in electric vehicles. Energy 2023/11/01;282:128317. https://doi.org/10.1016/j.energy.2023.128317.

[17] Sulaiman MH, Mustaffa Z, Zakaria NF, Saari MM. Using the evolutionary mating algorithm for optimizing deep learning parameters for battery state of charge estimation of electric vehicle. Energy 2023/09/15;279:128094. https://doi.org/10.1016/j.energy.2023.128094.

[18] Fu B, Wang W, Li Y, Peng Q. An improved neural network model for battery smarter state-of-charge estimation of energy-transportation system. Green Energy Intel Trans 2023/04/01;2(2):100067. https://doi.org/10.1016/j.geits.2023.100067.

[19] Mawonou KSR, Eddahech A, Dumur D, Beauvois D, Godoy E. State-of-health estimators coupled to a random forest approach for lithium-ion battery aging factor ranking. J Power Sources 2021/02/01;484:229154. https://doi.org/10.1016/j.jpowsour.2020.229154.

[20] Shibl MM, Ismail LS, Massoud AM. A machine learning-based battery management system for state-of-charge prediction and state-of-health estimation for unmanned aerial vehicles. J Energy Storage 2023/08/30;66:107380. https://doi.org/10.1016/j.est.2023.107380.

[21] Chen T, Wang Y-C. A modified random forest incremental interpretation method for explaining artificial and deep neural networks in cycle time prediction. Decision Analy J 2023/06/01;7:100226. https://doi.org/10.1016/j.dajour.2023.100226.

[22] Bhat D, Muench S, Roellig M. Application of machine learning algorithms in prognostics and health monitoring of electronic systems: a review. e-Prime Adv Elec Eng Electro Energy 2023/06/01;4:100166. https://doi.org/10.1016/j.prime.2023.100166.

[23] Budiman TA, James CR, Howlett NC, Wood RM. Near real-time prediction of urgent care hospital performance metrics using scalable random forest algorithm: a multi-site development. Healthcare Analytics 2023/11/01;3:100169. https://doi.org/10.1016/j.health.2023.100169.

[24] Kishino M, Matsumoto K, Kobayashi Y, Taguchi R, Akamatsu N, Shishido A. Fatigue life prediction of bending polymer films using random forest. Int J Fatig 2023/01/01;166:107230. https://doi.org/10.1016/j.ijfatigue.2022.107230.

[25] Ali MR, Nipu SMA, Khan SA. A decision support system for classifying supplier selection criteria using machine learning and random forest approach. Decision Analyt J 2023/06/01;7:100238. https://doi.org/10.1016/j.dajour.2023.100238.

[26] Khanna VV, Chadaga K, Sampathila N, Prabhu S, C. P R. A machine learning and explainable artificial intelligence triage-prediction system for COVID-19. Decision Analyt J 2023/06/01;7:100246. https://doi.org/10.1016/j.dajour.2023.100246.

[27] 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, doi: 10.3390/machines10100912.

[28] Jin Y, Su C, Luo S. Improved algorithm based on AEKF for state of charge estimation of lithium-ion battery. Int J Automot Technol 2022/08/01;23(4):1003–11. https://doi.org/10.1007/s12239-022-0087-x.

[29] Liu Q, Yu Q. The lithium battery SOC estimation on square root unscented Kalman filter. Energy Rep 2022/10/01;8:286–94. https://doi.org/10.1016/j.egyr.2022.05.079.

[30] Yang F, Shi D, Lam K-h. Modified extended Kalman filtering algorithm for precise voltage and state-of-charge estimations of rechargeable batteries. J Energy Storage 2022/12/01/2022;56:105831. https://doi.org/10.1016/j.est.2022.105831.

[31] Wu M, Qin L, Wu G. State of charge estimation of power lithium-ion battery based on an affine iterative adaptive extended kalman filter. J Energy Storage 2022/07/01;51:104472. https://doi.org/10.1016/j.est.2022.104472.

[32] Marques TMB, dos Santos JL, Castanho DS, Ferreira MB, Stevan SL, Illa Font CH, et al. An overview of methods and technologies for estimating battery state of charge in electric vehicles. Energies 2023;16(13):5050 [Online]. Available: https://www.mdpi.com/1996-1073/16/13/5050.

[33] Zeng Y, Li Y, Yang T. State of charge estimation for lithium-ion battery based on unscented kalman filter and long short-term memory neural network. Batteries 2023;9(7):358 [Online]. Available: https://www.mdpi.com/2313-0105/9/7/358.

[34] Hussein AA. Kalman filters versus neural networks in battery state-of-charge estimation: a comparative study. Int J Mod Nonlinear Theor Appl 2014;3(5):199–209. https://doi.org/10.4236/ijmnta.2014.35022.

[35] Zheng X, Fang H. An integrated unscented kalman filter and relevance vector regression approach for lithium-ion battery remaining useful life and short-term capacity prediction. Reliab Eng Syst Saf 2015/12/01;144:74–82. https://doi.org/10.1016/j.ress.2015.07.013.

[36] Breiman L. Random forests. Mach Learn 2001/10/01;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

[37] Li Y, Zou C, Berecibar M, Nanini-Maury E, Chan JCW, van den Bossche P, et al. Random forest regression for online capacity estimation of lithium-ion batteries. Appl Energy 2018/12/15/;232:197–210. https://doi.org/10.1016/j.apenergy.2018.09.182. 2018.

[38] da Silva DG, Geller MTB, Moura MSdS, Meneses AAdM. Performance evaluation of LSTM neural networks for consumption prediction. e-Prime Adv Elec Eng Electro Energy 2022/01/01;2:100030. https://doi.org/10.1016/j.prime.2022.100030.

[39] Cai Q, Li P, Wang R. Electricity theft detection based on hybrid random forest and weighted support vector data description. Int J Electr Power Energy Syst 2023/11/01;153:109283. https://doi.org/10.1016/j.ijepes.2023.109283.

[40] Yang B, Haghighat F, Fung BCM, Panchabikesan K. Season-based occupancy prediction in residential buildings using machine learning models. e-Prime Adv Elec Eng Electro Energy 2021/01/01;1:100003. https://doi.org/10.1016/j.prime.2021.100003.

[41] Cao M, Yin D, Zhong Y, Lv Y, Lu L. Detection of geochemical anomalies related to mineralization using the random forest model optimized by the competitive mechanism and beetle antennae search. J Geochem Explor 2023/06/01;249:107195. https://doi.org/10.1016/j.gexplo.2023.107195.

[42] Dinh TP, Pham-Quoc C, Thinh TN, Nguyen BKD, Kha PC. A flexible and efficient FPGA-based random forest architecture for IoT applications. Internet Things 2023/07/01;22:100813. https://doi.org/10.1016/j.iot.2023.100813.

[43] Guo W, Gao Z, Guo H, Cao W. Hydrogeochemical and sediment parameters improve predication accuracy of arsenic-prone groundwater in random forest machine-learning models. Sci Total Environ 2023/11/01;897:165511. https://doi.org/10.1016/j.scitotenv.2023.165511.

[44] Koohmishi M, Azarhoosh A, Naderpour H. Assessing the key factors affecting the substructure of ballast-less railway track under moving load using a double-beam model and random forest method. Structures 2023/09/01;55:1388–405. https://doi.org/10.1016/j.istruc.2023.06.027.

[45] Ma X, Chen Z, Chen P, Zheng H, Gao X, Xiang J, et al. Predicting the utilization factor of blasthole in rock roadways by random forest. Undergr Space 2023/08/01/2023;11:232–45. https://doi.org/10.1016/j.undsp.2023.01.006.

[46] Zhang H, Peng J, Wang R, Zhang M, Gao C, Yu Y. Use of random forest based on the effects of urban governance elements to forecast CO2 emissions in Chinese cities. Heliyon 2023/06/01;9(6):e16693. https://doi.org/10.1016/j.heliyon.2023.e16693.

[47] Mishra R, Kiran MSNS, Maheswaram M, Upadhyay A, Singh B. Investigation of optimal feature for milling chatter identification using supervised machine learning techniques. J Eng Res 2023/06/28:100138. https://doi.org/10.1016/j.jer.2023.100138.

[48] İfraz M, Aktepe A, Ersöz S, Çetinyokuş T. Demand forecasting of spare parts with regression and machine learning methods: application in a bus fleet. J Eng Res 2023/06/01;11(2):100057. https://doi.org/10.1016/j.jer.2023.100057.

[49] Lucchetta B. Battery state of charge estimation using a machine learning approach. CORSO DI LAUREA MAGISTRALE IN INFORMATICA, Dipartimento di Matematica "Tullio Levi-Civita". Italy: Universit`a degli Studi di Padova; 2021.

[50] M. S. J. B. D. Trifonov. Battery and heating data in real driving cycles, doi: 10.21227/6jr9-5235.

[51] Probst P, Wright MN, Boulesteix A-L. Hyperparameters and tuning strategies for random forest. WIREs Data Mining Knowledge Discovery 2019/05/01;9(3):e1301. https://doi.org/10.1002/widm.1301.

[52] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. Neurocomputing 2006/12/01;70(1):489–501. https://doi.org/10.1016/j.neucom.2005.12.126.

[53] Bai Z, Li F, Zhang J, Oko E, Wang M, Xiong Z, et al. Modelling of a post-combustion CO2 capture process using bootstrap aggregated extreme learning machines. In: Kravanja Z, Bogataj M, editors. Computer aided chemical engineering, 38. Elsevier; 2016. p. 2007–12.