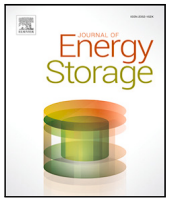
Journal of Energy Storage 68 (2023) 107577

Contents lists available at ScienceDirect

Journal of Energy Storage

journal homepage: www.elsevier.com/locate/est



Research papers

A novel smart feature selection strategy of lithium-ion battery degradation modelling for electric vehicles based on modern machine learning algorithms

Huzaifa Rauf [a,b,c,g,*], Muhammad Khalid [e,f,g], Naveed Arshad [c,d]

[a] Department of Electrical Engineering, Lahore University of Management Sciences (LUMS), Sector U, Phase 5 D.H.A, Lahore, 54000, Pakistan
[b] Center for Advanced Life Cycle Engineering, University of Maryland, College Park, MD, 54000, United States
[c] Department of Computer Science, Lahore University of Management Sciences (LUMS), Sector U, Phase 5 D.H.A, Lahore, 54000, Pakistan
[d] LUMS Energy Institute, Sector U, Phase 5 D.H.A, Lahore, 54000, Pakistan
[e] Electrical Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, 31261, Saudi Arabia
[f] Interdisciplinary Research Center for Renewable Energy and Power Systems (IRC-REPS), KFUPM, Dhahran, 31261, Saudi Arabia
[g] SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Dhahran, 31261, Saudi Arabia

ARTICLE INFO

Keywords:
Battery degradation
Li-ion batteries
Electric vehicles
Machine learning
Capacity loss
Prediction

ABSTRACT

Lithium-ion batteries are a key storage technology for electric vehicles and renewable energy applications. However, the complex degrading behaviour of batteries impacts their capacity and lifetime. Thus, battery capacity loss prediction is crucial for ensuring the longevity, safety, and reliable operation of the battery. This research proposes a smart feature selection (SFS) strategy-based machine learning framework for battery calendar and cyclic loss prediction. The presented methodology selects input parameters from the battery data of the current time step as well as the previous time step which are then utilized for model training and testing. Results demonstrate that the proposed SFS method in combination with the ML algorithms enhances the prediction accuracy and reduces the mean absolute error for all the machine learning algorithms applied in this study. The proposed SFS method is capable of excavating the useful features, therefore offering good generalization ability and accurate prediction results for capacity loss of the lithium-ion battery under real EV usage conditions. Furthermore, the results also depict that the performance accuracy of ML methods for battery calendar and cyclic loss prediction improves when combined with the SFS method. Greater improvement in prediction accuracy of battery capacity loss is observed for Gaussian Process Regression (GPR), random forest (RF), and XGBoost methods when applied in combination with the proposed SFS. This is the first-known feature selection-based ML application that is utilized to independently perform battery calendar and cyclic loss prognosis.

1. Introduction

The transition of personal transportation from internal combustion engine (ICE) vehicles to electric vehicles (EVs) is a vital step in achieving lower carbon emissions from the transportation sector [1]. EVs and renewable energy systems are widely promoted as clean alternatives to conventional vehicles and power generation and as promising solutions to effectively reduce GHG emissions and other environmental problems [2]. The rapid development of the EV and renewable energy industry as a clean alternative to fossil-fuel-based vehicles and power generation sources has increased the demand for energy storage technologies [3]. Among the available energy storage technologies, lithium-ion (Li-ion) batteries have detached as one of the solutions, which can meet the requirements imposed by both power grids and transportation sectors [4,5]. In recent years, a significant interest in battery-related applications has arisen globally due to reducing fuel consumption, mitigating dependence on imported oil, and decreasing greenhouse gas emissions [6].

Over the last few decades, battery technology has made significant progress in the area of energy storage and plays a key role in EVs and renewable energy systems [7]. The advancements in Li-ion batteries (LIBs) have attracted considerable attention due to their high energy density, low maintenance, and optimal performance [8]. Meanwhile, the reliability and safety assessment of LIBs has become an important issue, in particular for future EV performance [9]. The energy provision and consumption in LIB-related applications are highly dependent on the health condition of batteries and one main limitation of LIBs resides in battery ageing [10].

* Corresponding author at: Department of Electrical Engineering, Lahore University of Management Sciences (LUMS), Sector U, Phase 5 D.H.A, Lahore, 54000, Pakistan.
E-mail addresses: huzaifa.rauf@lums.edu.pk (H. Rauf), mkhalid@kfupm.edu.sa (M. Khalid), naveedarshad@lums.edu.pk (N. Arshad).

https://doi.org/10.1016/j.est.2023.107577
Received 4 June 2022; Received in revised form 12 April 2023; Accepted 28 April 2023
Available online 20 May 2023

LIBs are increasing in popularity, and there is an increased need to study and model their capacity degradation. The classical problem associated with the EV battery is that it undergoes a sophisticated degradation process during EV operations [11]. Battery degradation gradually happens over time under specific driving conditions and affects EV power consumption due to battery ageing. LIBs undergo operation periods that are substantially shorter than the idle intervals and have different stresses of C-rate, depth-of-discharge (DOD), temperature, and state-of-charge (SoC) [12,13]. LIBs undergo a process to store and provide electrical energy which can last over different time scales. This stationary and transient operation of the LIB causes calendar and cyclic loss, respectively [14]. Battery degradation takes place in every condition, but in different proportions as usage and external conditions interact to provoke degradation. When a defined degradation level is reached, the battery reaches its end-of-life (EOL) and has to be replaced. To address these difficulties, precise battery degradation models capable of accurately predicting the performance and lifetime of LIBs need to develop [15]. Battery lifetime models are used to predict the long-term degradation behaviour of LIB performance metrics such as capacity and internal resistance [16].

Generally, the phenomena of battery degradation can be classified into two categories: the calendar loss, which refers to the irreversible loss of battery capacity during storage, and the cyclic loss, which occurs due to battery charge and discharge cycles [17]. Cyclic ageing is one of the two main aspects used to model the battery degradation of a LIB. Battery cyclic loss is mainly dictated by the number of battery charging and discharging cycles and is defined as the loss in capacity of the battery when it undergoes a charging or discharging process. This is a direct consequence of the utilization mode, the temperature conditions, and the current use of the battery. Consequently, many factors are involved with cyclic ageing. In particular, the prediction of cyclic loss requires a large variety of activities concentrated on the analysis of cyclic loss behaviours of LIBs. In addition, calendar ageing is the other critical aspect used for battery degradation modelling of a LIB. However, unlike cyclic ageing, it comprises all ageing processes that lead to battery degradation independent of the charge–discharge cycle. Calendar prediction requires a substantial heterogeneous strategy concentrated on the analysis of the calendar loss behaviour of LIBs.

A comprehensive understanding of the battery ageing mechanisms and the ability to accurately predict the cyclic and calendar loss is crucial for battery degradation modelling. An accurate capacity loss prediction and battery degradation model allow for early detection of a battery's inadequate performance which facilitates timely maintenance of battery systems. To accurately model battery degradation and predict capacity loss, there is a need for effective techniques and methods to predict cyclic and calendar loss. There are many factors that affect the battery cyclic and calendar loss, which makes their prediction convoluted. Therefore, it is extremely significant to select a suitable prediction method and devise an accurate model. Among the data-driven techniques, ML is becoming more popular for predicting battery degradation trends due to the greater availability of battery data and improved computing [18,19]. ML methods have recently gained an appreciable research focus due to their finer data integrity, and have shown considerable promise in battery lifetime studies. ML methods can independently realize the relationship between battery capacity loss and external parameters, and establish a prediction model of the battery capacity loss. Various ML models are employed depending on the data quality, inputs and outputs, test conditions, battery types, and stated accuracy for battery calendar and cyclic loss prediction [20].

There are different processes linked to ML algorithms, which include data pre-processing, feature selection, model training, and testing [21]. The improvement in the outcomes of these processes considerably enhances the prediction capability. In particular, the accuracy of the capacity loss prediction is greatly affected by the feature selection of the battery data [22]. The model accuracy depends on the correlation between the feature data and the output label greater the correlation, the higher the accuracy of the cyclic and calendar loss prediction model. Nevertheless, different ML-based methods which typically include the aspects of data collection, data pre-processing, feature selection, and training/testing have been thoroughly studied for LIBs with their main objective to predict the battery capacity, health indicators, and lifetime [23,24]. For example, Zhang et al. [25] used a neural network (NN) to forecast battery lifetime using discharge capacity, terminal voltage, discharge current, and internal resistance. Yang et al. [26] utilized a gradient boosting regression tree (GBRT) model to predict battery life by considering voltage, capacity, and temperature characteristics. Li et al. [22] predicted the battery health by selecting the features from incremental capacity curves and applying a Gaussian process regression (GPR) model. Shu et al. [27] extracted characteristics from voltage curves and predicted battery health using a support vector machine (SVM) model. Xu et al. [28] proposed an online extreme ML method which is used to predict the capacity of LIBs. Zhao et al. [29] used two features of equal charging and discharging voltage difference time interval, and established their relationship model with capacity using support vector regression to evaluate online capacity. Ma et al. [30] applied NN that integrated a convolutional neural network (CNN) and long short-term memory (LSTM) to predict capacity loss. Guo et al. [31] selected 14 features, including charging time, temperature and voltage curve slope, as the feature vectors of battery degradation in the charging process of LIB, and predicted remaining capacity by the relevance vector machine (RVM). Li et al. [32] predicted the battery capacity using NN and the battery charging time, discharging time, and discharge capacity as characteristic features. Yang et al. [33] identified four features from constant-current charging curves and predicted the battery SOH using an enhanced GPR model. Wu et al. [34] applied a feed-forward neural network (FFNN) to predict the current battery cycle number after sampling the battery terminal voltages during the charging process. In these studies, the capacity loss is modelled using different feature selection techniques to predict battery health, and capacity to failure threshold, which is then used to predict remaining useful life (RUL). The literature discusses different methods for feature selection which are used in combination with ML algorithms. The existing approaches have shown satisfactory performance in predicting battery lifetime. However, due to the limitations and variations in battery datasets, the feature selection methods must be robust as extracting meaningful information from the raw data is extremely necessary.

Our proposed study aims to contribute in terms of devising and evaluating a smart feature selection (SFS) method which is utilized in the ML algorithms to predict battery cyclic and calendar loss. The study analyses the relationship between capacity loss and input features using cyclic and calendar loss prediction and introduces the SFS method that has enhanced the generalization ability and improved the predictive performance of the ML algorithms to accurately predict total capacity loss. A case study has been undertaken for the validation of the framework, in which features are extracted based on battery degradation data using the SFS method which selects features to reflect the battery ageing dynamics from different perspectives. Multiple indicators for battery capacity loss prediction are selected and various ML techniques have been extensively applied to predict cyclic and calendar loss. This leads to effectively predicting the battery capacity loss and demonstrates the effectiveness of the proposed framework. To manage feature irrelevancy and redundancy, SFS generates an optimal feature subset. The selected feature subset is then fed to eight representative ML algorithms involving linear regression, ridge regression, LASSO regression, elastic net, GPR, SVM, random forest (RF), and XGBoost. Comparative tests are carried out to demonstrate the efficiency of the proposed framework. The results suggest that the proposed strategy improves the predictive ability of ML models.

In summary, the contributions of this research work are as follows:

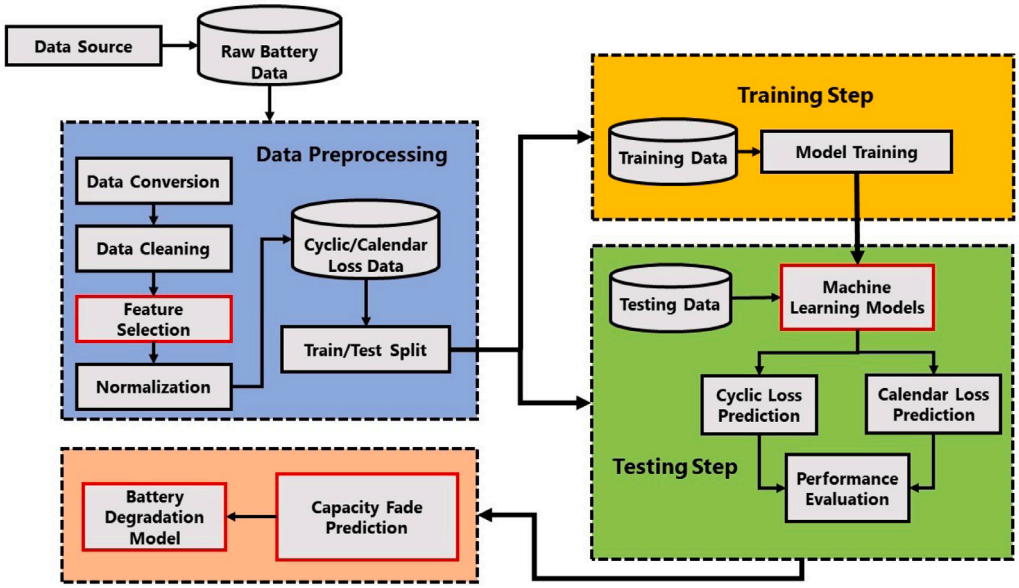


Fig. 1. Architecture of the ML-based framework for battery capacity loss prediction.

- A smart feature selection framework is developed which is utilized in combination with the ML algorithms to predict battery cyclic and calendar loss. The framework analyses the relationship between capacity loss and input features using the cyclic and calendar loss prediction and introduces a smart feature selection method which enhanced the generalization ability of the model and improved the predictive capability of ML algorithms to accurately predict battery capacity loss. The framework has been implemented with various types of ML models to verify the proposed framework by testing on battery data.
- The developed SFS-based ML framework for battery capacity loss prediction is evaluated and validated through a case study of a ten-year U.S-based EV battery dataset which includes features to reflect the battery ageing dynamics from different perspectives. Multiple indicators for battery capacity loss prediction are extracted and various ML techniques have been extensively applied to predict cyclic and calendar loss, which eventually leads to effectively predicting the battery capacity loss and demonstrates the effectiveness of the developed framework.

The paper is organized as follows: In Section 2, the methodology of the proposed SFS method and eight ML models is introduced and a comprehensive framework and application of SFS in combination with ML models for battery capacity loss prediction are presented. In Section 3, the description and specification of the experimental dataset under consideration and its subsequent pre-processing are discussed. In Section 4, the experimental results of battery cyclic and calendar prediction are demonstrated. In Section 5, the conclusion of the research study is presented.

2. Proposed methodology

2.1. Problem description

ML techniques are often used to train the complex non-linear degrading behaviour of LIBs based on historical data, and they do not necessitate a thorough grasp of the battery's internal activity [35]. The commonly used ML methods include neural networks (NNs) [36], Support vector machines (SVMs) [37], Relevance vector machine (RVM) [38], and Gaussian process regression (GPR) [22]. ML methods rely on input feature selection in the battery data to predict the battery cyclic and calendar loss. Typically, the problem lies in analysing the relationship between the battery input features, and the battery cyclic and calendar loss, which is critical to establishing an accurate capacity loss prediction model. In other words, the performance of cyclic and calendar loss prediction leads to accurate capacity loss prediction, and it mainly depends on the choice of input feature extraction. Due to the limited types of battery data, it is particularly important to extract useful information from the battery data, which is related to battery capacity loss. From a practical point of view, and while considering the difficulty involved in extracting useful features, the efficacy of battery capacity loss prediction is enhanced if the method to extract the features is made efficient and robust. The overall framework of the battery cyclic and calendar loss prediction and the subsequent battery capacity loss prediction using the smart feature selection method is illustrated in Fig. 1. The first step comprises the data pre-processing, in which the raw battery data undergoes the process of data conversion and data cleaning. The proposed smart feature selection method is then integrated to extract the mapping relationship between the selected features and the practical cyclic and calendar loss label and to predict the cyclic and calendar loss trend of the battery. The next step involves the processed data is split into training and testing parts for training and testing of the model, respectively. Various ML methods are applied to estimate prediction accuracy.

The main idea of the proposed research is to present a smart and improved feature selection strategy for the extraction of the actual characteristic features for battery cyclic and calendar loss prediction. The accurate prediction of battery cyclic and calendar loss subsequently leads to predicting the battery capacity and battery degradation modelling.

2.2. Smart feature selection (SFS) strategy

The smart feature selection (SFS) approach is proposed to accurately predict the cyclic and calendar loss of LIBs which is evaluated in combination with the ML algorithms. The conventional and most commonly used feature selection approach selects the input features of the current time set and applies ML methods considering these features. On the contrary, the proposed SFS feature method comprises the selection of all input features of the previous and the current time step. In addition, the SFS method also takes into consideration the previous time step output label as an input feature for model training, as illustrated in Fig. 2, and Fig. 3. The ML algorithms are subsequently applied considering those selected features.

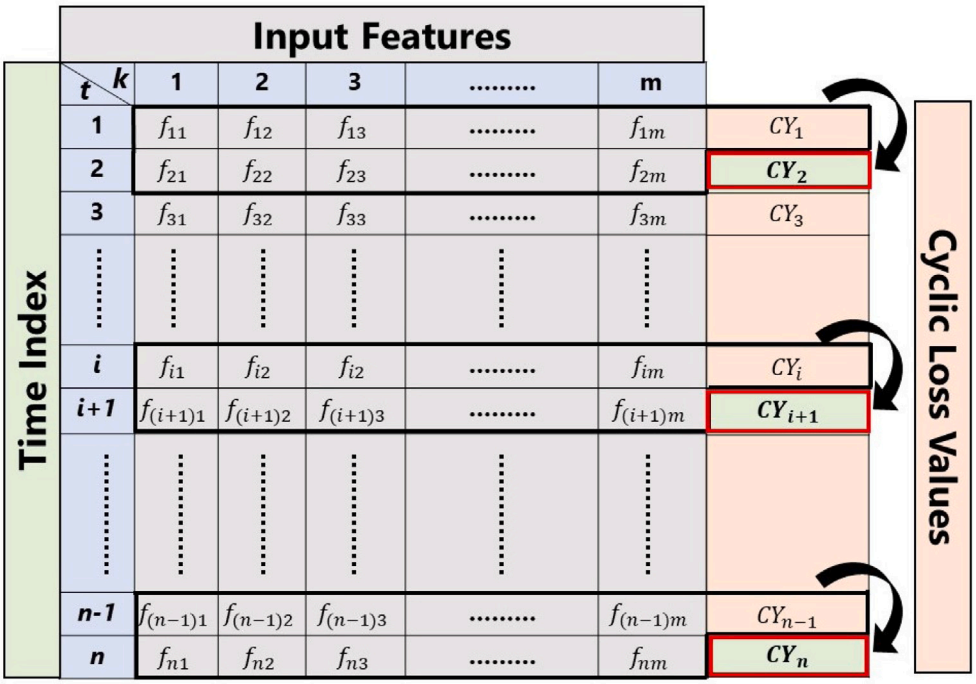


Fig. 2. SFS framework for battery cyclic loss feature selection.

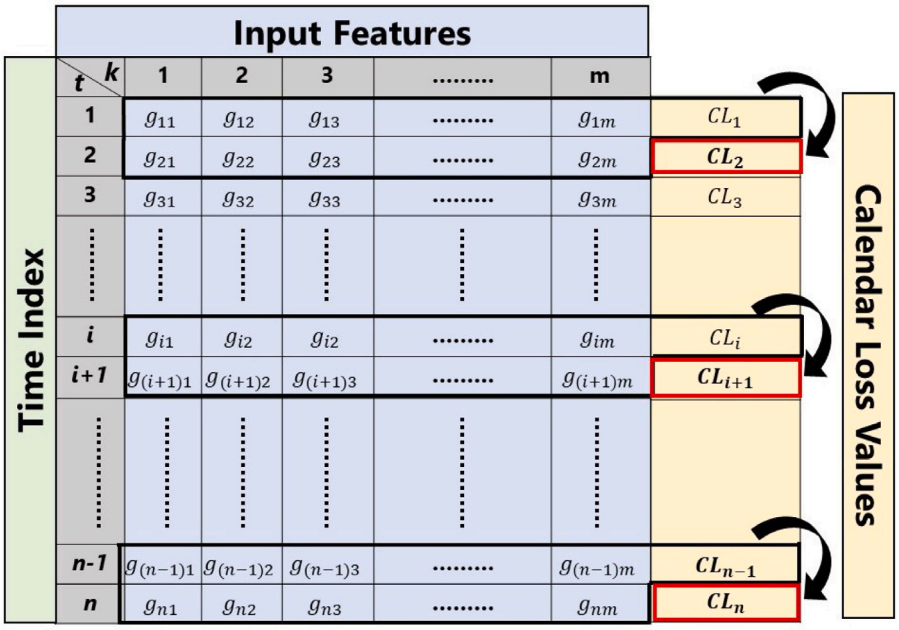


Fig. 3. SFS framework for battery calendar loss feature selection.

In this study, SFS is applied as a feature selection technique for different ML methods to predict the battery cyclic and calendar loss. The SFS method selects the cyclic and calendar loss features of previous and current time steps along with the previous time step's estimated cyclic loss and calendar loss output labels to devise the ML-based prediction framework. The current time step′s cyclic loss and calendar loss outputs are taken as an input feature to the ML model to predict the future cyclic and calendar loss values. This can limit the effect of the accumulated error, as the model is trained on both the features and recursive inputs. Predictions are made over a period of time by recursively feeding the model outputs from earlier time steps in as inputs for later time steps. The SFS method evaluates the current and previous time step features information to reflect the cyclic and calendar loss prediction better and resolve the problem of error accumulation by reducing the prediction error using the output labels of the previous time step. The SFS method is a direct approach which aims to avoid error accumulation by creating a separate model for each potential time horizon. The SFS method is integrated to extract the mapping relationship between the selected features and the cyclic/calendar loss, which leads to the prediction of the battery capacity loss.

2.2.1. SFS for battery cyclic loss

The feature selection-based framework of the SFS method for battery cyclic loss is given in Fig. 2. It is mathematically represented as:

$$C_{cyc(t)} = \{(CY_t, t) | CY_t \in \mathbb{R}, t \in \mathbb{N} \wedge t \leq n\} \quad (1)$$

where CY_t is the cyclic loss output that represents the battery cyclic loss value at the time index value t. Here t is represented as $t = 1, 2, \ldots, n$ and n represents the total number of time steps considered for the evaluation. For every cyclic loss value, CY_t, there are associated input features which are represented as f_{tk}. In the set notation form, the feature set is represented as:

$$F_{cyc(tk)} = \{(f_{tk}, t, k) | f_{tk} \in \mathbb{R}^{n \times m}, t \in \mathbb{N}, k \in \mathbb{N} \wedge t \leq n, k \leq m\} \quad (2)$$

Similarly, for future cyclic loss value CY_{t+1}, there are associated input features which are represented as $f_{(t+1)k}$. where k shows the feature index which is represented as $k = 1, 2, \ldots, m$ and m denotes the total number of input features for each cyclic loss output value. In the set notation form, the input feature set corresponding to cyclic loss is represented as:

$$F_{cyc(t+1)k} = \{f_{(t+1)k} | f_{(t+1)k} \in \mathbb{R}^{n \times m}\} \quad (3)$$

The feature sets of cyclic loss values are characterized as $f_{1k} \rightarrow CY_1, f_{2k} \rightarrow CY_2, \ldots, f_{nk} \rightarrow CY_n$. According to the SFS method, the input feature set which is used for the prediction of the future time step's cyclic loss CY_{t+1}, is denoted by $F_{cyc[c(t+1)]}$ and is given as a function of $f_{tk}, f_{(t+1)k}$ and $C_{cyc(t)}$:

$$F_{cyc[CY(t+1)]} = f(f_{tk}, f_{(t+1)k}, C_{cyc(t)}) \quad (4)$$

where f_{tk}, and $f_{(t+1)k}$ are the $n \times m$ input feature vector sets which correspond to the present time step's cyclic loss CY_t, and future time step's cyclic loss CY_{t+1}, respectively. The SFS-based input feature set $F_{cyc[CY(t+1)]}$ is used to predict the future time step's cyclic loss CY_{t+1}.

2.2.2. SFS for battery calendar loss

The framework of the proposed feature selection method for battery calendar loss is given in Fig. 3, in which CL_t is considered as the calendar loss output that represents the battery calendar loss value at the time index value t. Here t is represented as $t = 1, 2, \ldots, n$ and n represents the total number of time steps considered for the evaluation. It is mathematically represented as:

$$C_{cal(t)} = \{(CL_t, t) | CL_t \in \mathbb{R}, t \in \mathbb{N} \wedge t \leq n\} \quad (5)$$

For every calendar loss value, CL_t, there are associated input features which are represented as g_{tk}. In the set notation form, the feature set is represented as:

$$G_{tk} = \{(g_{tk}, t, k) | g_{tk} \in \mathbb{R}^{n \times m}, t \in \mathbb{N}, k \in \mathbb{N} \wedge t \leq n, k \leq m\} \quad (6)$$

For future calendar loss value CL_{t+1}, there are associated input features which are represented as $g_{(t+1)k}$. In the set notation form, the input feature set corresponding to calendar loss is represented as:

$$G_{cal(t+1)k} = \{g_{(t+1)k} | g_{(t+1)k} \in \mathbb{R}^{n \times m}\} \quad (7)$$

where k shows the feature index which is represented as $k = 1, 2, \ldots, m$ and m denotes the total number of input features for each calendar loss output value. The feature sets of calendar loss values are characterized as $f_{1k} \rightarrow CL_1, f_{2k} \rightarrow CL_2, \ldots, f_{nk} \rightarrow CL_n$. According to the proposed feature selection method, the input feature set used to predict the future time step's calendar loss CL_{t+1}, is denoted by $G_{cal[C(t+1)]}$ and is given as a function of $g_{tk}, g_{(t+1)k}$ and $C_{cal(t)}$:

$$G_{cal[C(t+1)]} = f(g_{tk}, g_{(t+1)k}, C_{cal(t)}) \quad (8)$$

where f_{tk}, and $f_{(t+1)k}$ are the $n \times m$ input feature vector sets which corresponds to the calendar loss CL_t and CL_{t+1}, respectively. The selected input feature set $G_{cal[c(t+1)]}$ for the prediction of the future time step's calendar loss CL_{t+1} is represented as the function of f_{tk}, $f_{(t+1)k}$, and $C_{cal(t)}$.

2.3. Machine learning algorithms with SFS method

ML techniques can be used to train the complex non-linear degradation behaviour of LIBs gathered from historical data, and they do not necessitate a thorough understanding of the battery's internal activity. ML employs a general fitting function with optimum parameters tailored to predict battery capacity loss and degradation behaviour. In this study, the SFS approach is used in combination with ML models to build an accurate battery cyclic and calendar loss prediction model. Eight representative ML models, including the Linear Regression (LR), Ridge Regression (RR), Lasso Regression (LSR), Support Vector Regression (SVR), Gaussian Process Regression (GPR), Random Forest (RF), ElasticNet, and XGBoost are investigated for the performance evaluation of the SFS strategy, and their application to battery cyclic and calendar loss prediction. The structure and framework of these methods are given as follows:

2.3.1. Linear regression (LR)

LR is a mathematical model that describes the relationship between explanatory feature variables and a target variable. LR aims to make predictions about the target variable based on the known feature variables according to the following equation: [39]:

$$y = r.x + c \quad (9)$$

where y is the target variable, x is the vector set of input feature variables, h is the vector of fitting parameters, and c is the y-intercept term. To predict the battery capacity loss using SFS, the model considers m number of features of the current time step as well as m number of features of the previous time step, along with the previous step's target variable, which makes a total of $2m + 1$ input features for the model evaluation. Assuming that the total number of selected input features $2m + 1$, is represented by h that is: $h = 2m + 1$. The model of LR with h number of feature variables, and n observations is as follows:

$$C_{i+1} = f_o + C_i + c_1 f_{i1} + c_2 f_{i2} \cdots + c_m f_{im} + c_{m+1} f_{(i+1)1} + c_{m+2} f_{(i+1)2} \cdots + c_{2m} f_{(i+1)m} + e_i \quad (10)$$

where $i = 1, 2, \ldots n$, C_{i+1} is the target variable, f_o is the y-intercept term, $[c_1, c_2, \ldots, c_{2m}]$ are the regression coefficients, $[f_{i1}, f_{i2}, \ldots, f_{(i+1)1}, \ldots f_{(i+1)m}]$ are the input feature variables which are selected through SFS method. e_i is the error term which is used to account for the difference between the actual value and the prediction. LR modelling is fast and simple, but when the number of features is large and the number of samples is small, it decreases the generalization performance of the model, resulting in the over-fitting [40].

2.3.2. Lasso regression (LSR)

Regularization lowers overfitting by penalizing parameter size during parameter prediction. To solve the over-fitting problem, a regularization term of the L_1 norm to the main function is added. If the parameter penalization is the $L1$-norm, the parameters are not only converged towards zero but are set to zero and thus employed as a feature selection approach. The method is known as the least absolute shrinkage and selection operator (LASSO) method. The LASSO regression estimates the coefficients by minimizing the following [41]:

$$\hat{\beta}^{lasso} = argmin_{\beta} \{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{h} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{h} |\beta_j| \} \quad (11)$$

where λ is a penalization parameter that controls the degree of regularization. y_i represents the predicted cyclic or calendar loss target variable, $[x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots x_{ih}]$ represents the SFS based feature set, and $\beta = (\beta_1, \ldots, \beta_h)$ is a h-dimensional row vector of parameters to be identified where $h = 2m + 1$ represents the number of the SFS based features. It can be seen from Eq. (11) that the goal is to find the β that minimizes $\hat{\beta}^{lasso}$, so when the λ is large, the more the size of the parameters is penalized, thereby, forcing more of the parameters to be zero.

2.3.3. Ridge regression

Ridge regression is the regularized form of LR, and adds a regularization term of the L_2 norm to the main function, as given in Eq. (12). Ridge regression shrinks the regression coefficients by imposing a L_2 penalty. The penalty is added to the least-squared algorithm, which is equal to the square of the coefficient. The ridge coefficients minimize the penalized residual sum of squares (SSE) as given in the following equation [42,43]:

$$\hat{\beta}^{ridge} = argmin_{\beta}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{h} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{h} \beta_j^2\} \quad (12)$$

where λ is a regularization parameter of the added penalty that controls the shrinkage of regression coefficients. y_i represents the predicted battery capacity loss target variable, $[x_{i1}, x_{i2},\ldots, x_{ij}, \ldots x_{ih}]$ represents the SFS based feature set, and $\beta = (\beta_1, \ldots, \beta_h)$ is a h-dimensional row vector of parameters to be identified. Ridge regression deliberately introduces bias into the prediction of β to reduce the variability in the battery cyclic and calendar loss prediction.

2.3.4. Elastic-Net regression

Elastic-Net regression is a regularized LR model that integrates both $L1$-norm and $L2$-norm regularization, known as Lasso and Ridge regression, respectively [44]. The elastic net takes the following form:

$$\hat{\beta}^{elastic} = argmin_{\beta}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{h} x_{ij}\beta_j)^2 + \lambda(\alpha \sum_{j=1}^{h} |\beta_j| + (1-\alpha)\sum_{j=1}^{h} |\beta_j^2|)\} \quad (13)$$

where the argmin function aims to find the value of β that minimizes the argument. The first term inside the square bracket is a form of least squares, y_i is an n-dimensional predicted battery cyclic/calendar loss, $[x_{i1}, x_{i2},\ldots, x_{ij}, \ldots x_{ih}]$ is an $n \times h$ matrix of features, and $\beta = (\beta_1, \ldots, \beta_h)$, is an $h \times 1$ vector of model coefficients. The second term is the regularization term, which contains two non-negative hyperparameters λ, and $alpha$, of the Elastic-Net model. λ is a regularization parameter, and α is a scalar between 0 to 1, which regulates the relative importance of the L1 and L2 norm penalties. For LASSO regression, specific feature coefficients are set to zero, whereas ridge regression shrinks feature weights closer to zero. For a value of α between 0 and 1, the elastic net combines both selection and shrinkage.

2.3.5. Gaussian process regression (GPR)

GPR is an effective technique for dealing with complicated battery degradation modelling problems due to its non-parametric nature, which allows for greater ability in capturing complex nonlinear relationships and quantifying the uncertainty in predictions [10]. GPR can predict the battery cyclic and calendar loss by using an appropriate combination of Gaussian processes (GP) to model their behaviour, which is denoted as:

$$f(x) \sim N(m(x), k(x_i, x_j)) \quad (14)$$

where $m(x)$ and $k(x_i, x_j)$ are the mean and covariance functions respectively, denoted by:

$$m(x) = E(f(x)) \quad (15)$$

$$k(x_i, x_j) = E[(f(x_i) - m(x_i))(f(x_j) - m(x_j))] \quad (16)$$

The GP $f(x)$ is derived by extending the multivariate Gaussian distribution to infinite dimensions and combining the mean function $m(x)$ and the covariance function $k(xi, xj)$. Because the GP is flexible enough to model the genuine mean, the mean function is commonly defined as $m(x) = 0$. The most common choice of co-variance function is the squared exponential kernel which is given as follows [45]:

$$k_{ij} = \theta_f^2 exp\left(-\frac{1}{2\theta_l^2}\|x_i - x_j\|^2\right) \quad (17)$$

where the covariance function parameters θ_f^2 and θ_l^2, are two hyperparameters to be tuned in the GPR, which control the y-scaling and x-scaling, respectively [33]. The GPR method delivers the training probability distribution of possible battery cyclic and calendar loss prediction, which is expressed through the following function [46]:

$$y \sim N(0, K(x_i, x_j) + \theta_n^2 I_n) \quad (18)$$

where y is a vector of predicted battery cyclic and calendar losses, x denotes the input features, $K(x_i, x_j) = (k_{ij})_{n\times n}$ is an n-dimensional symmetric positive definite matrix, I_n is an n-dimensional unit matrix, and $\theta_n^2 I_n$ is the noise covariance matrix. GPR is further used for the prediction of testing samples by computing the posterior distribution of y through Bayesian theory. The mean value of the posterior distribution of y is the predicted battery cyclic and calendar loss.

2.3.6. Support vector regression (SVR)

Battery health and capacity loss prediction problems are primarily classified as regression problems, and when support vector machine (SVM) is used for regression tasks such as battery cyclic and calendar loss prediction, it is referred to as support vector regression (SVR). SVR is suitable for prediction tasks because of its ability to describe the nonlinear correlation of input and output data. Kernels are commonly employed in SVM to aid in the evaluation of nonlinear issues with low feature space by changing them into linear problems with high feature space as formulated in Eqs. (19), and (20) [47].

$$y = \omega_n \phi(x) + b \quad (19)$$

$$y = \sum_{n=1}^{N} \omega_n K(x_i, x_j) + \varepsilon \quad (20)$$

where y is the predicted battery cyclic and calendar loss, and ω_n are the weights of the model connecting feature space to output. x, b, and $K(x_i, x_j)$ denote input features, intercept, and kernel function, respectively. The purpose of SVR is to develop a $epsilon$-insensitive error function in which the maximum deviation of predicted battery cyclic and calendar loss y in the training data is less than a preset threshold $epsilon$ while maintaining the function's smoothness to the greatest extent possible.

2.3.7. Random forest regression

Random forest (RF) regression is an ensemble learning method that integrates and averages decisions from numerous decision tree (DT) multiple decision trees (DT) models [48]. The RF training approach for battery cyclic and calendar loss prediction is to build N distinct decision trees, with each tree in RF being developed with a randomized subset of predictors. With the addition of such randomness, RF can expand the diversity of trees and capture more patterns in the data. RF regression can be expressed as follows [49]:

$$Y(x) = \frac{1}{N_{tree}} \sum_{i=1}^{N_{tree}} f_i(x) \quad (21)$$

where $Y(x)$ is the RF model, N_{tree} is the number of decision trees, and $f_i(x)$ is the i_{th} DT model. $f_i(x)$ is built by randomly sampling a training data subset for each decision tree. The battery cyclic and calendar loss is predicted using RF by averaging the predictions of N_{tree} trees in the forest. The predicted accuracy can be increased by averaging the multiple DT models on the appropriate sub-samples of the dataset.

2.3.8. Extreme gradient boosting regression (XGBoost)

Extreme gradient boosting (XGBoost) is a tree-based ensemble model that uses the boosting statistical approach. It is an implementation of gradient-boosted decision trees designed for speed and performance and is known for its excellent performance [50]. XGBoost generates a tree by combining split characteristics and aggregating multiple 'weak' trees to form a single 'strong' tree with greater stability. During the

XGBoost training process, a new simple tree is built in each step to compensate for prior simple trees' prediction residuals, therefore minimizing the loss function [51]. In addition, the prediction result of each tree is reduced by a learning rate factor to prevent over-fitting. The XGBoost algorithm uses advanced regularization techniques to suppress weights, prevent over-fitting, and enhance its performance in real-world scenarios. XGBoost aggregates the results of each decision tree along the way to calculate the final result. Finally, the cyclic and calendar loss output of the XGBoost is formed by aggregating predictions from t base trees using a weighted sum. It is clear that the error minimization performance of XGBoost is high enough and even with a little amount of data, the algorithm predicts with high accuracy.

2.4. Performance evaluation

The prediction accuracy of the aforementioned ML algorithms with the SFS technique can be evaluated by comparing the actual cyclic and calendar loss values from the data values with the expected ones. Mean absolute error (MAE) IS the metric applied in this work for evaluating the quality of ML methods with SFS method-based predictions. MAE averages the absolute differences between the tested and predicted values and is defined by Eq. (22). All the errors have the same weight in MAE, and it is evident that the smaller the MAE values, the more accurate the prediction result.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{22}$$

where n represents the number of observations, y_i represents the real cyclic and calendar loss values and $\hat{y_i}$ represents the predicted cyclic and calendar loss values.

3. Experimental data and analysis

3.1. Dataset description

In this study, a real-world dataset related to the lithium–manganese oxide (LMO) graphite-based EV battery and its battery usage has been considered and investigated [6]. LMO- graphite battery is extensively used in EVs such as the Nissan Leaf and the Chevrolet Volt. In the dataset under consideration, an EV battery pack which includes 192 cells with an initial capacity of 24.15 kWh is considered and the average voltage of each battery cell is 3.7 V, operating between 3.4 V and 4.1 V. The dataset consists of ten years of EV battery degradation data for the LMO–graphite battery, incorporating both cycle and calendar loss of an EV battery in each U.S. state. The dataset comprises parameters that are related to EV battery usage and capacity degradation in each U.S. state under different driving patterns and temperatures. For the current study, the data of five U.S. states which includes California (CA), Arizona (AZ), Alaska (AL), Arkansas (AR), and Alabama (AL) is considered for evaluation from the dataset under consideration. The dataset contains the battery cyclic loss and calendar loss percentage of the EV under the average driving conditions for each of the five U.S. states of ten years as illustrated in Figs. 4 and 5. The dataset also includes a monthly–hourly timescale of ambient temperature and separated travel demands for local and highway driving conditions. In addition, the driving factors in the dataset consist of the annual charging/discharging cycle number, which is dependent on the yearly travel demand and the driving range of the EV, variations in discharging rates relative to the power outputs required from the battery pack under different driving speeds of the EV, and the varying temperatures to which the battery is exposed all year round.

In order to precisely calculate the battery capacity loss in each state of the US, a comprehensive battery capacity loss model is used. The cycling capacity loss takes place during the EV charge–discharge cycles, which can be calculated by the following equation:

$$CL_{cyc} = \frac{\sum_{m-1}^{C} I(t_m - t_{m+1})}{I \times t_1} \tag{23}$$

where C is the charge–discharge cycle numbers of EV battery required in one year to meet the travel demand, I is the average charging current density, and t_m is the time needed to get the EV battery fully charged in mth cycle. The annual EV charge–discharge cycles are calculated using the National Oceanic and Atmospheric Administration ($NOAA$) data on the US monthly hourly local temperature distribution. It can be calculated by the following equation [52]:

$$C = \sum_{n=1}^{12}\sum_{h=1}^{24} C_{n,h} \tag{24}$$

where $C_{n,h}$ is state-level monthly hourly EV charge–discharge cycles, which is given as:

$$C_{n,h} = \frac{D_{n,h}}{R(T)} \tag{25}$$

$R(T)$ is the temperature-dependent EV driving range, which represents different load conditions needed by EV sub-systems and vehicle internal losses, T is the monthly hourly temperature, and $D_{n,h}$ is the monthly hourly travel demand [53]. The driving range of EVs is largely dependent on the EV driving conditions. In this study, the actual testing data of Nissan Leaf is used [54]. The data is fitted to calculate the EV driving range under various temperatures which correspond to the actual driving range data of 2013 and 2014 Nissan Leaf models collected by FleetCarma [55]. The driving range, $R(T)$ is given by following equation:

$$R(T) = -1.182\times10^{-4}\times T^4 + 3.754\times10^{-5}\times T^3 + 0.087\times T^2 + 2.838\times T + 111.542 \tag{26}$$

The calendar capacity loss takes place during battery energy storage and is mainly caused by battery self-discharge and side reactions. The battery calendar capacity loss follows Arrhenius-form kinetics [56], and an empirical expression based on the experimental data is formulated as:

$$CL_{cal} = 14876 \times exp(\frac{E_a}{RT})\psi_d(t_h)^{0.5} \tag{27}$$

where CL_{cal} is the percentage of calendar capacity loss, E_a is the activation energy i.e. E_a= 24.5kJ, R is the gas constant, ψ_d is the time adjustment function, t_h stands for hour.

3.2. Data pre-processing and feature selection

The data under consideration has been pre-processed by converting, normalizing, and combining the selection of feature values of previous and current time intervals, which corresponds to each target value of battery cyclic and calendar loss value. The data has been processed for datasets given a time length of ten years. The SFS method extracts cyclic and calendar loss indicators from the dataset. Yearly features with one value of each feature per year are selected. The mean of the respective feature value fills in missing values. Based on the quantitative correlation analysis, cyclic and calendar loss indicators that have a strong relationship with the practical battery cyclic and calendar loss, respectively, are adopted as the feature inputs to the model.

The features considered for the study include local and highway distance travelled, charging and discharging efficiency, internal resistance, energy consumption, temperature, and cyclic/calendar loss. Each of these features has a significant impact on battery performance and degradation, and the inclusion of these features improves the accuracy of battery degradation prediction. For instance, the local and highway distance travelled feature provides information about battery usage patterns, which is critical to predicting battery degradation. The charging and discharging efficiency feature is essential in assessing the battery's health, as it indicates the battery's ability to convert stored energy into usable energy. The internal resistance feature is an important indicator of battery health and degradation, as it can be used to estimate the state of health of the battery. The temperature feature is crucial to

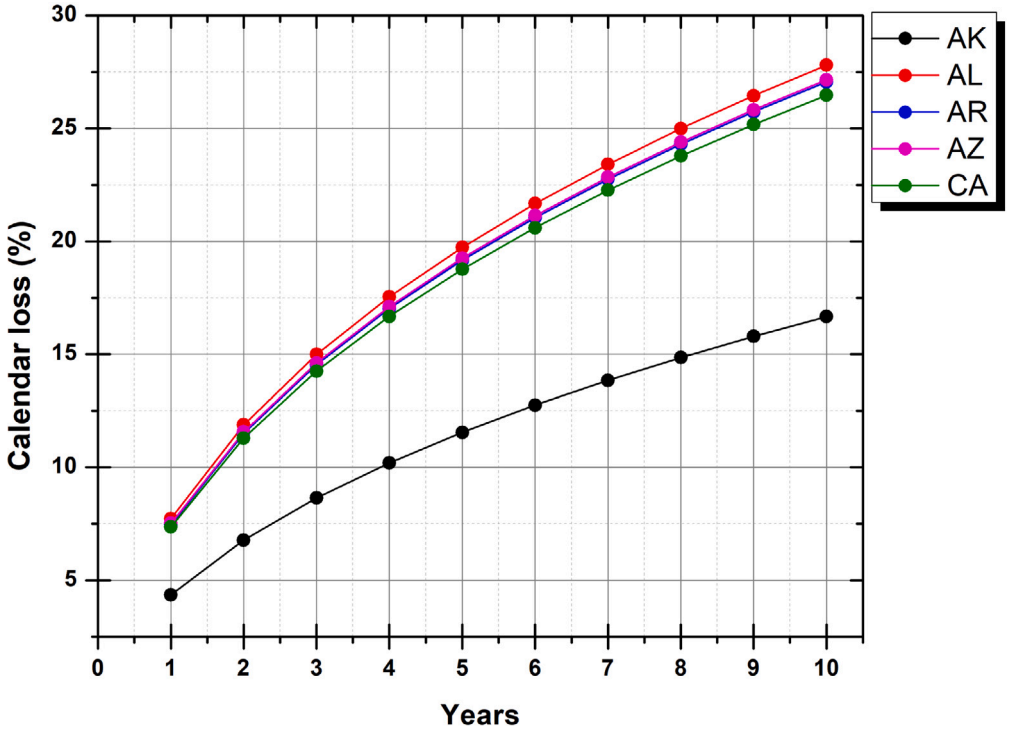


Fig. 4. EV Battery calendar loss percentage of five U.S states, California (CA), Arizona (AZ), Alaska (AL), Arkansas (AR), and Alabama (AL).

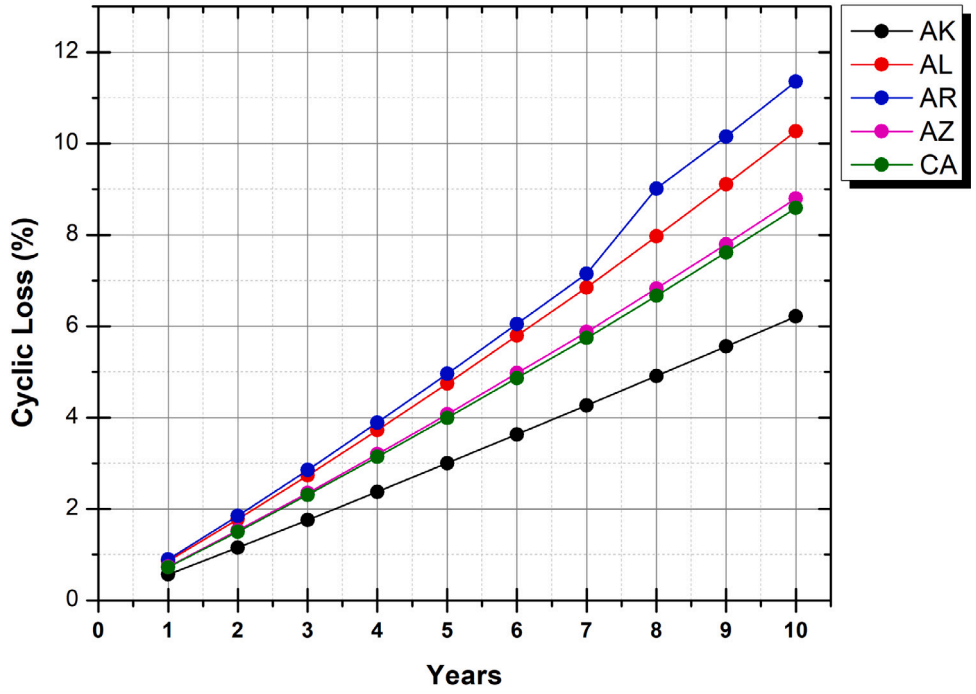


Fig. 5. EV Battery cyclic loss percentage of five U.S states, California (CA), Arizona (AZ), Alaska (AL), Arkansas (AR), and Alabama (AL).

predicting battery degradation, as temperature fluctuations can have a significant impact on battery performance and lifespan. Overall, the selected features represent the most relevant and informative data points for accurately predicting battery degradation in LIBs used in EVs.

Based on the proposed SFS method, a full set of 33 feature parameters is generated for the battery cyclic and calendar loss prediction which includes the previous time interval's cyclic and calendar loss and all the input features of the previous and current time interval. The 33 features are classified according to their extraction sources and techniques to reflect the battery cyclic and calendar ageing dynamics from different perspectives as listed in Table 1. The problem under study is a small-sample-size application with a total experimental dataset of only a few battery samples. Excessive features may cause prediction models to overfit. Furthermore, some of the retrieved features may be redundant, resulting in poor model performance. Given this case, SFS must be applied to the entire feature collection to generate an optimal feature subset selection. All these input features are used to predict the

Table 1
SFS based features for battery cyclic and calendar loss prediction.

Feature type	No.	Feature description
Distance	F1	Local distance travelled in previous time step
	F2	Local distance travelled in current time step
	F3	Highway distance travelled in previous time step
	F4	Highway distance travelled in current time step
	F5	Total distance travelled annually in previous year
	F6	Total distance travelled annually in current year
Charging	F7	Internal resistance while charging for previous time step
	F8	Internal resistance while charging for current time step
	F9	Charging efficiency for previous time step
	F10	Charging efficiency for current time step
Discharging	F11	Internal resistance while discharging for previous time step
	F12	Internal resistance while discharging for current time step
	F13	Discharging efficiency for previous time step
	F14	Discharging efficiency for current time step
Energy consumption	F15	Energy Consumption per charge for previous time step
	F16	Energy Consumption per charge for previous time step
	F17	Energy Consumption considering per travel demand for previous time step
	F18	Energy Consumption considering per travel demand for current time step
	F19	Energy Consumption considering battery degradation for previous time step
	F20	Energy Consumption considering battery degradation for current time step
Temperature	F21–F32	Average monthly temperature for month 1 to 12
Cyclic/Calendar loss	F33	Cyclic/Calendar loss for previous time step

Table 2
MAE comparison of the ML methods using the conventional and proposed feature selection for cyclic loss.

ML model	MAE		Percentage improvement (%)
	Conventional feature selection	SFS	
Linear Regression	0.029	0.023	20.68%
Ridge Regression	0.044	0.038	13.63%
Lasso Regression	0.213	0.193	9.38%
SVR	0.035	0.022	37.14%
GPR	0.027	0.014	48.14%
RF	0.018	0.010	44.44%
ElasticNet	0.213	0.154	27.69%
XGBoost	0.023	0.011	52.17%

Table 3
MAE Comparison of the ML methods using the conventional and proposed feature selection for calendar loss.

ML model	MAE		Percentage improvement (%)
	Conventional feature selection	SFS	
Linear Regression	0.059	0.027	54.23%
Ridge Regression	0.054	0.028	48.14%
Lasso Regression	0.239	0.073	69.45%
SVR	0.072	0.048	33.33%
GPR	0.081	0.036	55.55%
RF	0.031	0.016	48.38%
ElasticNet	0.201	0.186	7.46%
XGBoost	0.029	0.008	72.41%

current year's cyclic and calendar loss which is accounted for as an output label.

Each feature mentioned in Table 1 has a mechanistic explanation of its impact on battery degradation, which is given as:

(a) F2, F3, F4, F5, and F6: Distance travelled is a critical factor that can affect battery degradation. The amount of driving a battery experiences can result in chemical changes within the battery that can accelerate its degradation. Therefore, we consider the distance travelled in the current and previous time steps to account for the effect of distance on battery capacity loss [6].
(b) F7 to F14: Charging/Discharging related features, including internal resistance while charging and discharging can impact battery degradation. The internal resistance of the battery while charging and the efficiency of the charging process can lead to increased heat generation, which can accelerate the chemical changes within the battery that lead to capacity loss [57].
(c) F15, F16, F17, F18, F19, and F20: Energy consumption per charge, energy consumption per travel demand, and energy consumption considering battery degradation are all factors that can impact battery degradation. The energy consumption of the battery can cause temperature changes that can affect the battery's chemical composition and accelerate its degradation [58].
(d) F21-F32: Temperature is a critical factor that can impact battery degradation. By considering the average temperature, we can better understand the impact of temperature changes on battery capacity loss [59]
(e) F33: The cyclic and calendar loss from the previous time step is also considered as a feature to account for the impact of previous degradation on the current capacity loss [60]

4. Results and discussions

The SFS method is applied to the dataset to obtain an optimal feature set of 33 features which are selected based on 13 reference measurements of the cyclic and calendar loss dataset. To evaluate the effectiveness and performance of the SFS method, the derived feature subset is used with the ML algorithms to predict the battery capacity loss. The training and testing data are split into the ratio of 80% and 20%, respectively. The training model trains to predict the current year battery cyclic and calendar loss using the extracted features while the testing process validates the performance of the battery cyclic and calendar loss prediction model. The features obtained from the SFS method are used for battery cyclic and calendar loss prediction with the ML algorithms like Linear Regression (LR), Ridge Regression (RR), Lasso Regression (LSR), Support Vector Regression (SVR), Gaussian Process Regression (GPR), Random Forest Regression (RF), ElasticNet Regression, XGBoost.

Table 2 tabulates the average MAE results of battery cyclic loss for all of the ML methods applied to the testing data by applying the SFS method, while Table 3 demonstrates the average MAE of predicted calendar loss for all the ML methods by applying the SFS method. To

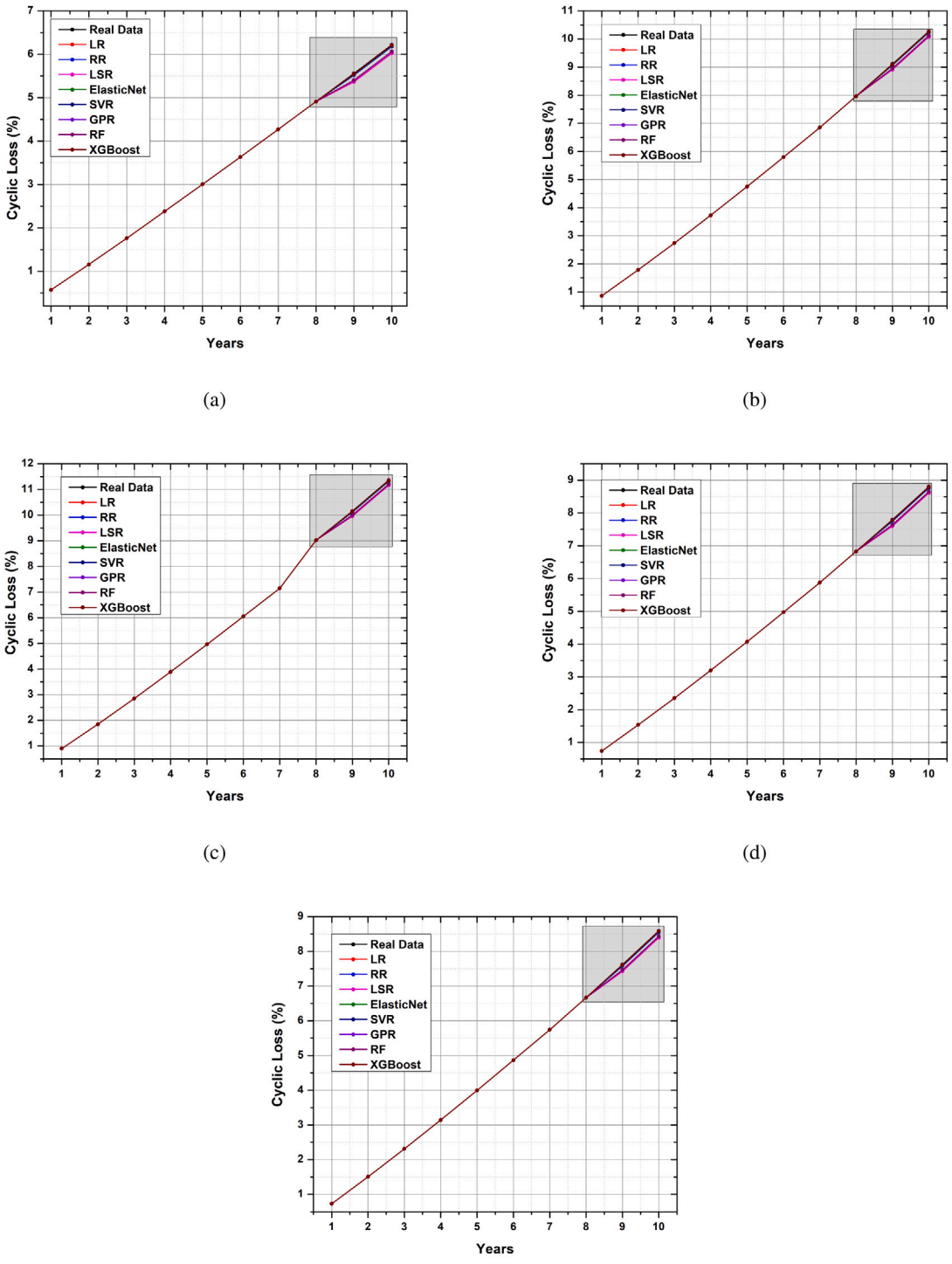


Fig. 6. SFS based ML battery cyclic loss prediction in the US states of (a) Alaska (b) Alabama (c) Arkansas (d) Arizona (e) California.

explore the impact of SFS on ML prediction models, the performance of the SFS method is compared with the conventional feature selection approach for battery cyclic loss as depicted in Table 2.

Fig. 6(a) to (e) illustrates the SFS-ML-based battery cyclic prediction results in the U.S. states of Alaska, Alabama, Arkansas, Arizona, and California, respectively. The prediction results of the SFS-based ML framework are depicted and compared for the last two years of battery cyclic loss data as the first eight years of battery data is designated for training, while the last two years of EV battery cyclic loss data is used for testing. The cyclic loss prediction results of the SFS-based ML framework are highlighted as the shaded area in Fig. 6.

MAE evaluation results of ML algorithms for SFS and conventional feature selection are shown in Fig. 7 while the percentage improvement in the prediction accuracy of the battery cyclic loss with the utilization

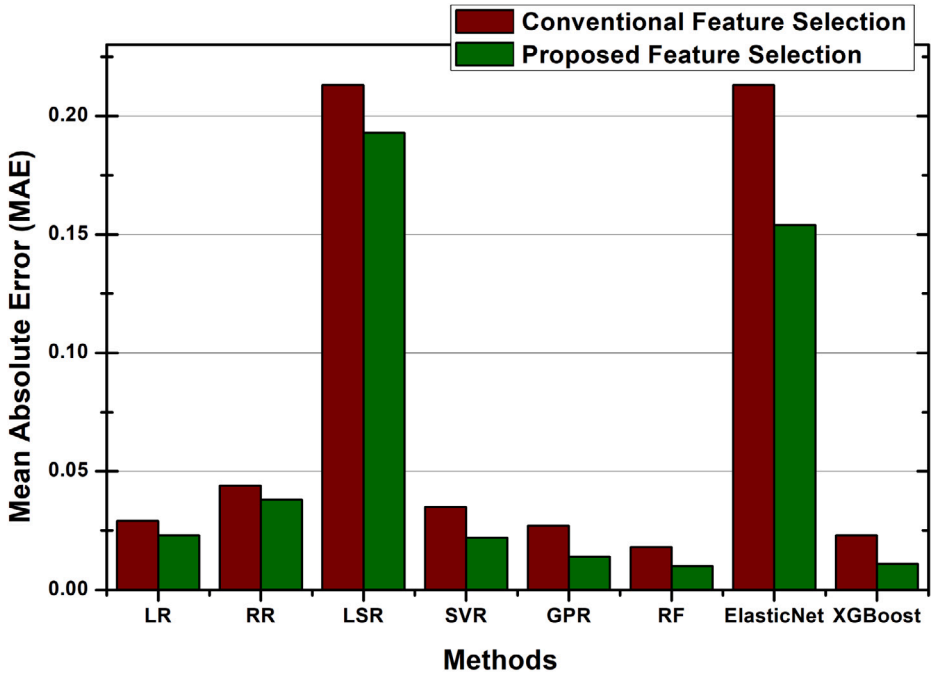


Fig. 7. Accuracy performance of ML models used with SFS method for predicting cyclic loss.

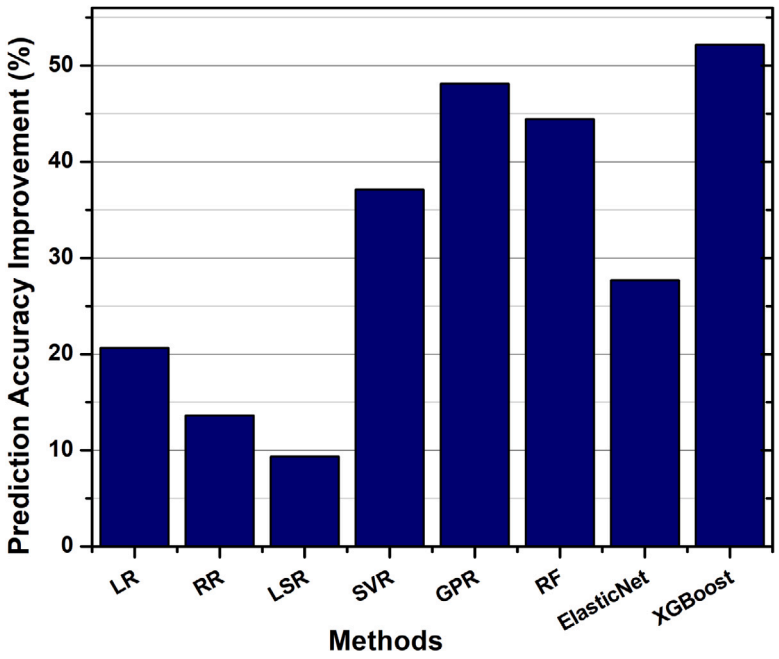


Fig. 8. Percentage improvement of the SFS method over conventional feature selection for battery cyclic loss prediction.

of the SFS method is represented in Fig. 8. It is observed that the performance accuracy of ML methods has improved with the SFS approach, and the highest improvement percentage in prediction accuracy is stated for XGBoost, and GPR which are 52.17%, and 48.14%, respectively. In addition, by applying SFS, RF and SVR algorithms, there is a respective performance improvement of 37.14% and 44.44% in the prediction accuracy. RF and XGBoost methods using features selected by the SFS method affords the best predictive performance for battery cyclic loss prediction with the MAE of 0.010, and 0.011, respectively. For battery cyclic loss prediction using conventional feature selection and the SFS method, RF, XGBoost, SVR and GPR outperform the other ML models as they show the lowest MAE. This may be due to the fact that they are simpler than the other ML models and hence more generalized to the small-sample-size problem. Furthermore, when all the 33 features obtained using the SFS method are used as input to the ML algorithms, prediction results depict an improvement which demonstrates that the SFS method enhances the prediction accuracy of each ML model.

Fig. 9(a) to (e) depicts the SFS-based ML battery calendar prediction results in the U.S. states of Alaska, Alabama, Arkansas, Arizona, and California, respectively. Similar to the battery cyclic prediction results, the first eight years of battery data are used for training, while the last two years of EV battery cyclic loss data are used for testing. The calendar loss prediction results of the SFS-based ML framework are

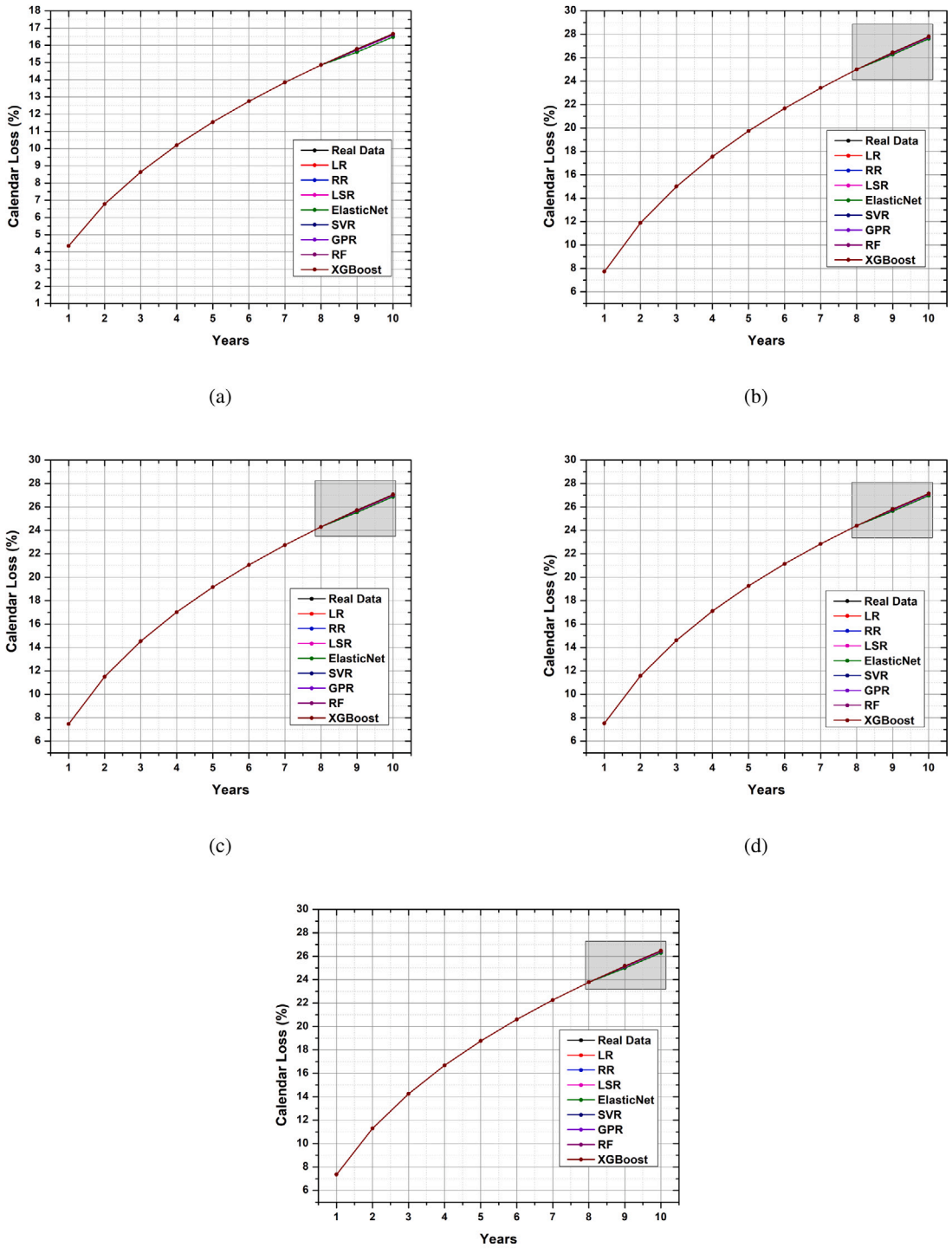


Fig. 9. SFS-ML based battery calendar loss prediction in the US states of (a) Alaska (b) Alabama (c) Arkansas (d) Arizona (e) California.

depicted and compared for the last two years of battery calendar loss data which is highlighted as the shaded area in Fig. 9.

Table 3 tabulates the MAE of the prediction results of battery calendar loss for the ML methods applied to the testing data by incorporating the SFS method. To explore the impact of SFS on ML prediction models, the performance of the SFS method is compared with the conventional feature extraction approach in terms of accuracy and performance for battery calendar loss prediction. It is observed that the XGBoost, and RF outperform the other ML models for both conventional feature selection and the proposed SFS method as they show the lower MAE of 0.008, and 0.016 respectively.

XGBoost showed better accuracy than RF and GPR while predicting the battery capacity loss with the SFS method. XGBoost is a sequential model, which means that each subsequent tree is dependent on the outcome of the last. XGBoost aggregates the results of each decision tree along the way to calculate the final result and does not aggregate the results at the end of the process. In addition, when features obtained from the SFS method are taken as input to the ML algorithms an

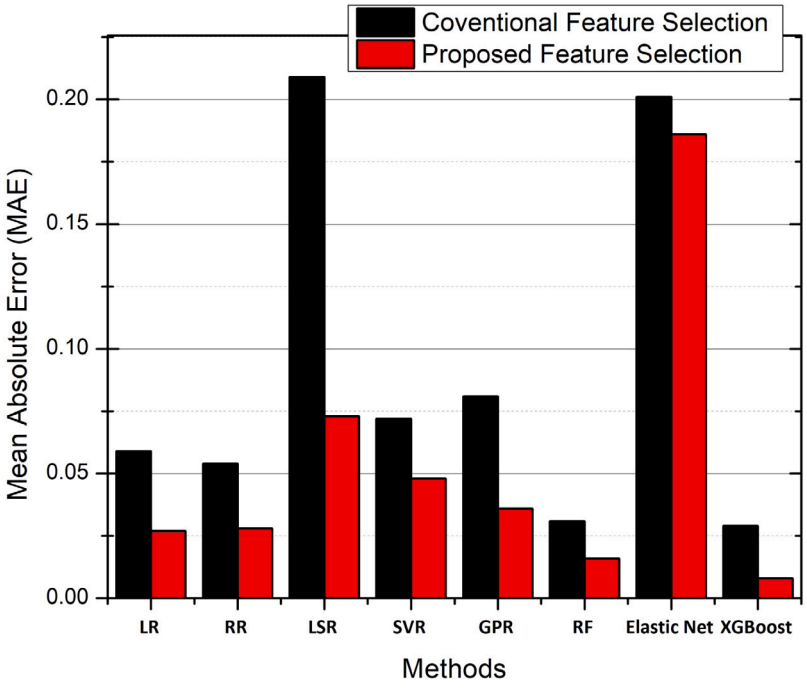


Fig. 10. Accuracy performance of ML models used with SFS method for predicting calendar loss.

improvement in the prediction accuracy is observed indicating that the SFS enhance the predictive ability of each ML. Table 3 also depicts that the ElasticNet algorithm showed the least improvement in the prediction results through the SFS as the calendar loss output values and corresponding input features selected through the SFS method do not have a very high correlation, which compels the ElasticNet algorithm to choose the entirety of input variables and does not shrink the coefficients. The grouping effect does not take place effectively in such cases as variables cannot be easily identified using the low correlation.

The MAE evaluation results of ML algorithms for SFS and conventional feature selection are compared and depicted in Fig. 10 while the percentage improvement in the prediction accuracy of the battery calendar loss with the utilization of the SFS method is represented in Fig. 11. It is observed that the performance accuracy of ML methods for battery calendar loss prediction has improved with the SFS approach, and the greater improvement percentage in prediction accuracy is stated for XGBoost, LASSO regression, GPR and linear regression which is 72.41%, 69.45%, 55.55%, and 54.23%, respectively. In addition, by applying SFS, RF and SVR algorithms have shown respective performance improvements of 48.38% and 33.33% in prediction accuracy. Using the SFS method, XGBoost and RF methods depict the best predictive performance for battery calendar loss prediction with the MAE of 0.008, and 0.016, respectively. It is also observed that without taking the past output label as the feature, the error is greater as compared to when the past output label is taken as an input feature, which leads to smaller accumulated errors over time. As shown in Tables 2 and 3, compared with the RF and XGBoost, the performance accuracy of the SVR and the GPR is relatively low. Based on the overall results, we can statistically conclude that the RF and XGBoost have the best predictive performance in terms of both battery cyclic and calendar loss prediction accuracy, as they have the lowest MAE. It is evident that the performance accuracy of ML methods has improved with the SFS approach.

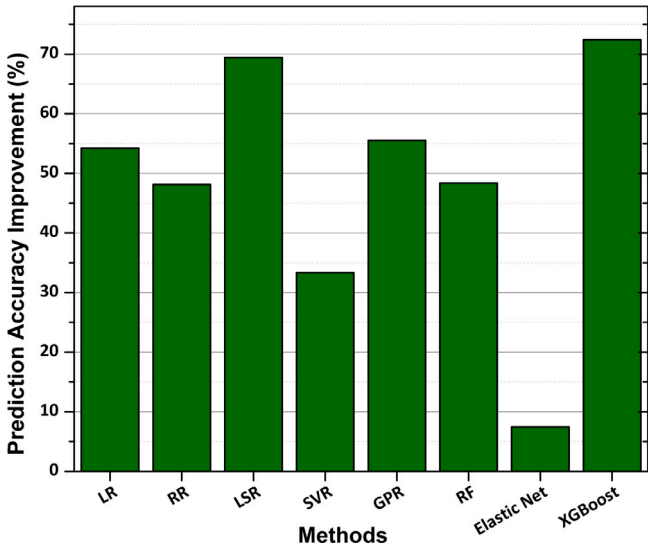


Fig. 11. Percentage improvement of the proposed SFS method over conventional feature selection for battery calendar loss prediction.

5. Conclusion

Accurate prediction of the battery capacity degradation could effectively enhance the safety and reliability of LIBs. ML draws a significant role in battery capacity loss prediction and degradation modelling. It has the potential to be widely applied in future EVs. Based on the utilization of the data pre-processing methods and ML algorithms, this paper presents a smart feature selection (SFS) method to extract characteristic input parameters for battery cyclic and calendar loss prediction, which plays an important role in battery capacity loss and degradation modelling. While devising a battery cyclic and calendar loss prediction model, appropriate indicators are selected as model inputs. The characteristic features for calendar and cyclic loss prediction are comprehensively extracted based on the intensive utilization of the SFS method on the battery datasets by coupling present and historical features. ML algorithms are applied in combination with the SFS method on the processed data to predict calendar and cyclic loss using the extracted features. The model trains on the designated

training data while the testing process validates the performance of the prediction model. A case study is performed on a diverse and dynamic EV dataset in the United States, where 33 features are extracted using the proposed feature selection method. It is worth mentioning that the features are extracted based on 13 reference measurements during the cyclic and calendar loss process mentioned in the dataset. The methodology is assessed using eight widely ML algorithms for battery cyclic and calendar loss prediction. The results depict that the proposed SFS method has improved the prediction accuracy and reduced the MAE for all the ML algorithms applied in this study. The highest improvement in prediction accuracy for the calendar is shown for XGBoost, GPR, and RF algorithm, which is 52.17%, 48.14%, and 44.44%, respectively. For calendar loss prediction, a significant improvement of 72.41%, 48.38%, and 33.33% is also depicted by XGBoost, GPR, and SVR algorithms when applied in combination with the SFS methods. The results also show that RF and XGBoost methods when applied with the proposed SFS method, have shown a higher accuracy for the battery capacity loss prediction. Our proposed study can be used as a reference for obtaining battery capacity loss models in practical applications as the features are obtained for a dynamic real-world EV dataset.

CRediT authorship contribution statement

Huzaifa Rauf: Conceptualization, Methodology, Software, Writing – original draft. **Muhammad Khalid:** Conceptualization, Investigation, Validation. **Naveed Arshad:** Writing – review & editing, Supervision.

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.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to acknowledge the funding support provided by the LUMS Energy Institute (LEI) at Lahore University of Management Sciences (LUMS), and the National Center of Big Data and Cloud Computing (NCBC) of the Higher Education Commission (HEC), Pakistan. The authors would also like to acknowledge the research support received from the Saudi Data and AI Authority (SDAIA), Saudi Arabia and King Fahd University of Petroleum and Minerals (KFUPM) under SDAIA-KFUPM Joint Research Center for Artificial Intelligence, Saudi Arabia, grant No. JRC-AI-RFP-08.

References

[1] Amaia González-Garrido, Andreas Thingvad, Haizea Gaztañaga, Mattia Marinelli, Full-scale electric vehicles penetration in the Danish Island of Bornholm—Optimal scheduling and battery degradation under driving constraints, J. Energy Storage 23 (April) (2019) 381–391.

[2] Huzaifa Rauf, Muhammad Shuzub Gull, Naveed Arshad, Complementing hydroelectric power with floating solar PV for daytime peak electricity demand, Renew. Energy 162 (2020) 1227–1242.

[3] Yunjian Li, Kuining Li, Yi Xie, Jiangyan Liu, Chunyun Fu, Bin Liu, Optimized charging of lithium-ion battery for electric vehicles: Adaptive multistage constant current–constant voltage charging strategy, Renew. Energy 146 (2020) 2688–2699.

[4] Hua Wang, De Zhao, Yutong Cai, Qiang Meng, Ghim Ping Ong, A trajectory-based energy consumption estimation method considering battery degradation for an urban electric vehicle network, Transp. Res. D 74 (2019) 142–153.

[5] Foad H. Gandoman, Joris Jaguemont, Shovon Goutam, Rahul Gopalakrishnan, Yousef Firouz, Theodoros Kalogiannis, Noshin Omar, Joeri Van Mierlo, Concept of reliability and safety assessment of lithium-ion batteries in electric vehicles: Basics, progress, and challenges, Appl. Energy 251 (C) (2019) 1.

[6] Fan Yang, Yuanyuan Xie, Yelin Deng, Chris Yuan, Predictive modeling of battery degradation and greenhouse gas emissions from US state-level electric vehicle operation, Nature Commun. 9 (1) (2018) 1–10.

[7] Miswar A. Syed, Muhammad Khalid, Neural network predictive control for smoothing of solar power fluctuations with battery energy storage, J. Energy Storage 42 (2021) 103014.

[8] Yuecheng Li, Hongwen He, Jiankun Peng, An adaptive online prediction method with variable prediction horizon for future driving cycle of the vehicle, IEEE Access 6 (2018) 33062–33075.

[9] Foad H. Gandoman, Joris Jaguemont, Shovon Goutam, Rahul Gopalakrishnan, Yousef Firouz, Theodoros Kalogiannis, Noshin Omar, Joeri Van Mierlo, Concept of reliability and safety assessment of lithium-ion batteries in electric vehicles: Basics, progress, and challenges, Appl. Energy 251 (2019) 113343.

[10] Huzaifa Rauf, Muhammad Khalid, Naveed Arshad, Machine learning in state of health and remaining useful life estimation: Theoretical and technological development in battery degradation modelling, Renew. Sustain. Energy Rev. 156 (2022) 111903.

[11] Miswar A. Syed, Muhammad Khalid, An intelligent model predictive control strategy for stable solar-wind renewable power dispatch coupled with hydrogen electrolyzer and battery energy storage, Int. J. Energy Res. x (2023) x.

[12] Xiaosong Hu, Yunhong Che, Xianke Lin, Simona Onori, Battery health prediction using fusion-based feature selection and machine learning, IEEE Trans. Transp. Electrification 7 (2) (2020) 382–398.

[13] Xiaosong Hu, Yunhong Che, Xianke Lin, Zhongwei Deng, Health prognosis for electric vehicle battery packs: A data-driven approach, IEEE/ASME Trans. Mechatronics 25 (6) (2020) 2622–2632.

[14] Assia Chadly, Elie Azar, Maher Maalouf, Ahmad Mayyas, Techno-economic analysis of energy storage systems using reversible fuel cells and rechargeable batteries in green buildings, Energy 247 (2022) 123466.

[15] Daniel-Ioan Stroe, Maciej Swierczynski, Søren Knudsen Kær, Remus Teodorescu, Degradation behavior of lithium-ion batteries during calendar ageing—The case of the internal resistance increase, IEEE Trans. Ind. Appl. 54 (1) (2017) 517–525.

[16] Xiaopeng Tang, Kailong Liu, Xin Wang, Furong Gao, James Macro, W. Dhammika Widanage, Model migration neural network for predicting battery aging trajectories, IEEE Trans. Transp. Electrification 6 (2) (2020) 363–374.

[17] Anthony Barré, Benjamin Deguilhem, Sébastien Grolleau, Mathias Gérard, Frédéric Suard, Delphine Riu, A review on lithium-ion battery ageing mechanisms and estimations for automotive applications, J. Power Sources 241 (2013) 680–689.

[18] Xiaosong Hu, Yunhong Che, Xianke Lin, Simona Onori, Battery health prediction using fusion-based feature selection and machine learning, IEEE Trans. Transp. Electrification PP (2020) 1.

[19] Maitane Berecibar, Floris Devriendt, Matthieu Dubarry, Igor Villarreal, Noshin Omar, Wouter Verbeke, Joeri Van Mierlo, Online state of health estimation on NMC cells based on predictive analytics, J. Power Sources 320 (2016) 239–250.

[20] Md Sazzad Hosen, Rekabra Youssef, Theodoros Kalogiannis, Joeri Van Mierlo, Maitane Berecibar, Battery cycle life study through relaxation and forecasting the lifetime via machine learning, J. Energy Storage 40 (2021) 102726.

[21] Haris Mansoor, Huzaifa Rauf, Muhammad Mubashar, Muhammad Khalid, Naveed Arshad, Past vector similarity for short term electrical load forecasting at the individual household level, IEEE Access 9 (2021) 42771–42785.

[22] Xiaoyu Li, Changgui Yuan, Xiaohui Li, Zhenpo Wang, State of health estimation for Li-ion battery using incremental capacity analysis and Gaussian process regression, Energy 190 (2020) 116467.

[23] Keisuke Ando, Tomoyuki Matsuda, Daichi Imamura, Degradation diagnosis of lithium-ion batteries with a LiNi0. 5Co0. 2Mn0. 3O2 and LiMn2O4 blended cathode using dV/dQ curve analysis, J. Power Sources 390 (2018) 278–285.

[24] Meinert Lewerenz, Jens Münnix, Johannes Schmalstieg, Stefan Käbitz, Marcus Knips, Dirk Uwe Sauer, Systematic aging of commercial LiFePO4| graphite cylindrical cells including a theory explaining rise of capacity during aging, J. Power Sources 345 (2017) 254–263.

[25] Yu Zhang, Zhen Peng, Yong Guan, Lifeng Wu, Prognostics of battery cycle life in the early-cycle stage based on hybrid model, Energy 221 (2021) 119901.

[26] Fangfang Yang, Dong Wang, Fan Xu, Zhelin Huang, Kwok-Leung Tsui, Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model, J. Power Sources 476 (2020) 228654.

[27] Xing Shu, Guang Li, Jiangwei Shen, Zhenzhen Lei, Zheng Chen, Yonggang Liu, A uniform estimation framework for state of health of lithium-ion batteries considering feature extraction and parameters optimization, Energy 204 (2020) 117957.

[28] Tingting Xu, Zhen Peng, Lifeng Wu, A novel data-driven method for predicting the circulating capacity of lithium-ion battery under random variable current, Energy 218 (2021) 119530.

[29] Qi Zhao, Xiaoli Qin, Hongbo Zhao, Wenquan Feng, A novel prediction method based on the support vector regression for the remaining useful life of lithium-ion batteries, Microelectron. Reliab. 85 (2018) 99–108.

[30] Guijun Ma, Yong Zhang, Cheng Cheng, Beitong Zhou, Pengchao Hu, Ye Yuan, Remaining useful life prediction of lithium-ion batteries based on false nearest neighbors and a hybrid neural network, Appl. Energy 253 (2019) 113626.

[31] Peiyao Guo, Ze Cheng, Lei Yang, A data-driven remaining capacity estimation approach for lithium-ion batteries based on charging health feature extraction, J. Power Sources 412 (2019) 442–450.
[32] Yuanyuan Li, Daniel-Ioan Stroe, Yuhua Cheng, Hanmin Sheng, Xin Sui, Remus Teodorescu, On the feature selection for battery state of health estimation based on charging–discharging profiles, J. Energy Storage 33 (2021) 102122.
[33] Duo Yang, Xu Zhang, Rui Pan, Yujie Wang, Zonghai Chen, A novel Gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve, J. Power Sources 384 (2018) 387–395.
[34] Ji Wu, Chenbin Zhang, Zonghai Chen, An online method for lithium-ion battery remaining useful life estimation using importance sampling and neural networks, Appl. Energy 173 (2016) 134–140.
[35] Xing Shu, Jiangwei Shen, Guang Li, Yuanjian Zhang, Zheng Chen, Yonggang Liu, A flexible state of health prediction scheme for lithium-ion battery packs with long short-term memory network and transfer learning, IEEE Trans. Transport. Electrification (2021).
[36] Sheng Shen, Mohammadkazem Sadoughi, Meng Li, Zhengdao Wang, Chao Hu, Deep convolutional neural networks with ensemble learning and transfer learning for capacity estimation of lithium-ion batteries, Appl. Energy 260 (2020) 114296.
[37] Jinhao Meng, Lei Cai, Daniel-Ioan Stroe, Guangzhao Luo, Xin Sui, Remus Teodorescu, Lithium-ion battery state-of-health estimation in electric vehicle using optimized partial charging voltage profiles, Energy 185 (2019) 1054–1062.
[38] Shun Jia, Bo Ma, Wei Guo, Zhaojun Steven Li, A sample entropy based prognostics method for lithium-ion batteries using relevance vector machine, J. Manuf. Syst. (2021).
[39] Man-Fai Ng, Jin Zhao, Qingyu Yan, Gareth J Conduit, Zhi Wei Seh, Predicting the state of charge and health of batteries using data-driven machine learning, Nat. Mach. Intell. 2 (3) (2020) 161–170.
[40] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Linear regression, in: An Introduction to Statistical Learning, Springer, 2021, pp. 59–128.
[41] Søren B. Vilsen, Daniel-Ioan Stroe, Battery state-of-health modelling by multiple linear regression, J. Clean. Prod. 290 (2021) 125700.
[42] Yan Jiang, Jiuchun Jiang, Caiping Zhang, Weige Zhang, Yang Gao, Na Li, State of health estimation of second-life LiFePO4 batteries for energy storage applications, J. Clean. Prod. 205 (2018) 754–762.
[43] Ji Wu, Xuchen Cui, Hui Zhang, Mingqiang Lin, Health prognosis with optimized feature selection for lithium-ion battery in electric vehicle applications, IEEE Trans. Power Electron. (2021).
[44] Kristen A. Severson, Peter M. Attia, Norman Jin, Nicholas Perkins, Benben Jiang, Zi Yang, Michael H. Chen, Muratahan Aykol, Patrick K. Herring, Dimitrios Fraggedakis, Martin Z. Bazant, Stephen J. Harris, William C. Chueh, Richard D. Braatz, Data-driven prediction of battery cycle life before capacity degradation, Nature Energy 4 (5) (2019) 383–391.
[45] Robert R. Richardson, Michael A. Osborne, David A. Howey, Gaussian process regression for forecasting battery state of health, J. Power Sources 357 (2017) 209–219.
[46] Zicheng Fei, Fangfang Yang, Kwok-Leung Tsui, Lishuai Li, Zijun Zhang, Early prediction of battery lifetime via a machine learning based framework, Energy 225 (2021) 120205.
[47] Yi Li, Kailong Liu, Aoife M. Foley, Alana Zülke, Maitane Berecibar, Elise Nanini-Maury, Joeri Van Mierlo, Harry E. Hoster, Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review, Renew. Sustain. Energy Rev. 113 (2019) 109254.
[48] Yi Li, Changfu Zou, Maitane Berecibar, Elise Nanini-Maury, Jonathan C.-W. Chan, Peter Van den Bossche, Joeri Van Mierlo, Noshin Omar, Random forest regression for online capacity estimation of lithium-ion batteries, Appl. Energy 232 (2018) 197–210.
[49] Yu-Wei Chung, Behnam Khaki, Tianyi Li, Chicheng Chu, Rajit Gadh, Ensemble machine learning-based algorithm for electric vehicle user behavior prediction, Appl. Energy 254 (2019) 113732.
[50] Tianqi Chen, Carlos Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[51] Fu Jiang, Jiajun Yang, Yijun Cheng, Xiaoyong Zhang, Yingze Yang, Kai Gao, Jun Peng, Zhiwu Huang, An aging-aware soc estimation method for lithium-ion batteries using xgboost algorithm, in: 2019 IEEE International Conference on Prognostics and Health Management, ICPHM, IEEE, 2019, pp. 1–8.
[52] National oceanic and atmospheric administration, 2020, U.S Climate Normals, Report No. 0003-0007.
[53] Xuefang Li, Chenhui Liu, Jianmin Jia, Ownership and usage analysis of alternative fuel vehicles in the united states with the 2017 national household travel survey data, Sustainability 11 (8) (2019) 2262.
[54] Kevin Stutenberg, Advanced technology vehicle lab benchmarking-level 1, in: 2014 US DOE Vehicle Technologies Program Annual Merit Review and Peer Evaluation Meeting, 2014.
[55] M. Allen, Electric range for the nissan leaf & chevrolet volt in cold weather, Fleet Carma 12 (2013).
[56] John Wang, Justin Purewal, Ping Liu, Jocelyn Hicks-Garner, Souren Soukazian, Elena Sherman, Adam Sorenson, Luan Vu, Harshad Tataria, Mark W. Verbrugge, Degradation of lithium ion batteries employing graphite negatives and nickel–cobalt–manganese oxide+ spinel manganese oxide positives: Part 1, aging mechanisms and life estimation, J. Power Sources 269 (2014) 937–948.
[57] Benedikt Lunz, Zexiong Yan, Jochen Bernhard Gerschler, Dirk Uwe Sauer, Influence of plug-in hybrid electric vehicle charging strategies on charging and battery degradation costs, Energy Policy 46 (2012) 511–519.
[58] Wesley D. Connor, Yongqiang Wang, Andreas A. Malikopoulos, Suresh G. Advani, Ajay K. Prasad, Impact of connectivity on energy consumption and battery life for electric vehicles, IEEE Trans. Intell. Veh. 6 (1) (2020) 14–23.
[59] Yudi Qin, Jiuyu Du, Languang Lu, Ming Gao, Frank Haase, Jianqiu Li, Minggao Ouyang, A rapid lithium-ion battery heating method based on bidirectional pulsed current: Heating effect and impact on battery life, Appl. Energy 280 (2020) 115957.
[60] Ian Mathews, Bolun Xu, Wei He, Vanessa Barreto, Tonio Buonassisi, Ian Marius Peters, Technoeconomic model of second-life batteries for utility-scale solar considering calendar and cycle aging, Appl. Energy 269 (2020) 115127.