A Hybrid Ensemble Deep Learning Approach for Early Prediction of Battery Remaining Useful Life

***Abstract*—Accurate estimation of the remaining useful life (RUL) of lithium-ion batteries is critical for their large-scale deployment as energy storage devices in electric vehicles and stationary storage. A fundamental understanding of the factors affecting RUL is crucial for accelerating battery technology development. However, it is very challenging to predict RUL accurately because of complex degradation mechanisms occurring within the batteries, as well as the dynamic operating conditions in practical applications. Moreover, due to insignificant capacity degradation in early stages, early prediction of battery life with early cycle data can be more difficult. In this paper, we propose a hybrid deep learning model for early prediction of battery RUL. The proposed method can effectively combine handcrafted features with domain knowledge and latent features learned by deep networks to boost the performance of RUL early prediction. We also design a non-linear correlation-based method to select effective domain knowledge-based features. Moreover, a novel snapshot ensemble learning strategy is proposed to further enhance the model generalization ability without increasing any additional training cost. Our experimental results show that the proposed method not only outperforms other approaches in the primary test set having a similar distribution as the training set, but also generalizes well to the secondary test set having a clearly different distribution with the training set. The PyTorch implementation of our proposed approach is available at https://github.com/batteryrul/battery_rul_early_prediction.**

***Index Terms*—Remaining useful life, Lithium-ion battery, Deep learning, Early prediction**

I. Introduction

Lithium-ion batteries (LIBs) are widely used as energy storage devices in various commercial applications such as electric vehicles (EVs), stationary storage and portable electronic devices due to their low costs, high energy densities and long cycle lives [1], [2], [3], [4]. Precisely monitoring the capacity degradation process and estimating the remaining useful life (RUL) of LIBs are crucial since the failure of LIBs will result in system performance degradation or even catastrophic hazards. What's more, accurately predicting battery RUL with early cycle data would benefit battery manufacturing. For instance, prediction with early cycle data would accelerate battery development cycle, allow manufacturers to perform rapid validation of their new manufacturing processes, and grade new batteries by their expected lifetimes [5]. However, due to nonlinear degradation mechanisms caused by cycling and varied operation conditions, accurately predicting battery RUL is very challenging. Moreover, making predictions only with early cycle data is much more difficult as a lithium-ion battery often degrades with a very low rate at the early stage and then goes through an accelerated degradation after a certain time point or cycle number, which is called the knee point [6]. In other words, the degree of degradation is negligible from cycle to cycle in the early stage. Fig. 1 demonstrates the LIBs' capacity degradation process over cycles. In particular, the early RUL prediction in this paper refers to only utilizing the first 100 cycle measurements before rapid capacity degradation or the knee point occurring to predict the lifetimes of LIBs, following [5].

Generally, battery RUL is the number of cycles when a battery reaches 80% of its initial capacity, which is defined as the end of life (EOL) of the battery [7]. Previous works on battery RUL prediction can be classified into two categories: physics-based approaches and data-driven-based approaches.

Physics-based approaches such as the single particle model (SPM) [8], [9] and pseudo-two-dimensional (P2D) model [10], [11], [12], [13], [14] are based on the electrochemical principles underlying LIBs and can simulate a battery's current and voltage characteristics from kinetics and transport equations. Another more general physics-based approach is the multiphase porous electrode theory that uses nonequilibrium thermodynamics to account for important microscopic physics such as phase separation [15], [16]. Those approaches are used for parameter estimation and cycle life prediction of LIBs. While such models are typically accurate and interpretable, they are often computationally complex and have many parameters and interactions that might be unknown [17]. In other words, an accurate physical model often requires strong domain knowledge on battery degradation mechanisms.

On the contrary, data-driven-based approaches do not assume battery degradation mechanisms a priori [18] but leverage battery historical cycling data. Various approaches such as the support vector machine (SVM) [19], Box-Cox transformation [20] and other machine learning (ML)-based models [18], [5], [21], [22] have already been widely applied to battery RUL prediction tasks. For instance, Wu *et al.* employed a Feed Forward Neural Network (FFNN) to model the relationship between battery RUL and the difference of constant-current charge voltage curves under different cycles [21]. Zhang *et al.* trained a long short-term memory (LSTM) network to learn the long-term dependencies among the degraded capacity [22]. Although the aforementioned methods have achieved satisfactory performances, they still have some inevitable limitations. On one hand, their model performances heavily rely on the quality of handcrafted features based on domain knowledge. These features are usually pre-devised based on the specialized knowledge of the field and have certain specific physical meanings. Involving unqualified features or excluding informative features can result in a poor prediction accuracy. For instance, Severson *et al.* proposed a set of 21 domain specific features from the first 100 cyclic measurements of batteries in [5] and the accuracy of their model varies largely with the different feature subsets. It reveals the great difficulty of feature engineering for the battery RUL prediction task. Furthermore,

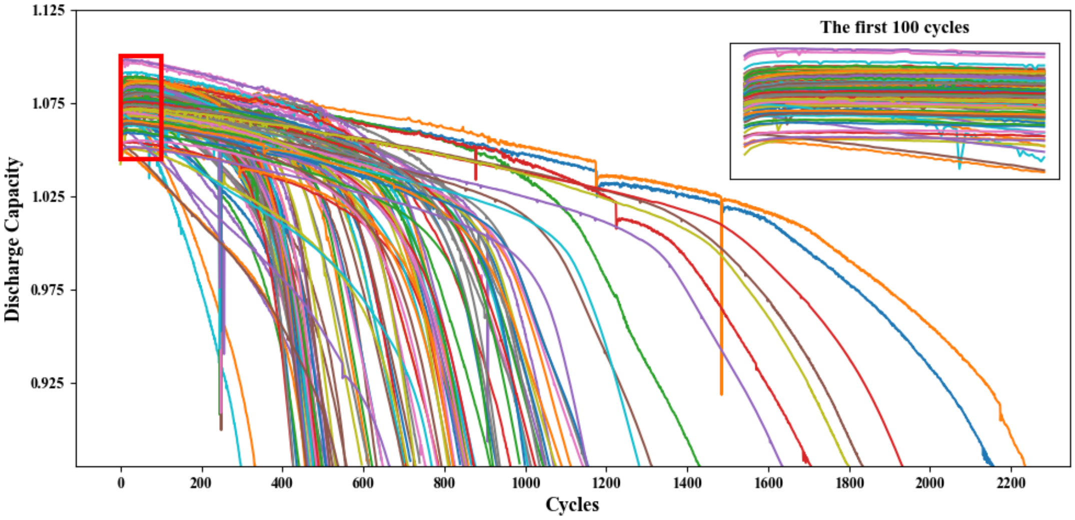


Fig. 1: Battery Capacity Degradation Curves

it can be even more challenging for early prediction since some features employed in previous works may be no longer informative in the early stage of battery operation before rapid degradation. On the other hand, most of previous works tend to leverage derived data like discharge capacity or handcrafted features aggregated from different cycles. They usually avoid leveraging raw measurements (*e.g.*, terminal voltage, current, and temperature) from each cycle due to their tremendous data volumes. Therefore, some important intrinsic physical information embedded in these direct measurements might be missed out due to limited domain knowledge. Moreover, existing works evaluate their proposed methods within a single battery, in which they partition single battery measurements into training and evaluation segments, but seldom across different batteries. The problem of distribution discrepancy among different batteries has not been taken into account. In other words, the generalization capability of their methods is not sufficiently verified.

To address the above issues, in this paper we propose an innovative hybrid deep learning method to make full use of both handcrafted features with domain knowledge and latent features learned by a deep neural network for early prediction of battery RUL. In order to select the most effective domain knowledge-based features, a non-linear correlation-based feature selection method is developed. Furthermore, a novel snapshot ensemble learning strategy is designed to further improve the generalization performance of the proposed hybrid model. Extensive experiments have been conducted to verify the performance of the proposed hybrid deep learning method on battery RUL prediction. The main contributions of this paper are summarized as below:

- We propose an innovative hybrid deep model which can not only leverage prior domain knowledge but also learn latent features from raw measurements. The two types of features can compensate each other, resulting in a superior performance for battery RUL prediction. To the best of our knowledge, we are the first to propose the hybrid deep learning method that can leverage two types of information for early prediction of battery RUL.
- A non-linear correlation-based approach is proposed for feature selection from excessive domain knowledge-based features, which is simple but can effectively improve the performance of battery RUL prediction.
- We develop a novel snapshot ensemble learning strategy upon the proposed deep learning framework to further enhance the generalization capability of the model without increasing any additional training cost.
- Our experimental results show that our proposed method outperforms state-of-the-art methods for LIBs' early prediction task. It can not only achieve good performance in the test set having a similar distribution as the training set, but also generalize well in the test set that has a clearly different distribution with the training set.

The rest of the paper is organized as follows. Section II reviews some related works on battery RUL prediction. Section III introduces the details of the proposed method, including feature generation and selection, the structure of proposed hybrid deep learning model and how to improve model generalization ability. Section IV describes the dataset used for evaluation and experimental setup, followed by the experimental results and ablation study. Section V concludes this paper and presents potential future works.

II. Related Works

In recent years, data-driven approaches have been widely adopted in lithium-ion battery RUL prediction applications. Particle filter (PF) is a commonly used method in battery RUL prediction. Zhang *et al.* leveraged battery capacity degradation curve data and a PF to identify key parameters in a battery exponential model, which was then used for forecasting battery RUL [23]. Song *et al.* proposed a hybrid method of PF algorithm and an enhanced autoregressive (AR) model, which used a nonlinear degradation factor and an iterative updating approach to improve long term prediction performance [24].

Pang *et al.* jointly utilized the Kalman filter (KF) and the expectation-maximization (EM) algorithm to estimate battery degradation state and model parameters [25]. Although data-driven approaches are widely used, a precise model description of battery degradation is a prerequisite.

In addition to the approaches described above, ML-based methods have also received much attention. Patil *et al.* proposed a two-stage prediction approach [19]. Particularly, a set of parameters were extracted from voltage, temperature and time curves from each discharge cycle. Then, a SVM-based classification model was trained to estimate a gross RUL value at the early stage and Support Vector Regression (SVR) was employed to predict the accurate RUL when the battery gradually reached its EOL. Chang *et al.* developed a hybrid model that used a Relevance Vector Machine (RVM) to compensate the prediction error of an Unscented Kalman Filter (UKF) with discharge capacity curves [26]. In [27], Chen *et al.* utilized a SVR-based model to predict battery RUL. Meanwhile, they leveraged phase space reconstruction (PSR) to obtain the optimal input sequence from the reconstructed capacity and discharging voltage difference of equal time interval curves, which were reconstructed by the ensemble empirical mode decomposition (EEMD). Severson *et al.* generated a set of domain specific features from the early 100 cycles data and then a regularized elastic net was employed to map those features to battery cycle life [5]. Ren *et al.* proposed to extract geometric features from charging and discharging processes and used an auto-encoder neural network for feature fusion before feeding them into a neural network (NN) for predicting battery RUL [28]. The performances of these conventional ML-based methods, to a large extent, depend on the quality of extracted features, which are either linear or non-linear transformations of raw measurements based on specific domain knowledge.

To extract as much representative information as possible, deep learning methods have also been introduced for battery RUL prediction. Li *et al.* applied Empirical Mode Decomposition (EMD) to decompose the capacity data into variance high-frequency and low-frequency sub-layers that were respectively fit into an Elman neural network and LSTM for RUL prediction [29]. Similarly, Liu *et al.* also decomposed the capacity data and fitted them into a hybrid model, which used a LSTM network to capture the long-term dependence of capacity degradation and a Gaussian Process Regression (GPR) to capture the prediction uncertainty caused by capacity regeneration phenomena [18]. Ma *et al.* leveraged a fusion model of convolutional neural network (CNN) and LSTM (denoted as CNN-LSTM) that can leverage not only CNN's automatic feature extraction capability but also LSTM's capability of capturing temporal dependency [7]. Ren *et al.* designed an encoder to augment feature dimensions and then leveraged a CNN-LSTM hybrid model to mine deeper useful information from the augmented data [30]. A Gated Recurrent Unit Recurrent Neural Network (GRU-RNN) approach was proposed in [31], which extracted sufficient statistical features from voltage, current and temperature measurements at each cycle. Additionally, the authors used the linear correlation and random forest to further reduce the size of features. Although the effectiveness of deep learning methods has been demonstrated in above works, they highly rely on the information provided by battery degradation curves, especially requiring the segments where the batteries start to degrade. Unlike them, our research focuses on the early stage data for battery RUL prediction. Moreover, as aforementioned, most approaches in previous works are trained and evaluated within a single battery data but seldom across different batches (e.g., collected in different time periods). Due to the deviation between different batteries' initial states, it will be more challenging and require the proposed model to have good generalization capability such that it can achieve good accuracy on unseen batteries as well.

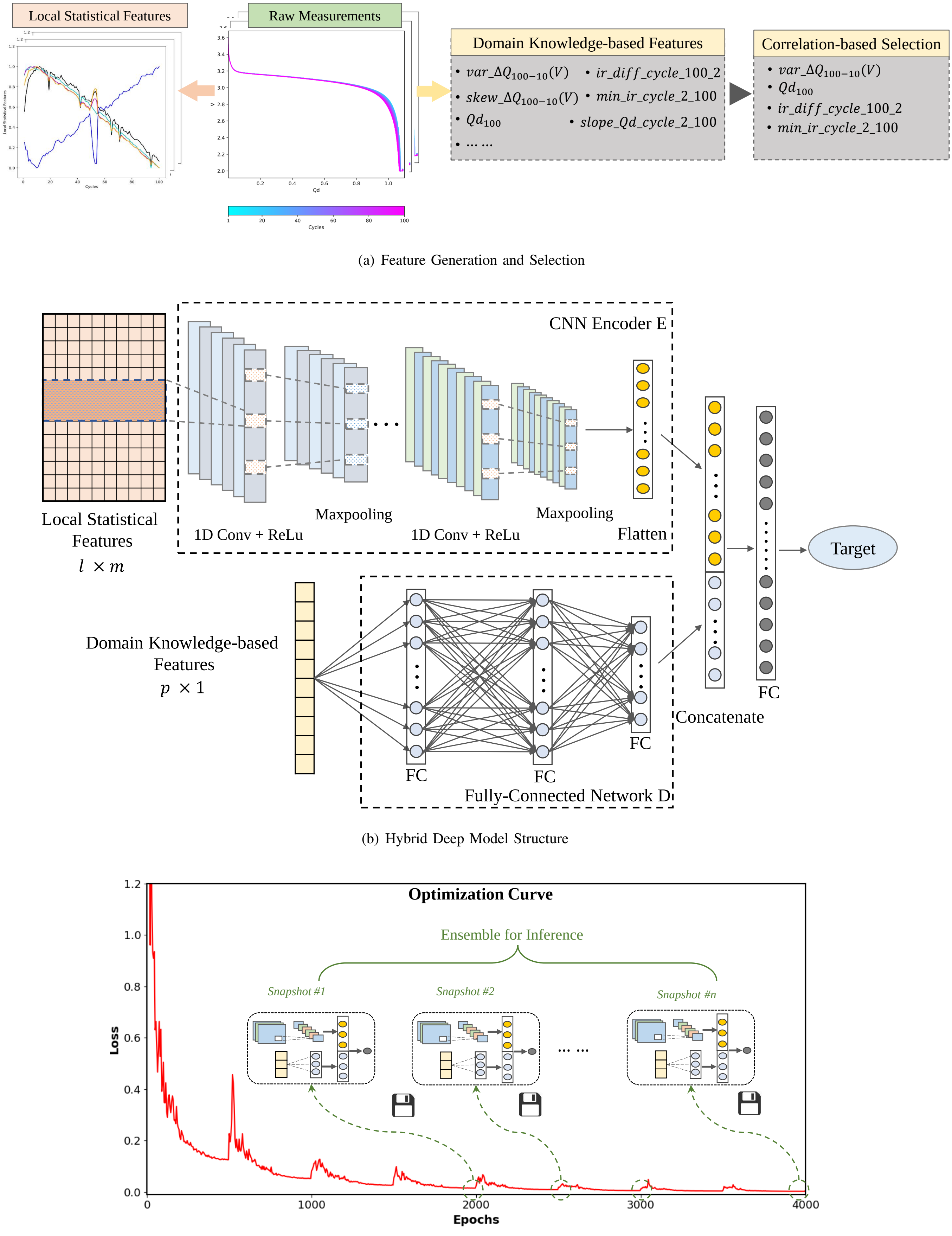
It is worth noting that our work is related to [5] but significantly differentiates from it. Although authors in [5] proposed various features based on their domain knowledge, the feature selection process was not explicitly stated. Their performance varied largely with different feature subsets. Moreover, the shallow elastic net they employed tends to only perform well on the primary test set but generalize poorly on the secondary test set, which has a different distribution from the training set. It reveals the shortcoming of using handcrafted features only for battery RUL prediction. On the contrary, we first design a non-linear correlation feature selection method together with the recursive feature elimination (RFE) technique to generate an appropriate domain knowledge feature subset. Secondly, a hybrid model, which can integrate local statistical features with domain knowledge-based features, is proposed to mitigate the data distribution discrepancy over training and test datasets and thus improves prediction accuracy. Moreover, the snapshot ensemble training strategy is developed to further enhance the generalization performance of our model.

III. Methodology

A. Overview

For a system like LIBs, the degradation process may span over thousands of cycles and for each cycle there are many variables collected by different sensors. In general, it is impractical to directly feed those tremendous raw data into a deep learning model. The noise and redundant information among those raw measurements can result in slow model convergence. It is even more challenging when only utilizing data from a few cycles to conduct an early prediction due to LIBs' non-linear degradation characteristics as aforementioned. In addition, the inconsistent distribution between training and test data further requires a good generalization ability for RUL prediction.

To cope with the above challenges, we propose to utilize features extracted based on cycle-level statistical characteristics (named as *Local Statistical Features*) and domain knowledge (named as *Domain Knowledge-based Features*) as shown in Fig. 2(a). Then, a hybrid model is developed to leverage both local statistical information and domain knowledge for battery RUL prediction as shown in Fig. 2(b). Moreover, a training strategy called snapshot ensemble is designed to further improve the model generalization ability as illustrated in Fig. 2(c).


Local Statistical Features
Raw Measurements
Domain Knowledge-based Features
• $var\_\Delta Q_{100-10}(V)$
• $ir\_diff\_cycle\_100\_2$
• $skew\_\Delta Q_{100-10}(V)$
• $min\_ir\_cycle\_2\_100$
• Qd_{100}
• $slope\_Qd\_cycle\_2\_100$
• … …
Correlation-based Selection
• $var\_\Delta Q_{100-10}(V)$
• Qd_{100}
• $ir\_diff\_cycle\_100\_2$
• $min\_ir\_cycle\_2\_100$
Cycles
Qd
V
Local Statistical Features
CNN Encoder E
1D Conv + ReLu
Maxpooling
1D Conv + ReLu
Maxpooling
Flatten
Local Statistical Features $l \times m$
Domain Knowledge-based Features $p \times 1$
FC
FC
FC
Fully-Connected Network D
Concatenate
FC
Target
Optimization Curve
Ensemble for Inference
Snapshot #1
Snapshot #2
Snapshot #n
… …
Loss
Epochs
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1000
2000
3000
4000

(a) Feature Generation and Selection

(b) Hybrid Deep Model Structure

(c) Snapshot Ensemble

Fig. 2: (a) Local Statistical Features are extracted at individual cycle level and Domain Knowledge-based Features are extracted at entire cycle level. A non-linear correlation-based method is leveraged to select features; (b) The hybrid model takes local statistical features and selected domain knowledge features as input; (c) Snapshot Ensemble Learning Strategy.

B. Feature Generation and Selection

The pipeline of feature generation and selection is crucial to the prediction performance of our proposed approach. The details are given as follows.

1) Local Statistical Features: Directly feeding cycle-level battery measurements to a deep neural network is impractical as the number of sampled data per cycle can be tremendous. Generally for time series analysis, the high dimensional raw measurements are often aggregated at the cycle level by applying different statistical metrics. For instance, for variable $Q(V)$ which is the discharge capacity as a function of voltage, we summarize the minimum value at each cycle and then generate a new feature vector. Similarly, other statistical metrics like maximum value, mean value, variance and skewness are also applied to the $Q(V)$ curve per cycle. The aggregation of those local statistical features represents the cycle-to-cycle degradation trend of battery cells from different perspectives. Moreover, it requires less prior knowledge compared with domain knowledge-based features.

2) Domain Knowledge-based Features: There is no doubt that model performance can be significantly improved by integrating features based on domain knowledge. In [5], the authors have shown that battery RUL is highly correlated to the variance of $\Delta Q_{100-10}(V)$, with a correlation coefficient of 0.93. Here, $\Delta Q_{100-10}(V)$ represents the difference of the discharge capacity curves as a function of voltage between 100^{th} and 10^{th} cycles. Only using this variable, they have achieved a promising accuracy on battery RUL prediction. Based on their domain knowledge, they proposed a set of different domain specific features summarized from battery cycle-to-cycle raw measurements.

However, a procedure to properly select the key features from those domain knowledge-based features is missing in [5]. On one hand, it requires researchers' prior knowledge to carefully generate a subset from them as involving irrelevant information or inadvertently removing important features can lead to performance degradation. On the other hand, the selected features may even contrarily reduce model performance in situations where training and test data have different distributions.

To mitigate the above issues, we propose a non-linear correlation-based feature selection method. Considering the non-linear degradation characteristics at LIBs' early stage, we adopt the Spearman's correlation coefficient to measure monotonic relationships instead of the commonly used linear correlation coefficient like Pearson. The Spearman's coefficient is calculated as Equation (1). Here, $R(X_i)$ and $R(Y_i)$ are the ranks of each data point in vector X and Y, respectively, and n is the number of samples.

$$\rho(X, Y) = 1 - \frac{6\sum_{i=1}^{n}(R(X_i) - R(Y_i))}{n(n^2-1)}. \tag{1}$$

To be specific, we first calculate the correlation between training features f_{train} and target y, denoted as $\rho(f_{train}, y)$ and generate a primary feature subset A with correlation $\rho(f_{train}, y)$ above a pre-defined threshold. Then, among features in subset A, we further recursively eliminate the features via an extreme gradient boosting regressor (XGBRegressor) and obtain a sub-subset B. In section IV, we will demonstrate how the features selected by this method contribute to the model performance.

C. Hybrid Model

In order to mine as much useful information as possible from raw measurements, we propose a hybrid model to combine both the domain knowledge-based features and the local statistical features as inputs as shown in Fig. 2(b). Particularly, the one-dimensional (1D) CNN architecture is adopted to capture the temporal dependency among the time series data. Compared to recurrent neural networks, 1D CNNs are more computational efficient and have stable back propagation characteristics. We employ the 1D CNN Encoder E to encode those local statistical features to feature vectors as shown in Equation (2). Here, X_{local} represents the local statistical features whose dimension is $n \times l \times m$. n is the number of samples, l is the total number of statistical characteristics and m is the cycle number.

Meanwhile, the selected domain knowledge-based features are fed into a fully-connected network D consisting of a series of fully connected layers as shown in Equation (3). X_{domain} represents the selected domain-specific features from the same cycle period as the local statistical features and its dimension is $n \times p$, where n is the number of samples and p is the total number of selected domain knowledge-based features. Lastly, the high-dimensional feature maps H_l and H_d are flattened and concatenated together (denoted as $H_l \| H_d$ in Equation (4)). To prevent overfitting, two dropout layers are also integrated just after CNN encoder E and fully-connected network D, before the concatenation operation. Then, a fully-connected layer is used to map the concatenated features to final target $\hat{y}$. The hybrid model is optimized by minimizing the loss between $\hat{y}$ and ground truth label y.

$$H_l = E(X_{local}), \tag{2}$$

$$H_d = D(X_{domain}), \tag{3}$$

$$\hat{y} = (H_l \parallel H_d) \times W + B. \tag{4}$$

D. Improving Model Generalization Ability

The ensemble method is a commonly used approach for better generalization performance as it can achieve consensus among models trained with different initialization and regularization configurations [32]. Among various ensemble methods, snapshot ensemble [33] has the advantage of being able to learn an ensemble of multiple neural networks without any additional training cost. As illustrated in Fig. 2(c), instead of independently training multiple models from scratch, the snapshot ensemble method saves several intermediate models along one optimization path. To prevent the saved models to be similar, a cosine annealing learning rate schedule is adopted. To be specific, during the training period of one snapshot, a large learning rate is first applied to Stochastic Gradient Descent (SGD) optimizer in order to let the model escape from the current local minimum. Then, the learning rate gradually decreases to a predefined value, which allows the model converge to some other local optimal point. We adopt the cyclical cosine annealing schedule with warm restart technique [34] for the learning rate adjustment. Fig. 3 illustrates the learning rate schedule over the whole training progress and Equation (5) gives the details of how to calculate the learning rate at a specific training epoch. Here, t is the training epoch, $lr(t)$ is the learning rate at epoch t, $lr_{\max}$ and $lr_{\min}$ are the maximum and minimum learning rates, T is the total number of training epochs for one snapshot and mod is the modulo operation. Although T could start with a small value and increase by a factor suggested in [34], we empirically found that a fixed T could yield a better result. Via the aforementioned snapshot strategy, we can obtain multiple models converging to different local minima, which can significantly improve model generalization ability over ensembling.

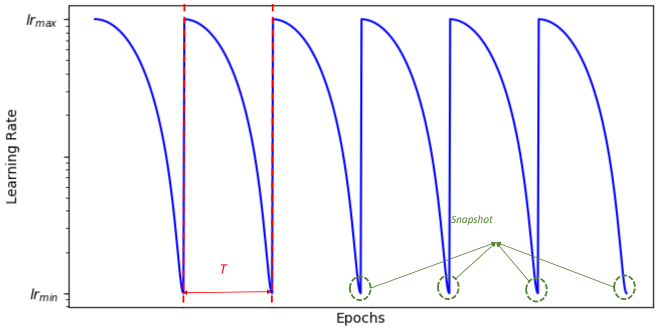


Fig. 3: Snapshot Ensemble with Cosine Annealing Learning Rate Schedule

$$lr(t) = lr_{\min} + \frac{1}{2}(lr_{\max} - lr_{\min})(1 + \cos(\frac{\mathrm{mod}(t, T)}{T}\pi)). \quad (5)$$

It is worth noting that the optimization curve in Fig. 2(c) shows that the first several intermediate models tend to perform worse as it needs some training epochs for the model converging to some local minimum points. Therefore, we only take the last several models out of all saved models for final ensemble prediction.

IV. Experiments

In this section, we evaluate the performance of our proposed hybrid model with a public dataset for the early prediction task of battery remaining useful life.

A. Dataset and Experimental Setup

1) A123 Battery Dataset: The “A123 dataset” used in this paper is generated by Severson *et al.* [5] and consists of 124 commercial lithium-ion battery cells in total. The cells were cycled under various fast-charging policies but the same discharging rate until reaching EOL, which is defined as 80% of the nominal capacity. Cycle measurements of voltage, current, charge capacity, discharge capacity, temperature, internal resistance and charge time were recorded. There are three batches of batteries in this dataset. Following [5], for the first two batches, we alternatively select one for training and the following one for primary testing. And all batteries in third batch are selected for secondary testing. Thus, the whole dataset was separated into three subsets: training, primary test and secondary test sets. Note that battery cells in the training and primary test sets were collected at the same time period but cells in the secondary test set were collected about one year later. Due to calendar aging, bias was introduced to the secondary test set. Fig. 4 compares the battery cell distributions in terms of cycle life and initial discharge capacity among these three subsets. It is clear that the training and primary test sets have similar distributions, while the secondary test set has a different distribution from the training and primary test sets. Our target is to design a hybrid model that can not only perform well in the primary test set but also generalize well to the secondary test set. (Check https://data.matr.io/1 for more details about this A123 battery dataset.)

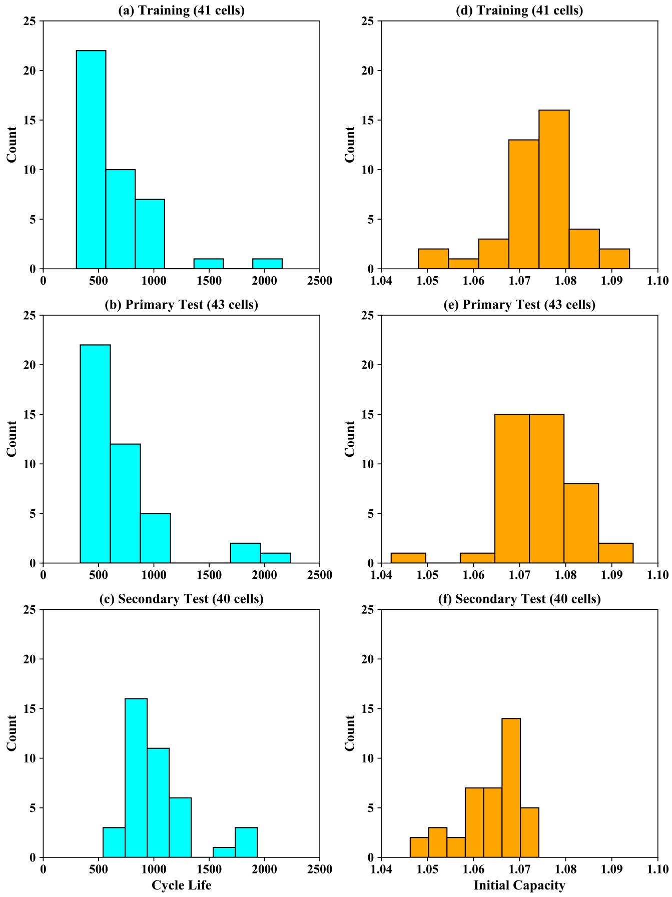


Fig. 4: (a)-(c) Battery cell distribution in terms of cycle life for training, primary test and secondary test sets; (d)-(f) Battery cell distribution in terms of initial capacity for training, primary test and secondary test sets.

2) Selecting Domain Knowledge-based Features: Severson *et al.* have done an explicit research on the relation between RUL and domain specific features in [5] and proposed 21 domain knowledge-based features for the A123 battery dataset. However, as the code for the calculation of these features used in [5] has not been released, we have to implement these

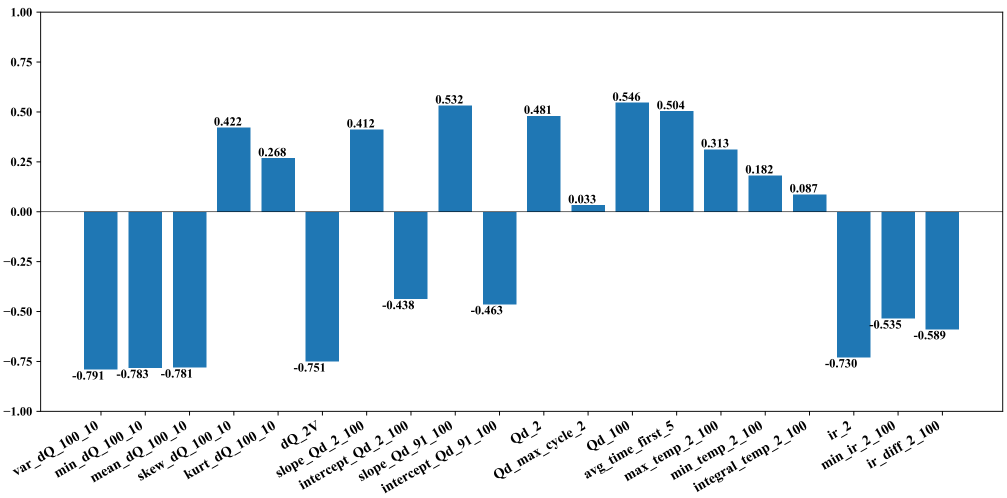


Fig. 5: Correlation between domain knowledge-based features and the logarithm value of battery RUL.

domain knowledge-based features according to the paper as shown below.

- $var\_dQ\_100\_10$, $min\_dQ\_100\_10$, $mean\_dQ\_100\_10$, $skew\_dQ\_100\_10$, $kurt\_dQ\_100\_10$: the variance, minimum, mean, skewness and kurtosis values of difference of discharge capacity curves $\Delta Q_{100-10}(V)$. Here, $\Delta Q_{100-10}(V)$ represents the difference among discharge capacity curves as a function of voltage between the 100^{th} and 10^{th} cycles, which equals $Q_{100}(V) - Q_{10}(V)$;
- dQ_2V: value at 2V;
- $slope\_Qd\_cycle\_2\_100$, $intercept\_Qd\_cycle\_2\_100$, $slope\_Qd\_cycle\_91\_100$, $intercept\_Qd\_cycle\_91\_100$: slope and intercept values of the linear fit to the capacity fade curve from cycle 2 to 100 and cycle 91 to 100.
- Qd_2, Qd_100: discharge capacity at cycle 2 and 100;
- $Qd\_max\_cycle\_2$: difference between maximum discharge capacity and discharge capacity at cycle 2;
- $avg\_time\_first\_5$: the average charge time of first 5 cycles;
- $max\_temp\_cycle\_2\_100$, $min\_temp\_cycle\_2\_100$, $integral\_temp\_cycle\_2\_100$: the maximum and minimum temperatures from cycle 2 to 100, the integral of temperature over time from cycle 2 to 100;
- ir_2, $min\_ir\_cycle\_2\_100$, $ir\_diff\_cycle\_100\_2$: internal resistance at cycle 2, minimum internal resistance from cycle 2 to 100, internal resistance difference between cycle 100 and 2.

The Spearman's coefficients $\rho(f_{train}, y)$ between training features and targets are presented in Fig. 5. We generate a subset A including 12 features whose $|\rho| > 0.45$. Within the selected features in A, we perform recursive feature elimination via an XGBRegressor estimator on the training set to further remove the noisy features. The final subset with 8 features we use in our model is $\{var\_dQ\_100\_10$, $min\_dQ\_100\_10$, $mean\_dQ\_100\_10$, $slope\_Qd\_cycle\_91\_100$, Qd_2, Qd_100, $min\_ir\_cycle\_2\_100$, $ir\_diff\_cycle\_100\_2\}$.

3) Experimental Setup: Same as [5], we only use the first 100 cycle measurements for the early prediction of battery cell RUL. To accelerate the learning process of the proposed hybrid model, we apply min-max normalization to both local statistical features and domain knowledge-based features on the training set. The normalization scalers are then applied to the primary and secondary test sets as well. The logarithm value of cycle life is the prediction target of the hybrid model. The model is trained on the training set and then evaluated on the primary and secondary test sets.

We use k-fold cross-validation with grid search to tune the hyper-parameters of the proposed hybrid model. Particularly, the training set is randomly split into 5 folds, where 4 folds are utilized for training and the remaining one for validation. We use the validation performance to select the best hyper-parameters. After selection, we first identify the network configuration for our proposed hybrid model as shown in Table I. Here, "Conv1D(5,1)" represents 1-D convolutional operation with a kernel size $= 5$ and stride $= 1$. "FC-16" means a fully connected layer with the output dimension $= 16$. In addition, the dropout rates of the two dropout layers after the CNN encoder E and deep network D are set to be 0.5 and 0.1, respectively. For the settings of snapshot ensemble, we finally set $lr_{max} = 0.1$, $lr_{min} = 1e-6$ and $T = 200$ in Equation (5). Hence, the learning rate starts at 0.1 and decays with a cosine annealing until reaching $1e-6$ within 200 epochs, then we save the intermediate model and reset the learning rate to 0.1 and repeat it again. We take the last 10 saved models for ensemble learning and evaluate them on test sets. Please refer to the released code for more implementation details.

Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) are chosen to evaluate the model performance. They are defined in Equations (6) and (7), respectively. Here, n is the total number of samples, $\hat{y}_i$ is the predicted cycle life and y_i is the ground truth.

TABLE I: Network Configuration of Proposed Hybrid Model

Layers	Local Statistical Features	Domain Knowledge-based Features
#1	Conv1D(5,1); ReLU; MaxPool1d	FC-16; Sigmoid
#2	Conv1D(5,1); ReLU; MaxPool1d	
#3	Conv1D(5,1); ReLU; MaxPool1d	FC-16; Dropout
#4	Flatten; FC-16; Dropout	
#5	Concatenate	
#6	FC-1	

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{6}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i} \times 100\% \tag{7}$$

B. Performance Comparison

To evaluate the effectiveness of our proposed hybrid model, we compare it with various methods as shown in Table II. As aforementioned, we implemented the domain feature extraction part according to the original paper [5] and trained an elastic net with them in the same manner as the original paper. The models **Elastic-V**, **Elastic-D** and **Elastic-F** represent the elastic net trained with “Variance”, “Discharge” and “Full” feature sets, respectively. Particularly, the “Variance” feature set contains var_dQ_100_10 feature only; the “Discharge” feature set contains more features from discharge capacity fade curves, namely { var_dQ_100_10, min_dQ_100_10, $skew_dQ$_100_10, $kurt_dQ$_100_10, Qd_2, Qd_max_cycle_2 }; the “Full” feature set consists of all the 9 features, namely {var_dQ_100_10, min_dQ_100_10, $slope_Qd_cycle$_2_100, $intercept_Qd_cycle$_2_100, Qd_2, avg_time_first_5, $integral_temp_cycle$_2_100, min_ir_cycle_2_100, ir_diff_cycle_100_2}. We take these models as the baselines in our experiments.

Other machine learning based methods like SVR [35], NN [36], LSTM [22] and CNN-LSTM [7] are also evaluated with the same dataset. We tried different combinations of models and feature sets, and only listed the models that perform best in Table II. In particular, SVR-V and NN-F mean that the models are trained with the “Variance” and “Full” feature set, respectively. LSTM and CNN-LSTM models are trained with local statistical features. Moreover, we also investigate two more methods, namely GRU-RNN [31] and PSR-SVR [27]. For GRU-RNN, we generate the exact same features from voltage, current and temperature measurements as [31], and then apply the random forest method to select 15 best features. We adopt their GRU-RNN model to evaluate the performance on RUL prediction. For PSR-SVR, we extract the discharge capacity and discharging voltage difference of equal time intervals from the first 100 cycles. The EEMD is then employed to reconstruct the signal and the same PSR process is utilized to obtain an optimal input sequence. Particularly, we set embedding dimension $m = 6$ and delay time $\tau = 3$ which are the same as [27]. Through explicit experiments, we intend to explore how to properly design a model and generate features so as to not only achieve a good performance but also generalize well.

TABLE II: Performance Comparison among Various Methods

Methods	Primary Test		Secondary Test	
	RMSE	MAPE	RMSE	MAPE
Elastic-V [5]	138.39	13.19	196.01	11.41
Elastic-D [5]	170.35	10.99	179.64	14.20
Elastic-F [5]	117.64	9.20	225.72	12.85
SVR-V [35]	170.72	14.72	226.79	12.11
NN-F [36]	116.54	9.13	225.83	12.87
LSTM [22]	166.87	18.05	380.85	22.66
CNN-LSTM [7]	176.63	14.85	375.47	25.25
GRU-RNN [31]	127.65	9.94	356.31	34.17
PSR-SVR [27]	191.83	20.3	404.57	32.93
Proposed	**114.05**	**8.54**	**177.88**	**11.31**

From Table II, we can see that different feature sets generally result in different performance on the two test sets. For instance, Elastic-D performs better in the secondary test set in terms of RMSE but performs worse in the primary test set in terms of both RMSE and MAPE, when compared with Elastic-F. SVR-V leverages the same feature set as Elastic-V but performs worse than it. These observations indicate the importance and difficulty of selecting suitable domain knowledge-based features and proper models. Moreover, it also shows the poor generalization ability of simple models like the elastic, which fails to achieve good performance on the secondary test dataset. It is not surprising that deep neural networks like LSTM and CNN-LSTM also perform poorly in both test sets as they are originally designed for the capacity degradation curve, but such degradation is too small to be captured in the early stage of the whole battery cycle life. The results of GRU-RNN and PSR-SVR reveal the difficulty of generating representative features from classical characteristics like discharging voltage for early prediction of battery RUL. Our proposed hybrid model outperforms others in both primary and secondary test sets in terms of RMSE and MAPE, indicating the effectiveness of combining local statistical and domain knowledge-based features to improve model generalization ability.

Moreover, we have selected some representative methods with different network architectures (i.e., Elastic-F, SVR-V, CNN-LSTM and GRU-RNN) and our proposed method

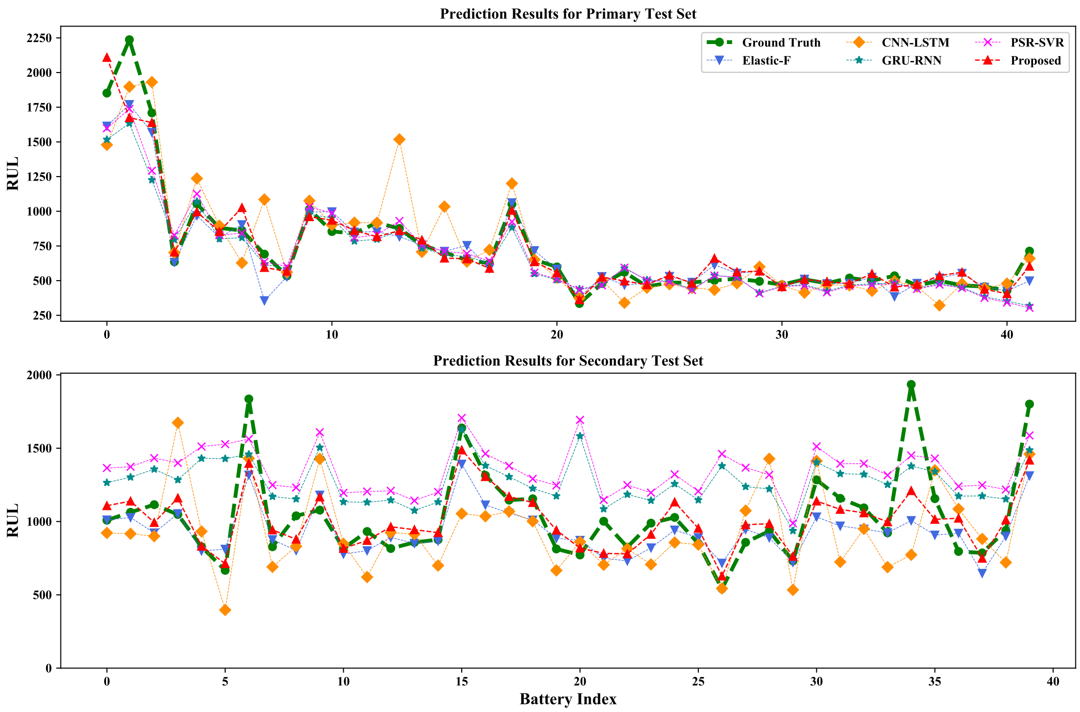


Fig. 6: Visualization of Prediction Results for Different Methods.

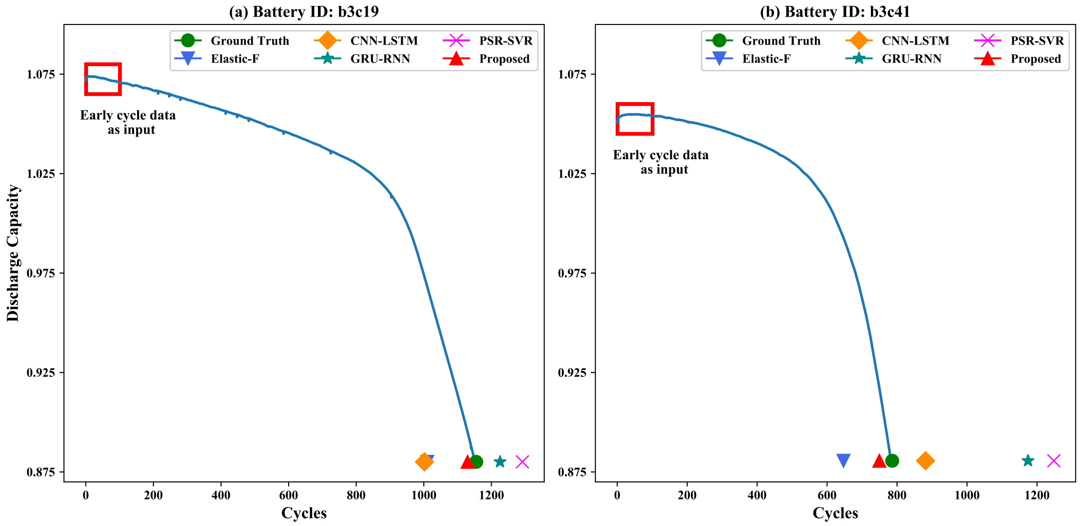


Fig. 7: Visualization of Prediction Results for Single Battery.

to demonstrate the final RUL prediction results of testing samples. Fig. 6 visualizes the final RUL prediction results of these methods for primary and secondary test sets. As aforementioned, most of methods can achieve relatively good performance on primary test set but generalize worse on secondary test set. Fig. 7 depicts the early prediction results of two randomly selected batteries. By only utilizing the early cycle data (the first 100 cycles) before the degradation of battery capacity, our proposed hybrid model could achieve more accurate RUL prediction results than other methods.

C. Ablation Study

To investigate the contributions of local statistical features and domain knowledge-based features to model performance, we conduct an ablation study as shown in Fig. 8. Here, CNN model is trained with local statistical features only, and NN-C is a network consisting of two fully connected layers and trained with domain knowledge features selected based on correlations. It is obvious that features based on domain knowledge contribute more to the accuracy and only using local statistical features would result in very poor performance as there are redundant information and noise underlying them.

TABLE III: Performance Comparison between model averaging and snapshot ensemble.

Methods	Primary Test		Secondary Test	
	RMSE	MAPE	RMSE	MAPE
Model Averaging	120.56±10.04	8.71±0.24	184.09±6.47	12.91±1.42
Weighted Averaging	117.99±4.24	8.62±0.16	192.45±3.54	12.04±0.82
Snapshot Ensembles	114.05±1.31	8.54±0.08	177.88±1.36	11.31±0.33

However, combining local statistical information with domain knowledge indeed helps to enhance model performance on both test sets.

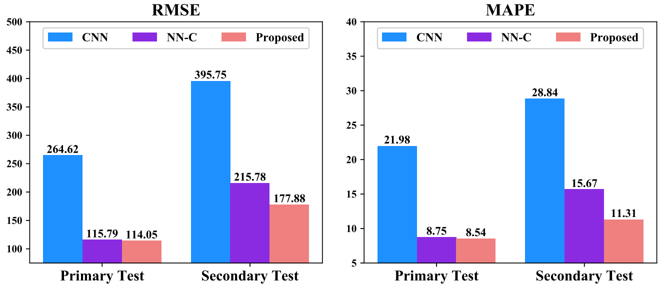


Fig. 8: Model performance with different features.

We also compare different ensemble learning methods on prediction performance. A common ensemble method to improve model generalization ability is to train several models initialized with different random seeds and take the average value of these model predictions as the final result. It is denoted as Model Averaging in Table III. In addition, we also investigate the weighted average ensemble by assigning different weights to each individual model based on their performance on the validation set. The lower RMSE the model achieves, the higher weight it is assigned with. This method is denoted as Weighted Averaging. Note that both Model Averaging and Weighted Averaging methods require training of multiple models for ensemble learning, while our proposed snapshot ensemble only requires the model to be trained once. The mean and standard deviation values are calculated over 5 iterations for all the ensemble learning methods. From Table III, we can find that the proposed snapshot ensemble outperforms both methods with smaller RMSE and lower variance, which indicates the effectiveness and robustness of the proposed ensemble learning.

Here, we also test different correlation methods for feature selection. In addition to the non-linear Spearman's correlation coefficient for measuring the monotonic relationship between features and RUL, we also explore another linear correlation method (*i.e.,* Pearson correlation coefficient). We compare the two correlation methods with/without RFE to evaluate the impact of different correlation methods and the usefulness of RFE. The results are shown in Table IV. We can find that the feature subset selected by non-linear Spearman's correlation consistently performs better than the subset selected by linear correlation with and without RFE. In addition, the RFE can further improve the performance in both scenarios.

TABLE IV: Performance Comparison of Different Correlation Methods.

Features selected by	Primary Test		Secondary Test	
	RMSE	MAPE	RMSE	MAPE
Pearson	144.72	11.38	222.46	17.10
Pearson + RFE	129.45	10.41	220.43	14.48
Spearman	116.45	8.55	206.88	11.67
Spearman + RFE	114.05	8.54	177.88	11.31

V. Conclusion

In this paper, we propose a hybrid deep learning model that can integrate the handcrafted features with domain knowledge and the latent features derived by deep networks to boost the performance and generalizability of battery RUL prediction. Moreover, we also explore different correlation methods for feature selection and different ensemble strategies that would affect model generalization ability. An exhaustive comparison with other SOTA approaches and ablation study have been conducted in the paper. The experimental results show that our proposed hybrid model outperforms SOTA approaches in terms of two evaluation criteria in the primary test set, and also has better generalization ability to the secondary test set. In addition, the proposed non-linear correlation-based feature selection and snapshot ensemble strategy can clearly contribute to model prediction accuracy and generalization ability.

In the future, we intend to integrate physics-based models with deep learning models as the physics-based models are well-known for its physical interpretability and outstanding capability on accurately modeling LIBs' degradation processes [37], [38]. The combination of these two types of models would result in novel hybrid models that are more physically consistent, explainable and accurate. Besides, we also intend to investigate the minimal required cycles for accurate battery RUL prediction in our future work, since it can significantly reduce the experimental cost for battery manufactures in practical applications.

References

[1] Y. Ma, X. Zhou, B. Li, and H. Chen, "Fractional modeling and soc estimation of lithium-ion battery," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 3, pp. 281–287, 2016.

[2] K. Liu, Z. Wei, C. Zhang, Y. Shang, R. Teodorescu, and Q.-L. Han, "Towards long lifetime battery: Ai-based manufacturing and management," *IEEE/CAA Journal of Automatica Sinica*, 2022.

[3] Y. Ma, B. Li, G. Li, J. Zhang, and H. Chen, "A nonlinear observer approach of soc estimation based on hysteresis model for lithium-ion battery," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 2, pp. 195–204, 2017.

[4] T. Meng, Z. Lin, and Y. A. Shamash, "Distributed cooperative control of battery energy storage systems in dc microgrids," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 3, pp. 606–616, 2021.

[5] K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis *et al.*, "Data-driven prediction of battery cycle life before capacity degradation," *Nature Energy*, vol. 4, no. 5, pp. 383–391, 2019.

[6] P. M. Attia, A. Bills, F. B. Planella, P. Dechent, G. d. Reis, M. Dubarry, P. Gasper, R. Gilchrist, S. Greenbank, D. Howey, O. Liu, E. Khoo, Y. Preger, A. Soni, S. Sripad, A. G. Stefanopoulou, and V. Sulzer, ""Knees" in lithium-ion battery aging trajectories," *arXiv:2201.02891 [cond-mat, physics:physics]*, Jan. 2022, arXiv: 2201.02891. [Online]. Available: http://arxiv.org/abs/2201.02891

[7] G. Ma, Y. Zhang, C. Cheng, B. Zhou, P. Hu, and Y. Yuan, "Remaining useful life prediction of lithium-ion batteries based on false nearest neighbors and a hybrid neural network," *Applied Energy*, vol. 253, p. 113626, 2019.

[8] M. Guo, G. Sikha, and R. E. White, "Single-particle model for a lithium-ion cell: Thermal behavior," *Journal of The Electrochemical Society*, vol. 158, no. 2, p. A122, 2011.

[9] S. Santhanagopalan, Q. Guo, P. Ramadass, and R. E. White, "Review of models for predicting the cycling performance of lithium ion batteries," *Journal of Power Sources*, vol. 156, no. 2, pp. 620–628, 2006.

[10] M. Doyle, T. F. Fuller, and J. Newman, "Modeling of Galvanostatic Charge and Discharge of the Lithium/Polymer/Insertion Cell," *Journal of The Electrochemical Society*, vol. 140, no. 6, pp. 1526–1533, Jun. 1993. [Online]. Available: http://jes.ecsdl.org/content/140/6/1526

[11] T. F. Fuller, M. Doyle, and J. Newman, "Simulation and Optimization of the Dual Lithium Ion Insertion Cell," *Journal of The Electrochemical Society*, vol. 141, no. 1, pp. 1–10, Jan. 1994. [Online]. Available: http://jes.ecsdl.org/content/141/1/1

[12] A. Jokar, B. Rajabloo, M. Désilets, and M. Lacroix, "Review of simplified pseudo-two-dimensional models of lithium-ion batteries," *Journal of Power Sources*, vol. 327, pp. 44–55, 2016.

[13] P. Kemper, S. E. Li, and D. Kum, "Simplification of pseudo two dimensional battery model using dynamic profile of lithium concentration," *Journal of Power Sources*, vol. 286, pp. 510–525, 2015.

[14] K. Liu, Y. Gao, C. Zhu, K. Li, M. Fei, C. Peng, X. Zhang, and Q.-L. Han, "Electrochemical modeling and parameterization towards control-oriented management of lithium-ion batteries," *Control Engineering Practice*, vol. 124, p. 105176, 2022.

[15] R. B. Smith and M. Z. Bazant, "Multiphase porous electrode theory," *Journal of The Electrochemical Society*, vol. 164, no. 11, p. E3291, 2017.

[16] R. B. Smith, E. Khoo, and M. Z. Bazant, "Intercalation kinetics in multiphase-layered materials," *The Journal of Physical Chemistry C*, vol. 121, no. 23, pp. 12 505–12 523, 2017.

[17] J. M. Reniers, G. Mulder, and D. A. Howey, "Review and performance comparison of mechanical-chemical degradation models for lithium-ion batteries," *Journal of The Electrochemical Society*, vol. 166, no. 14, pp. A3189–A3200, 2019.

[18] K. Liu, Y. Shang, Q. Ouyang, and W. D. Widanage, "A data-driven approach with uncertainty quantification for predicting future capacities and remaining useful life of lithium-ion battery," *IEEE Transactions on Industrial Electronics*, 2020.

[19] M. A. Patil, P. Tagade, K. S. Hariharan, S. M. Kolake, T. Song, T. Yeo, and S. Doo, "A novel multistage support vector machine based approach for li ion battery remaining useful life estimation," *Applied Energy*, vol. 159, pp. 285–297, 2015.

[20] Y. Zhang, R. Xiong, H. He, and M. G. Pecht, "Lithium-ion battery remaining useful life prediction with box–cox transformation and monte carlo simulation," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1585–1597, 2018.

[21] J. Wu, C. Zhang, and Z. Chen, "An online method for lithium-ion battery remaining useful life estimation using importance sampling and neural networks," *Applied Energy*, vol. 173, pp. 134–140, 2016.

[22] Y. Zhang, R. Xiong, H. He, and M. G. Pecht, "Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 5695–5705, 2018.

[23] L. Zhang, Z. Mu, and C. Sun, "Remaining useful life prediction for lithium-ion batteries based on exponential model and particle filter," *IEEE Access*, vol. 6, pp. 17 729–17 740, 2018.

[24] Y. Song, D. Liu, C. Yang, and Y. Peng, "Data-driven hybrid remaining useful life estimation approach for spacecraft lithium-ion battery," *Microelectronics Reliability*, vol. 75, pp. 142–153, 2017.

[25] Z. Pang, X. Si, C. Hu, and Z. Zhang, "An age-dependent and state-dependent adaptive prognostic approach for hidden nonlinear degrading system," *IEEE/CAA Journal of Automatica Sinica*, 2021.

[26] Y. Chang, H. Fang, and Y. Zhang, "A new hybrid method for the prediction of the remaining useful life of a lithium-ion battery," *Applied energy*, vol. 206, pp. 1564–1578, 2017.

[27] L. Chen, Y. Zhang, Y. Zheng, X. Li, and X. Zheng, "Remaining useful life prediction of lithium-ion battery with optimal input sequence selection and error compensation," *Neurocomputing*, vol. 414, pp. 245–254, 2020.

[28] L. Ren, L. Zhao, S. Hong, S. Zhao, H. Wang, and L. Zhang, "Remaining useful life prediction for lithium-ion battery: A deep learning approach," *IEEE Access*, vol. 6, pp. 50 587–50 598, 2018.

[29] X. Li, L. Zhang, Z. Wang, and P. Dong, "Remaining useful life prediction for lithium-ion batteries based on a hybrid model combining the long short-term memory and elman neural networks," *Journal of Energy Storage*, vol. 21, pp. 510–518, 2019.

[30] L. Ren, J. Dong, X. Wang, Z. Meng, L. Zhao, and M. J. Deen, "A data-driven auto-cnn-lstm prediction model for lithium-ion battery remaining useful life," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 3478–3487, 2021.

[31] R. Rouhi Ardeshiri and C. Ma, "Multivariate gated recurrent unit for battery remaining useful life prediction: A deep learning approach," *International Journal of Energy Research*, 2021.

[32] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[33] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get m for free," *arXiv preprint arXiv:1704.00109*, 2017.

[34] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[35] C. Weng, Y. Cui, J. Sun, and H. Peng, "On-board state of health monitoring of lithium-ion batteries using incremental capacity analysis with support vector regression," *Journal of Power Sources*, vol. 235, pp. 36–44, 2013.

[36] S. Zhang, B. Zhai, X. Guo, K. Wang, N. Peng, and X. Zhang, "Synchronous estimation of state of health and remaining useful lifetime for lithium-ion battery using the incremental capacity and artificial neural networks," *Journal of Energy Storage*, vol. 26, p. 100951, 2019.

[37] M. Aykol, C. B. Gopal, A. Anapolsky, P. K. Herring, B. v. Vlijmen, M. D. Berliner, M. Z. Bazant, R. D. Braatz, W. C. Chueh, and B. D. Storey, "Perspective—Combining Physics and Machine Learning to Predict Battery Lifetime," *Journal of The Electrochemical Society*, vol. 168, no. 3, p. 030525, Mar. 2021, publisher: The Electrochemical Society. [Online]. Available: https://doi.org/10.1149/1945-7111/abec55

[38] J. Lin, Y. Zhang, and E. Khoo, "Hybrid physics-based and data-driven modeling with calibrated uncertainty for lithium-ion battery degradation diagnosis and prognosis," *arXiv:2110.13661 [physics]*, Oct. 2021, arXiv: 2110.13661. [Online]. Available: http://arxiv.org/abs/2110.13661