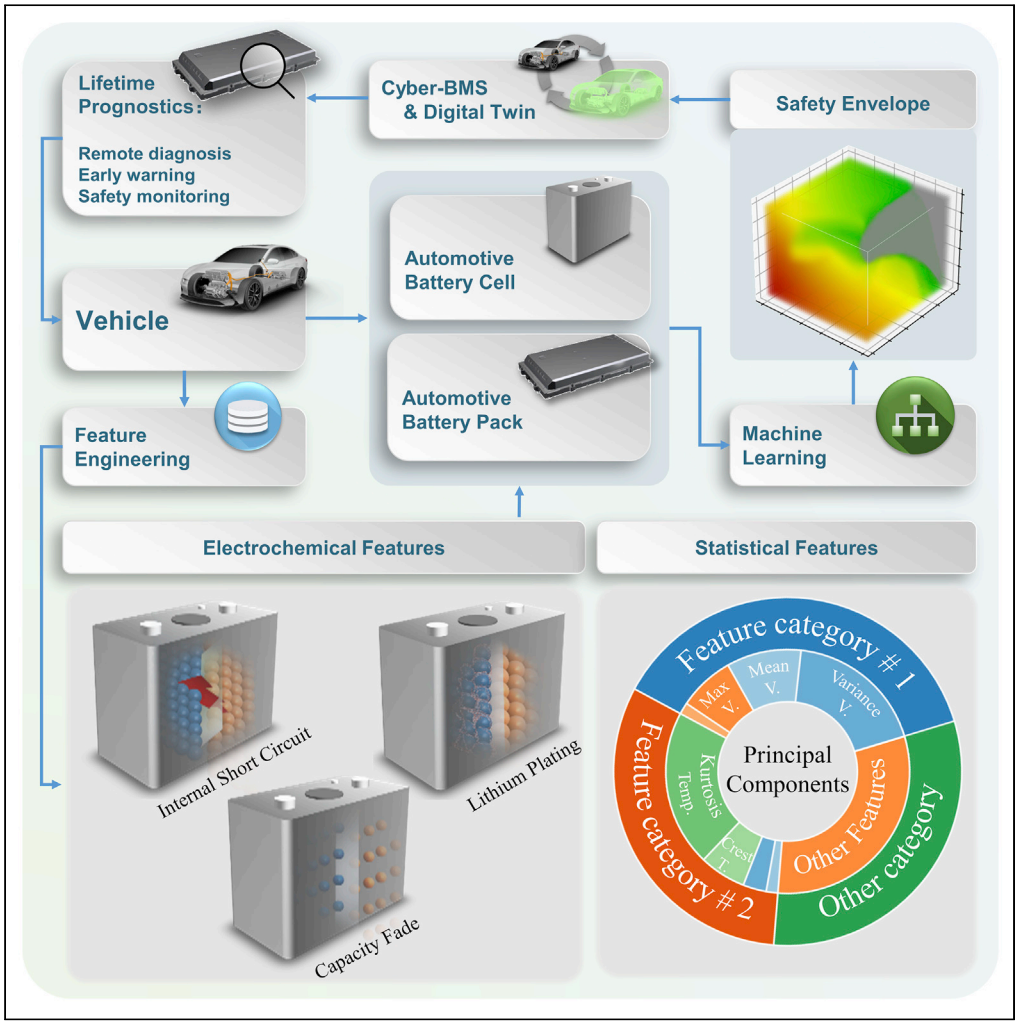
iScience

CellPress
OPEN ACCESS

Article

Data-driven prediction of battery failure for electric vehicles



Jingyuan Zhao, Heping Ling, Junbin Wang, Andrew F. Burke, Yubo Lian

jyzhao_scholar@yeah.net (J.Z.)
lian.yubo@byd.com (Y.L.)

Highlights

A well-integrated machine learning technique is applied to failure prediction

A cloud-based closed-loop framework is established for real-world EV applications

Cloud-based AI solution is based on an in-depth analysis of the field data

Both electrochemical and statistical feature engineering are established

Zhao et al., iScience 25, 104172
April 15, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.104172

iScience

CellPress
OPEN ACCESS

Article

Data-driven prediction of battery failure for electric vehicles

Jingyuan Zhao,[1,2,3,*] Heping Ling,[1] Junbin Wang,[1] Andrew F. Burke,[2] and Yubo Lian[1,*]

SUMMARY

Despite great progress in battery safety modeling, accurately predicting the evolution of multiphysics systems is extremely challenging. The question on how to ensure safety of billions of automotive batteries during their lifetime remains unanswered. In this study, we overcome the challenge by developing machine learning techniques based on the recorded data that are uploaded to the cloud. Using charging voltage and temperature curves from early cycles that are yet to exhibit symptoms of battery failure, we apply data-driven models to both predict and classify the sample data by health condition based on the observational, empirical, physical, and statistical understanding of the multiscale systems. The best well-integrated machine learning models achieve a verified classification accuracy of 96.3% (exhibiting an increase of 20.4% from initial model) and an average misclassification test error of 7.7%. Our findings highlight the need for cloud-based artificial intelligence technology tailored to robustly and accurately predict battery failure in real-world applications.

INTRODUCTION

The increase in environmental awareness and development of high-energy rechargeable batteries, as well as policy incentives, greatly stimulated the growth of electric vehicles (EVs) (Foulds and Christensen, 2016; Plötz et al., 2019).Innovation initiative to accelerate the progress on clean energy research and EV technology is currently succeeding in its quest to support carbon mitigation. However, to compare favorably with conventional vehicles using internal combustion engines, EVs should have a long driving range, a short recharging time, and an affordable cost. Thus, huge efforts have been devoted worldwide by engineers and researchers focusing on the improvement of energy storage capability of battery energy storage technology (Roberts et al., 2014; Nitta et al., 2015; Zeng et al., 2019; Gao and Lu, 2021; Li et al., 2021b; Manthiram and Goodenough, 2021; Thackeray and Amine, 2021). One price for this reasonable action are safety issues when squeezing more active materials in the cell or aggressive operation windows (Burke, 2014; Finegan and Cooper, 2019). From the serious consequences of the fire incidents (Sun et al., 2020), it can be concluded that safety issues are a top priority and they are highly related to EV industrial security, which is of great significance to create positive and sustainable transformations into a greener future. Therefore, we must address safety issues of EV batteries in real-world applications, even if the absolute probability of such an event remains very low.

Over the past decades, great efforts have been made to improve the safety of battery systems in the safety enhancement at the material and cell levels (Wu et al., 2014; Liu et al., 2018b; Deyab et al., 2021), data-driven prediction methods using physics-based models (Deng et al., 2018; Finegan et al., 2020), and investigation of failure mechanism and mitigating solutions (Feng et al., 2017; Galushkin et al., 2018; Wang et al., 2019; Bla et al., 2020; Chombo and Laoonual, 2020; Feng et al., 2020; Xiong et al., 2020; Chen et al., 2021; Li et al., 2021a). Nevertheless, safety issues related to battery failures in real-world applications are still a serious issue for EVs. Most early efforts in battery safety modeling are at the material or cell level with a focus on a single abuse condition such as mechanical or thermal failure. These efforts may not completely reflect the failure of batteries in automotive applications (Deng et al., 2018). The plausible, cutting-edge methods being considered can provide insights into the early detection of battery failure; however, it is still difficult to find a robust method to accurately predict battery failure in real-world EV applications. One reason is that although data-driven, machine learning-based methods offer opportunities for advancement in the predictive capabilities in the energy storage realm (Correa-Baena et al., 2018; Sendek et al., 2019; Aykol et al., 2020; Ng et al., 2020; Lombardo et al., 2021; Roman et al., 2021). The absence of high-quality,

[1]BYD Automotive Engineering Research Institute, Shenzhen 518118, China

[2]Institute of Transportation Studies, University of California, Davis, CA 95616, USA

[3]Lead contact

*Correspondence: jyzhao_scholar@yeah.net (J.Z.), lian.yubo@byd.com (Y.L.)

https://doi.org/10.1016/j.isci.2022.104172

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

full-scale datasets is one of the most critical challenges facing the accurate prediction of battery failure in EV use (Finegan et al., 2020). One recent study can be considered an excellent starting point in the prediction of battery safety for EVs, in which a large matrix of mechanical loading conditions in terms of geometry, mass, loading speed, and angle is simulated, and safety boundary conditions, named a safety envelope, is established using the machine learning technique (Li et al., 2019a). However, in practical terms, placing sensors in each cell to detect the loading condition will greatly increase the costs of EVs, which is undesirable for both manufacturers and consumers. In addition, mechanical abuse conditions are only one of the detonators that may trigger hazardous failure. Besides, in addition to various abuse conditions, unintentional defects or contamination occurring during manufacturing and subtle and undetected differences in synthesis conditions make the gap between experimental data and field failure under real-world conditions larger than we expected.

Establishing a high-accuracy detailed computational model of cells to cover all the abuse conditions is a good scientific method, although not feasible (Finegan and Cooper, 2019; Finegan et al., 2020). The experimental datasets that cover the complete picture of battery failure and underlying mechanisms under various conditions of failure occur very infrequently, which makes them experimentally very difficult (if not impossible) to obtain even for an effective collaboration between academia and industry. In addition battery failure is becoming more rare with ever-improving production techniques and packaging designs; this makes it much more difficult to create the test matrix covering all the types of failure mechanisms and loading patterns for collection of high-quality and robust data. Another reason why accurate prediction of battery failure in real-world application is very challenging is because of the absence of precise knowledge of field failure mechanisms, uncertainties in materials and manufacturing processes, and dynamic environmental and operation conditions. The challenge is daunting: the realistic conditions in EV applications have made the issue quite a different proposition. Degradation mechanisms and health conditions greatly depend on the operating conditions. Besides, finding methods to design and develop representative charge/discharge cycles and rest/shelf cycles from large volumes of battery data to simulate field conditions in EV use is also very time-consuming because of their complex operational lifetime. There may be a huge and unrecognized risk for just a few days or weeks of operation in real-world applications.

Even if all the aforementioned problems are solved, another, far more pressing matter must be faced is the onboard battery management strategy (BMS) including voltage balancing and thermal management (i.e., heating and cooling). There is no possible way to generate full-scale relevant data covering all the failure scenarios and mechanisms at the pack level even for a large EV manufacturer. In addition, random noise, unpredictable driving cycles, and dynamic charging protocols in real-world EV use also cover the information that is extremely valuable to reflect the real conditions of the automotive battery. Unfortunately, very few studies have considered the impact of the strategies encoded in the BMS on the predictive modeling of automotive batteries in real-world applications. These practical issues not only make huge efforts-based experimental data and physics-based methods useless, but they also make interpretation of the field failure much more difficult than can be imagined, based on the analysis of the field data in EV applications.

The question that needs to be asked is whether there exists a method that holds excellent robustness and high accuracy. The efforts in understanding multiscale physics by numerically solving partial differential equations (PDEs) provide an effective method in the modeling and predicting the evolution of nonlinear electrochemical energy storage systems. However, modeling and predicting the nonlinear dynamics of multiphysics and multiscale systems remains a daunting problem. High-dimensional multiphysics problems governed by parameterized PDEs can be a formidable task, particularly in the case of solving real-life physical problems with uncertain parameters or noisy data. Recent advances in machine learning opened up a new world of opportunities to accelerate the pace of continuous innovation in material and battery technologies. Such probabilistic machine learning offers an elegant way of integrating noisy data and statistical models and has a chance to achieve better prediction accuracy and improved generalization performance.

In this study, we proposed a creative cloud-based closed loop solution for robustly and accurately predicting battery failure, with the maturity of the technologies on cloud-computing (Drake, 2014) and internet of things (IoT) (Al-Fuqaha et al., 2015). To best represent and manipulate uncertainty about models and predictions, a well-integrated machine learning model with rapid evolution and predictive robustness is developed as a key ingredient of the closed loop system. The cloud-based framework using data-driven models

CellPress
OPEN ACCESS

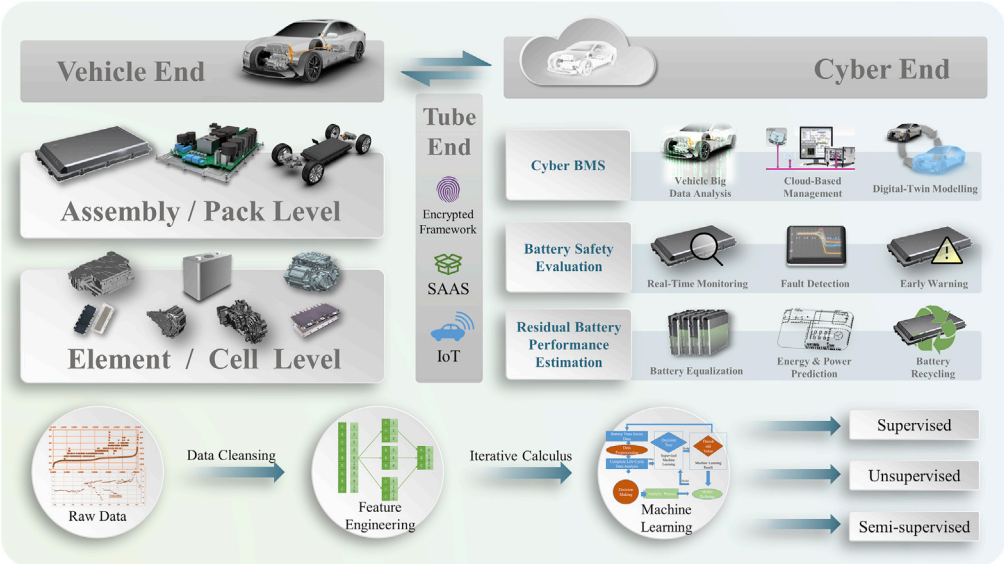


Figure 1. Schematic illustration of cyber-BMS for EVs

The cyber chain is composed of the vehicle-end, tube-end, and cyber-end. Vehicle-end represents a complex automotive system with the kinematics and dynamics characteristics, in which an enormous number of power units can be considered as a bridge and dynamic force of co-operation, creating opportunities and mechanisms for safe, effective, and smart operation of the vehicles under realistic conditions. The tube-end is a communicating channel with service-oriented pipeline architecture and security equipment, which offers opportunities to iteratively update the strategies encoded in on-spot BMS. In cyberspace, a battery digital twin is created, enabling the self-controlling system with machine learning models to work efficiently and effectively. Periodically, the longitudinal electronic health records for the battery systems are established by the cyber-BMS on the cloud platform based on the fault detection and failure prediction.

can seamlessly learn from the historical automotive battery data and generate longitudinal electronic health records in cyberspace, thereby achieving continual learning for the prediction task (Figure 1). In addition, the cloud-based framework is an efficient solution for promoting a greater understanding on battery failure in vehicle applications and accelerates the implementation of artificial intelligence and machine learning techniques for cloud-based battery management systems (i.e. Cyber-BMS). Along with the introduction to the methodology and modeling, we also discussed the most intractable problems that need to be addressed for establishing the framework and early warning mechanism in real-world applications. This work demonstrates that the combination of cloud-based data generation and machine learning modeling is a more practical approach for ensuring automotive battery safe operation under realistic conditions. The authors hope that this exnovation will create reference value to the EV and battery industries for ameliorating the hurdles for the prediction of battery failure and attract more initiatives in academia and industry to fully exploit its potential.

A CLOUD-BASED CLOSED LOOP FRAMEWORK FOR THE PREDICTION OF BATTERY FAILURE IN EV APPLICATIONS

Data generation

Over the past decade, the availability of large datasets facilitates faster acceptance of artificial intelligence in various communities. More recently, in the field of energy storage, a number of innovative technologies have been launched and are now starting to shape battery research in terms of performance evaluation, such as cycle life prediction (Severson et al., 2019), charging protocols optimization (Attia et al., 2020), and safety modeling (Deng et al., 2018; Li et al., 2019a). These new innovations open a new avenue to develop highly accurate and detailed computational models of lithium-ion batteries by using machine learning techniques based on the data collected in the laboratory. However, in the case of prediction of automotive battery failure in real-world applications, it will make experimental design covering the entire spectrum of failure conditions under various operating windows to simulate real-world environments a very challenging task (if not impossible), even for a battery industry that is used to being proficient in battery design and management. In addition, the gap between experimental environment (a small number of tests

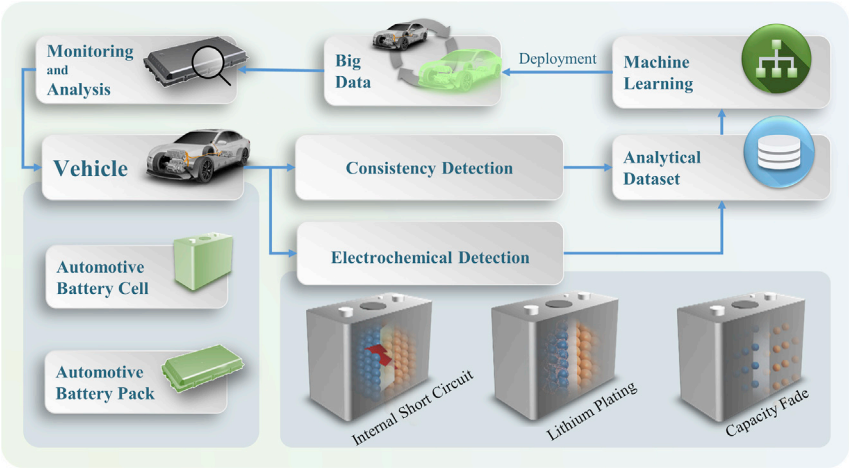


Figure 2. Schematic illustration of the cloud-based closed loop method for data generation

at the cell level) and field scenarios (huge amounts of data at the pack level) under real-world conditions makes it difficult to transfer academic progress to industrial applications (Aykol et al., 2020). In a typical example, field failure caused by a foreign object (e.g., metal particle) or a vehicle crash may evolve into hazardous failure with little warning during what appears to be normal operation of the battery. Offline safety modeling provides a greater understanding on battery safety issues. However, the problem is to update the model under complex, nonlinear, and random operating conditions in automotive applications, such as operating voltage window, charging/discharging rate, and operating temperature. As Finegan stated in a more recent perspective, the operational history of the cell plays a pivotal role in the prediction of the health state, because the degradation mechanisms hinge largely on the operating conditions (Finegan et al., 2020). In this formidable task, one of the most tactical approaches, to circumvent these limitations, is establishing digital twins for the battery system in cyberspace. Cloud-based cyber-physical-systems and platform technologies can serve as a best practice for creating self-improved AI models that are competent and resourceful to handle matters. In this study, we totally generate a comprehensive dataset consisting of failure/positive data (36 samples) and safety/negative data (58 samples) for the model training, calibration, and testing. Figure 2 illustrates how we use the cloud-based digital solution to facilitate date generation and achieve data-driven machine learning-based predictions.

Feature engineering

Over the past decade, data-driven machine learning-based techniques have proven capable of providing effective tools for scientific discovery and optimization in the field of energy storage. In the scope of machine learning tasks, feature engineering plays a pivotal role in the process of predictive modeling (Severson et al., 2019; Duboue, 2020). Insights into physical failure mechanisms would help guide physics-based data-driven models and establish feature engineering. However, efforts have been devoted worldwide to establishing performance metrics for various energy storage devices (Zhang and Pan, 2015; Cao et al., 2019; Noori et al., 2019; Zhao and Burke, 2020; Burke and Zhao, 2021). There are no well-defined metrics for evaluating the batteries in terms of safety and risk, which makes it much harder to automate feature engineering for an accurate description of the detailed physical failure mechanisms. Therefore, prior knowledge is a necessary part in the case of development of appropriate performance metrics/feature engineering in this prediction task. In this regard, a practice to establish an all-around, comprehensive evaluation method is the combination of electrochemical metrics at the cell level and statistical engineering at the pack level for predicting the state and risk of automotive batteries (Figure 3). A number of parameters have been amply demonstrated that they may trigger hazardous failure after long-term incubation, including internal short circuit (Liu et al., 2018a), lithium-plating (2019b), and capacity degradation (Fleischhammer et al., 2015). Therefore, they are used as electrochemical feature engineering (Figures S1–S3) in this study. Considering that a number of field failures occur without any noticeable symptoms during what appears to be a normal operation of the battery (Burke, 2014; Wang et al., 2019; Finegan et al., 2020; Chen et al., 2021), which makes the electrochemical-based prediction so singularly useless, in particular with the online historical battery data. In this case, the statistical features provide accurate numerical

CellPress
OPEN ACCESS

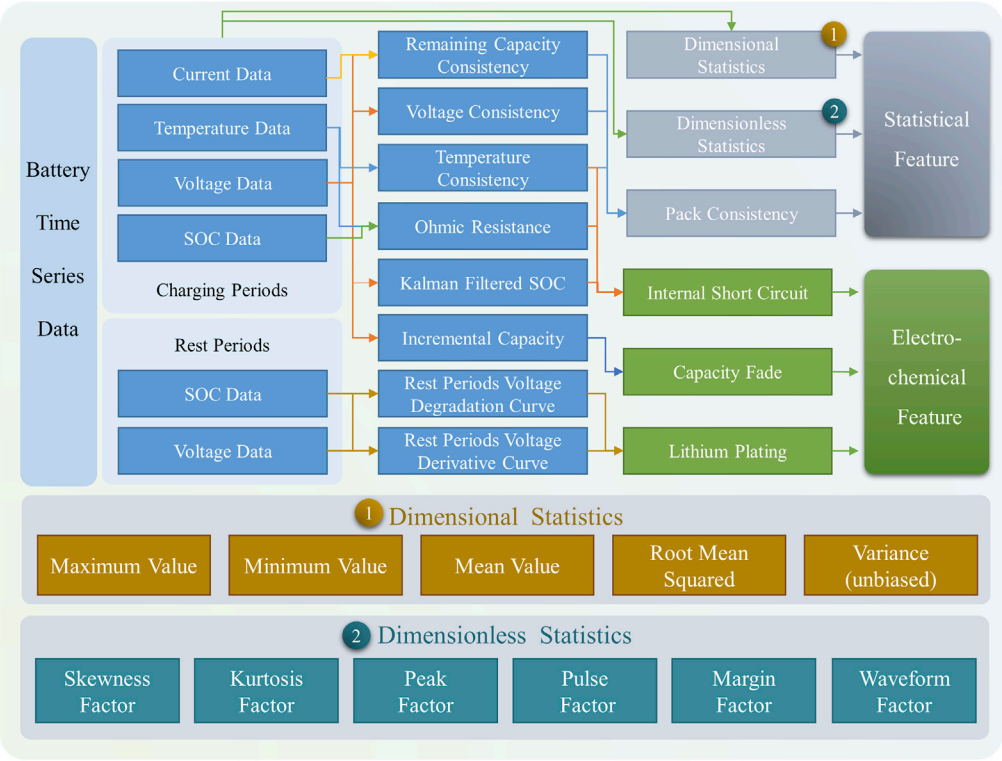


Figure 3. Performance metrics for the prediction of battery failure

Electrochemical metrics include internal short circuit, capacity fade, and lithium plating, which will be used for physics-guided supervised learning modeling. Statistical metrics includes dimensional, dimensionless statistics, and pack consistency for unsupervised learning. Both electrochemical and statistical metrics will be used for semi-supervised learning modeling.

calculations based on deviations and outliers of cell behavior at the pack level, and hence it can contribute to detect seemingly trivial details in the task of predicting battery failures. Battery pack consistency (Figure S4), in terms of voltage, temperature, and capacity, as well as nondimensional features such as kurtosis, skewness, crest, margin, impulse, and waveform factor during the operational lifetime offer advantages in analysis of dynamic signals, which contain much more information on a battery's state than a quasi-steady-state detection and analysis (Krewer et al., 2018). For a more detailed description of the measurement of electrochemical feature engineering and the formula for the statistical feature engineering, please see the STAR Methods.

A well-integrated machine learning technique

Machine learning using probabilistic frameworks are well suited for prediction tasks and have hence emerged as one of the principal theoretical and practical solutions to designing self-controlled machines to automatically recognize and organize patterns from training data (Ghahramani, 2015). In this study, a well-integrated machine learning technique using both the electrochemical-based and statistical feature engineering (see Feature engineering section) is developed to achieve robust and accurate prediction of automotive battery failure in real-world applications. The well-integrated technique consists of three different machine learning models: physics-guided supervised learning, high-dimensional unsupervised learning, and disagreement-based semi-supervised learning.

Generally, the sample data can be clearly labeled as positive samples (failure) or negative samples (safety) after a full examination in the laboratory. But it is not only impractical but is also impossible to check each vehicle in a laboratory under strictly controlled conditions. To achieve this goal, the prediction model must provide an acceptable level of accuracy. As such, a cloud-based closed loop framework is established to achieve the stated aims by using the data-driven, machine learning-based models. In short, the training

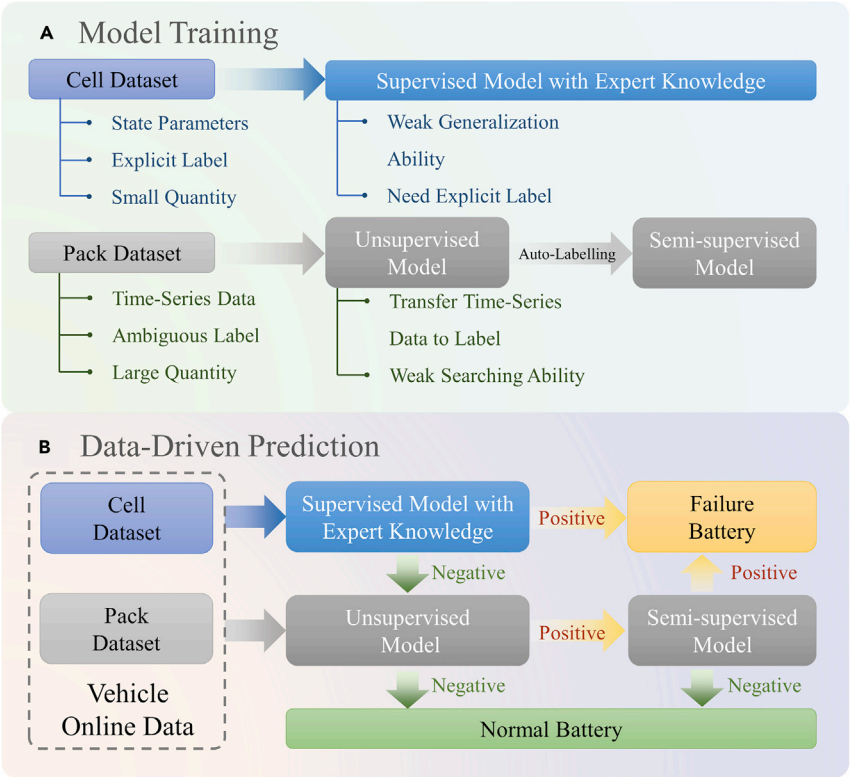


Figure 4. Data-driven prediction of automotive battery failure

(A) Data generation and model training. All data is stored on the cloud servers for cloud computing using the developed computational data-driven models.

(B) Online fault detection and early-warning is achieved by a well-integrated model consisting of three different probabilistic machine learning models.

data are used for physics-guided supervised learning modeling and verifying the effectiveness of the electrochemical performances. An unsupervised learning model is trained to discern multiple categories and classify samples into either a safety or a failure group in a high-dimensional space. To handle the unclear, inadequate input–output pairs, a semi-supervised learning model is developed. The structure of the machine learning modeling process is illustrated in Figure 4.

Physics-guided supervised learning

In the case of battery failure prediction, supervised learning offers advantages in designing the safety models to identify the electrochemical behavior of cells that may trigger failure through incorporating domain-specific knowledge into machine learning models. It would be intriguing and potentially beneficial to combine the knowledge with machine intelligence to automatically distinguish safe and unsafe cells in a data-driven manner. In the prediction of battery failure, the physics-guided machine learning model significantly improves the interpretability of the prediction reasoning, which is of great value for uncovering the underlying physics and mechanisms. An expert system consisting of engineering experience and domain knowledge is developed for supervised learning, which can extract what experts know—a process called knowledge acquisition (Russell and Norvig, 2020). With such a physics-guided learning, the model not only recognizes faulty information but also provides reasonable interpretation on what happens to the observed data (i.e., battery cell).

The most intuitive way to recognize battery failure is to build a collection of rules. In this study, fuzzy set theory and multi-criteria decision-making approach have been employed in the supervised learning modeling. The application of the fuzzy set theory is used to determine the type of fundamental parameters, such as the abnormal temperature increment, abnormal voltage fluctuation, and drop. Further, as discussed earlier, the underlying electrochemical abuse conditions include internal short circuit, lithium

CellPress
OPEN ACCESS

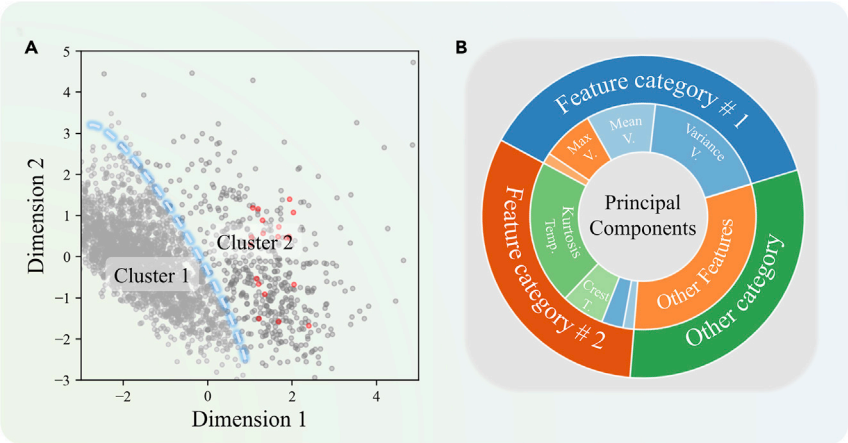


Figure 5. Unsupervised machine learning

(A) The unsupervised clustering is achieved by K-Means in a 2D space after a reconstruction using principal component analysis (PCA) method. Batteries were categorized into two clusters, which are characterized by high-density and low-density regions with plenty of discrete points, respectively. All failure samples (red points) are clustered in the second classification.

(B) Feature engineering used for unsupervised learning.

plating, and abnormal capacity fade. Therefore, these key parameters are used as feature engineering for developing predictive models. Symptoms are analyzed and evaluated by the expert system using the analytic hierarchy process. Every symptom is given a weight to calculate the fuzzy membership grade of a certain kind of failure. Symptom inspection is achieved by a rule established in the decision tree. The rule-making for the supervised learning model follows the principle of "information entropy," which gives priority to the rules with high differentiation.

However, in some cases, the classification test exhibits that the domain knowledge-based supervised learning model fits the observed data very well but shows a poor generalization performance because of extrapolation or observational biases and most importantly, lack of insufficient labeled data. The labeling effort and the analysis work are very time-consuming, making the modeling infeasible when the unlabeled dataset is in a tremendous volume.

High-dimensional unsupervised learning

The physics-guided supervised learning offers advantages in the interpretation of what we can specifically learn from the model. Although recent studies have shed some light on the common characteristics of battery failure, the failure mechanism is still unclear in particular for the cases in real-world applications. Therefore, domain knowledge-based supervised systems are inherently limited to fairly low-dimensional problems for a small amount of the sample data in this case. In addition, a number of field failures occur without any noticeable symptoms during what appears to be normal operation of the battery, which makes the electrochemical-based prediction so singularly useless. Therefore, supervised learning from the perspective of electrochemical performance evaluation is far from a guarantee of battery safety during its operational lifetime. This also shows how crucial seemingly trivial details could be. In this regard, statistical feature-based probabilistic machine learning offers distinct advantages in this prediction task. To solve this problem, unsupervised learning provides an effective solution to discerning multiple categories from a high-dimensional space in a collection of unlabeled data (Figure 5). It can effectively recognize outliers using clustering analysis, which is a promising approach to divide safe and risky cells into different categories from a form of plausible reasoning based on degree of similarity and convergence. In practice, this unsupervised learning offers unexpected earnings when adopting the metrics that are often utilized in the time domain analysis of signals. However, both lack of labels and noise create a continuum between supervised and unsupervised learning. Although unsupervised learning can effectively categorize the samples into different groups, it can be seen that the failed samples cannot be accurately separated from the large samples of data. Merging the best properties of each of the learning methods opens up opportunities to develop a robust prediction of battery failure that possesses the advantages of the individual machine

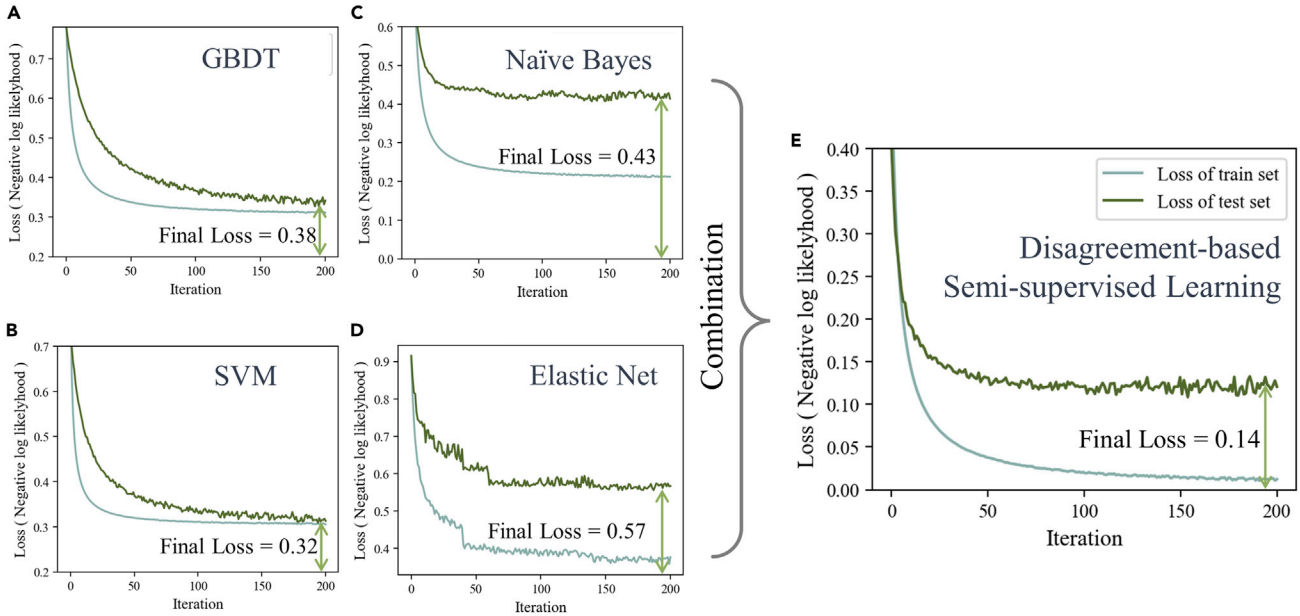


Figure 6. The NLLP of the disagreement-based learning model

The quadratic loss function of the supervised models using GBDT, SVM, naïve Bayes, and EN are shown in the (A), (B), (C), and (D) respectively. The final loss for the training set and testing set are both much lower than those losses of the four supervised learning models, as shown in the (E).

blocks, in addition to eliminating the associated shortcomings. Hence, semi-supervised learning is developed to tackle the challenges in this subject, which is fully discussed in the next section.

Disagreement-based semi-supervised learning

To achieve more significant improvements in the accuracy of machine learning, it is of great importance to address the challenge-lack of sufficient labeled data. In this case, disagreement-based semi-supervised learning is developed to overcome the challenges of generating plentiful class-labeled data (Figure S5). In the development of disagreement-based models, four different supervised machine learning methods are used, namely, including gradient boosting (GB), Bayesian regression (BR), support vector machine (SVM), and elastic net (EN). All of these methods have proven to be effective for the prediction of battery failure, and each of these techniques have a particular advantage in different situations of data distribution. GB achieves a good balance between bias and variance by combining multiple 'base' classifiers to produce a form of committee, in which the weighting coefficient for each data point depends on the performance of the previous classifiers (known as low-bias weak learners). SVMs are better suited for the prediction of big data that have high sampling rate (e.g., 1 s) and high dimensional feature space. Bayes regression in general has a better robust and efficient performance if training data contains both continuous and discrete features, because each can be represented separately using appropriate probability assumptions. EN is preferred in the case of simply distributed training data with noise or there are high correlations between the features.

The utilization of disagreement-based learning has several advantages (He et al., 2020): (1) the more learners in this method, the more accuracy pseudo labels have. (2) The model can effectively reduce variance of each learner, in addition to removing overall bias among the whole dataset. (3) The final model can ignore invalid data because of the "Majority Voting" mechanism. To test the disagreement-based model, the negative log likelihood probability (NLLP) is used as a parameter for multimode process monitoring (Wu et al., 2020). As shown in Figure 6, the disagreement-based semi-supervised machine learning model achieves a much better "loss" and results in improved predictions than the four separate supervised models. Please see STAR Methods for a more detailed description of the machine learning modeling and the established safety envelope using each machine learning technique could be found in Figures S6–S9.

CellPress
OPEN ACCESS

Table 1. Model metrics for training with a primary modeling and model refining

	Train Data	1st Refining	2nd Refining	3rd Refining
Positive samples	8	4	3	9
Negative samples	50	2	3	2
Total samples	58	6	6	11
Training accuracy rate	75.9%	89.1%	94.3%	96.3%
Precision in validation	66.7%	50.0%	81.8%	90.0%

Training and refining refer to the data used to learn the model and improve model performance, respectively.

Cloud-based digital solution

In practical application scenarios such as EVs, the processor onboard can be used to control and manage the EV battery system in real time. However, the onboard processor in general has a limited resource of data storage and processing capability; therefore, the use of large datasets combined with pretrained data-driven models is better suited for off-board use. To achieve the task of predicting battery failure, there is an urgent need for more efficient operation of the complex computational models. In this case, the cloud platform effectively alleviates the computational burden of onboard microprocessors on the vehicle because of its exponential increase in the computational power and storage space. In addition, cloud-based computing and storage also has a tremendous amount of flexibility on how and where to perform the mission unconstrained by physical space, benefiting from its distributed nature. This high-speed, low-cost computational environment provides fertile ground for training date-driven artificial intelligence and machine learning models to achieve the tasks at the low unit cost.

Initially, raw battery data during the daily operation of the EVs are seamlessly uploaded to the cloud server and hence the battery's digital twin can be built in cyberspace. As an up-to-date virtual technology, the digital twin opens up new avenues of investigation, in which it not only provides full information about the battery system from the sensors but also offers advantages in structuring the computational models to accomplish the task of prediction of battery failure. One benefit of this: the cloud-based data-driven framework can generate longitudinal electronic health records periodically based on the machine learning models developed. In addition, the cloud-based solution enables continuous improvement and iteration in the battery failure prediction for EVs with the increase of battery data under realistic conditions. Thus, remote monitoring, fault diagnosis and early warning is available because of the technological innovation, namely, IoT.

RESULTS AND DISCUSSION

In this study, the dataset for machine learning modeling is established using the data that are uploaded to a cloud monitoring platform to best address the safety issues of automotive batteries in real-world applications. Totally, we generate a comprehensive dataset consisting of failure/positive data (36 samples) and safety/negative data (58 samples), which is based on the monitoring and analysis using onboard diagnostics provided by the closed loop self-taught machine learning system. Initially, the training dataset (8 positive samples and 50 negative samples) were used to develop an early-prediction machine learning model. Failure is a necessary part in the predictive coding process because from failure comes learning, iteration, adaptation, and the establishment of new machine learning models through an iterative learning process in a high-dimensional state-space. Using the early-prediction machine learning models, the cloud-based closed solution achieves a continuous lifelong learning based on the confusion matrix by statistical analysis. We calibrate the model 3 times to accelerate training and enhance generalization, which generates 16 positive samples and achieves a training classification accuracy of 96.3%. Table 1 displays the performance of the well-integrated machine learning model applied to our one training dataset and three refining datasets. We then evaluate the model on two testing datasets (the primary one is 10 samples, the secondary one is 3 samples), which attains an average test classification precision of 92.3% (only one sample is misclassified out of the 13 samples), as shown in Table 2. Besides, the well-integrated model provides an accurate and robust prediction of the health condition of the samples, as shown in Figure 7.

Dynamic and multidimensional analysis on the cloud server-farms opens up a new world for developing a continual learning method to deal with the emergence of new data sources at unknown time points during

Table 2. Model metrics for testing

	1st Test	2nd Test
Positive samples	9	3
Negative samples	1	0
Total samples	10	3
Precision of the classifier	90.0%	100.0%

Primary/secondary test refers to the data used to evaluate model performance.

the daily operation of the vehicles while enabling models to adapt to new operating environment (e.g., different vehicle models but same battery cell, various driving and using behaviors) with excellent robustness. The focus of continual lifelong learning using machine learning techniques is to enhance the ability to handle new tasks in a closed loop model, where the model must learn online over a continuous stream of nonstationary data (Harari, 2017; Parisi et al., 2019). In this case, the continual lifelong learning is achieved by using the cloud-based closed loop system with nonstop action and operation. When the machine learning model tends to be stable and shows an overall uniform convergence, it is able to form a safety envelope of battery failure in real-world applications while classifying samples data as either a safety or a failure group, as shown in Figure 8. In addition, regression models can be trained using the labeled data to quantify the risk of battery failure, which is illustrated in Figure 9.

Conclusions

New technologies are driven primarily by their ability to provide affordable, efficient solutions to problems previously impossible to solve. Considering the challenge of accurately predicting complex, nonlinear systems, closed-based machine learning techniques using probabilistic frameworks are well suited for prediction tasks. This study highlights the difficulties of realizing robust and accurate prediction of EV battery failure in real-world applications and offers a cloud-based well-integrated machine learning solution. We develop a data-driven early-prediction model using the training data generated from commercial battery packs using NMC/graphite cells. The proposed cloud-based closed-loop framework using machine learning models can learn from the historical EV battery data and generate longitudinal electronic health records in cyberspace through the establishment of digital twins for battery

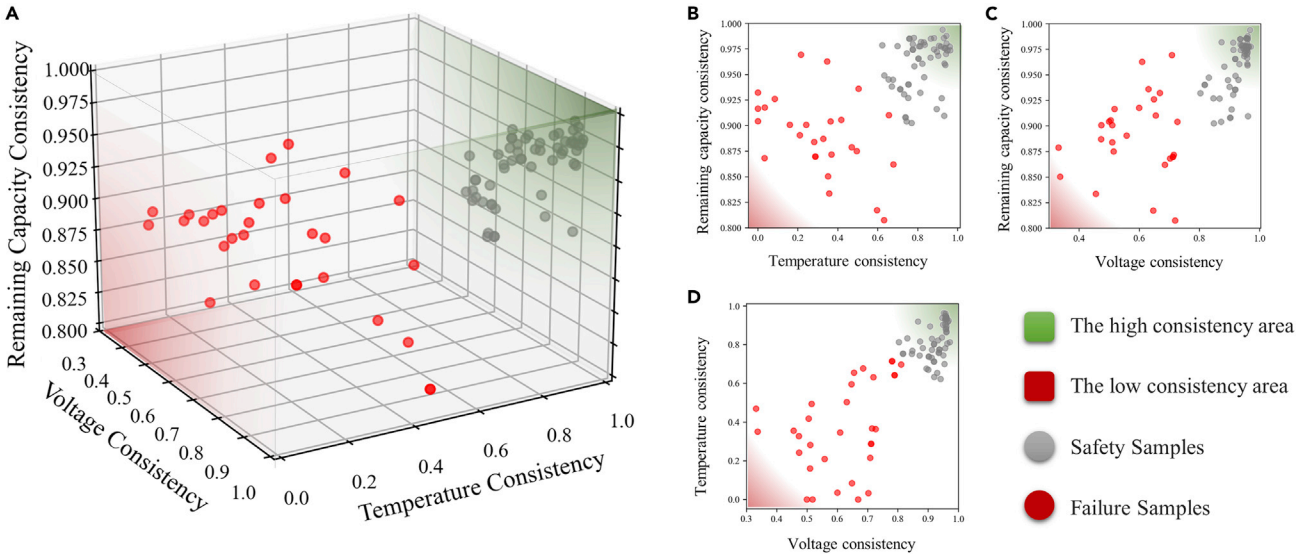


Figure 7. The well-integrated machine learning model for the classification

The green and red points represent negative samples (safety) and positive samples (failure), respectively.

(A) Scatter plot in the 3D state space.

(B–D) are the scatter plots in the two-dimensional plane.

CellPress
OPEN ACCESS

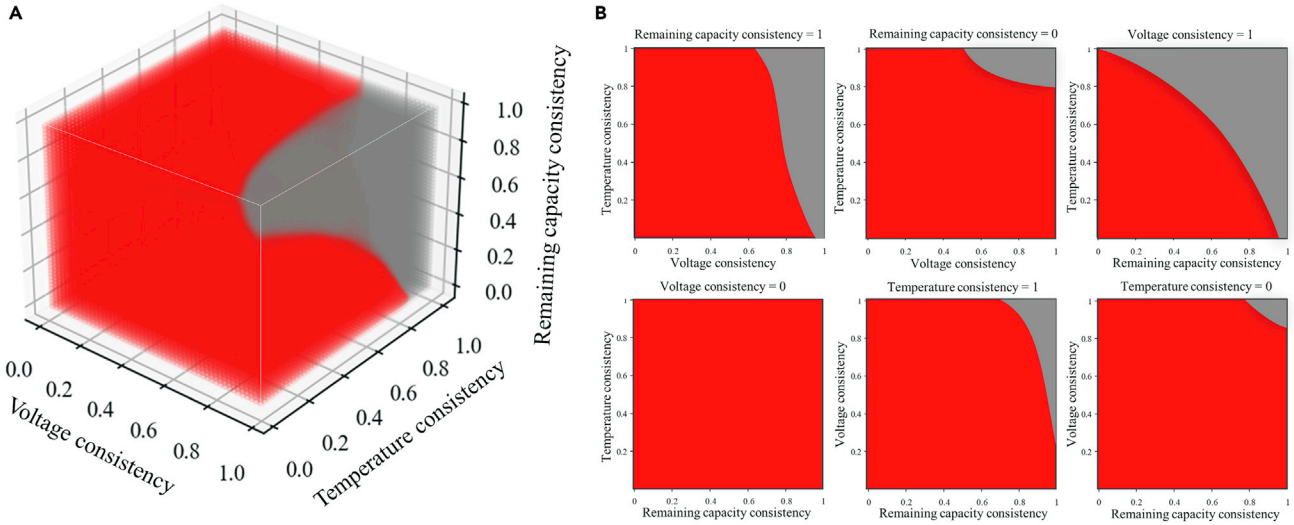


Figure 8. Classification model for the development of safety envelope in the prediction of battery failure

The red region and gray region represent the unsafe and the safe ranges, respectively.

(A) Classification in three-dimensional space.

(B) Classification in each two-dimensional plane.

systems. In the classification setting, we obtain a test error of 7.7% using the field data. This level of accuracy is achieved by a well-integrated machine learning model, and observational, empirical, physical, and statistical understanding of the electrochemical energy storage systems. The authors believed that the closed-based machine learning techniques not only provide an innovative method for achieving accurate prediction of battery failure, but they also offer a very grounded view of the multipronged approach to become continual lifelong learning taking advantage of the self-taught and self-improved artificial intelligence techniques with nonstop availability to the server-farms. We hope this effort will provide a useful contribution to the energy-storage and the electrification of the transportation community

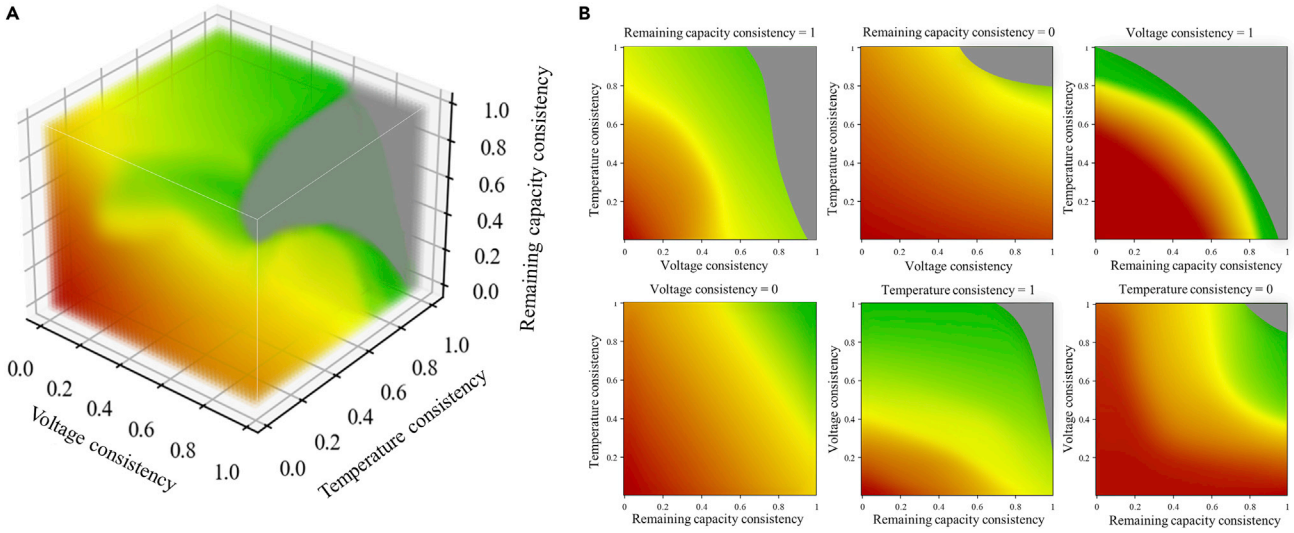


Figure 9. Regression model for the development of safety envelope in the prediction of battery failure

The colored region and the gray region represent the unsafe and the safe ranges, respectively.

(A) Regression analysis shown in the three-dimensional space.

(B) Regression analysis in each two-dimensional plane.

as a roadmap for ameliorating the hurdles for the prediction of automotive battery failure in real-world applications.

Limitations of the study

It should be noted that the present study does have some limitations. First, although the closed loop machine learning-based framework provides a robust and accurate method in both predicting and classifying automotive battery cells by their risk, the study did not provide new insights into physical failure mechanisms and how specific phenomena govern risk. More effort needs to be focused on the investigation of failure mechanisms at the molecular, atomic, single-atom level, and establishing methods for mechanism identification. However, it is a significantly positive compromise and up-to-now an essential one because of the huge cost of the establishment of the test matrix that can cover all types of failure mechanisms, loading patterns, and field scenarios. Second, the as-developed safety envelope is visualized with two types of phase diagrams: a classifier that can fast distinguish safety and unsafety in terms of the consistency of the battery pack for the online EVs in a data-driven manner and a regressor that can quantitatively the risk and exhibits the information that rises to the level of concern. However, the safety envelope was developed for one type of battery pack using prismatic NMC/graphite cells in-series with specific cell balancing and thermal management strategy and is therefore not universal.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Electrochemical feature engineering at the cell level
  - Statistical feature engineering at the pack level
  - Supervised learning methodology
  - Unsupervised learning methodology
  - Semi-supervised learning methodology
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104172.

ACKNOWLEDGMENTS

The authors acknowledge support from BYD electric vehicle technology development (EVTD) center.

AUTHOR CONTRIBUTIONS

Y.L. and H.L. conceived the idea and secured funding for the establishment of cloud technology and internet of things. J.Z. designed the research and wrote the paper. J.W. prepared the figures and supplemental information. J.Z. and J.W. developed the ML models. A.F.B. proposed scientific hypothesis and contributed to the revising of the manuscript. All authors approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 25, 2021
Revised: February 27, 2022
Accepted: March 23, 2022
Published: April 15, 2022

CellPress
OPEN ACCESS

REFERENCES

Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., and Ayyash, M. (2015). Internet of things: a survey on enabling technologies, protocols, and applications. IEEE Commun. Surv. Tutor. *17*, 2347–2376.

Attia, P.M., Grover, A., Jin, N., Severson, K.A., Markov, T.M., Liao, Y.H., Chen, M.H., Cheong, B., Perkins, N., Yang, Z., et al. (2020). Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature *578*, 397–402.

Aykol, M., Herring, P., and Anapolsky, A. (2020). Machine learning for continuous innovation in battery technologies. Nat. Rev. Mater. *5*, 725–727.

Bla, B., Yja, B., Cya, B., Lwa, B., Xiang, G., and Sha, Y. (2020). Safety issues and mechanisms of lithium-ion battery cell upon mechanical abusive loading: a review. Energy Storage Mater. *24*, 85–112.

Burke, A.F., and Zhao, J. (2021). Past, present and future of electrochemical capacitors: technologies, performance and applications. J. Energy Storage *35*, 102310.

Burke, A.F. (2014). Battery safety for lithium batteries in vehicle applications. In Encyclopedia of Automotive Engineering (John Wiley & Sons, Inc.), pp. 1–15.

Campbell, I.D., Marzook, M., Marinescu, M., and Offer, G.J. (2019). How observable is lithium plating? differential voltage analysis to identify and quantify lithium plating following fast charging of cold lithium-ion batteries. J. Electrochem. Soc. *166*, A725–A739.

Cao, Y., Li, M., Lu, J., Liu, J., and Amine, K. (2019). Bridging the academic and industrial metrics for next-generation practical batteries. Nat. Nanotechnol. *14*, 200–207.

Chen, Y., Kang, Y., Zhao, Y., Wang, L., Liu, J., Li, Y., Liang, Z., He, X., Li, X., Tavajhi, N., et al. (2021). A review of lithium-ion battery safety concerns: the issues, strategies, and testing standards. J. Energy Chem. *59*, 83–99.

Chombo, P.V., and Laoonual, Y. (2020). A review of safety strategies of a li-ion battery. J. Power Sources *478*, 228649.

Correa-Baena, J.P., Hippalgaonkar, K., van Duren, J., Jaffer, S., Chandrasekhar, V.R., Stevanovic, V., Wadia, C., Guha, S., and Buonassisi, T. (2018). Accelerating materials development via automation, machine learning, and high-performance computing. Joule *2*, 1410–1420.

Deng, J., Bae, C., Marcicki, J., Masias, A., and Miller, T. (2018). Safety modelling and testing of lithium-ion batteries in electrified vehicles. Nat. Energy *3*, 261–266.

Deyab, M.A., Mele, G., and Bloise, E. (2021). Novel nanocomposites of Ni-Pc/polyaniline for the corrosion safety of the aluminum current collector in the Li-ion battery electrolyte. Sci. Rep. *11*, 12371.

Drake, N. (2014). Cloud computing beckons scientists. Nature *509*, 543–544.

Duboue, P. (2020). The Art of Feature Engineering: Essentials for Machine Learning (Cambridge University Press).

Feng, X., Weng, C., Ouyang, M., and Sun, J. (2016). Online internal short circuit detection for a large format lithium ion battery. Appl. Energy *161*, 168–180.

Feng, X., Ouyang, M., Liu, X., Lu, L., Xia, Y., and He, X. (2017). Thermal runaway mechanism of lithium ion battery for electric vehicles: a review. Energy Storage Mater. *10*, 246–267.

Feng, X., Yue, P., He, X., Li, W., and Ouyang, M. (2018). Detecting the internal short circuit in large-format lithium-ion battery using model-based fault-diagnosis algorithm. J. Energy Storage *18*, 26–39.

Feng, X., Weng, C., He, X., Han, X., Lu, L., Ren, D., and Ouyang, M. (2019). Online state-of-health estimation for li-ion battery using partial charging segment based on support vector machine. IEEE Trans. Veh. Technol. *68*, 8583–8592.

Feng, X., Ren, D., He, X., and Ouyang, M. (2020). Mitigating thermal runaway of lithium-ion batteries. Joule *4*, 743–770.

Finegan, D.P., and Cooper, S.J. (2019). Battery safety: data-driven prediction of failure. Joule *3*, 2594–2601.

Finegan, D.P., Zhu, J., Feng, X., Keyser, M., and Cooper, S.J. (2020). The application of data-driven methods and physics-based learning for improving battery safety. Joule *5*, 316–329.

Fleischhammer, M., Waldmann, T., Bisle, G., Hogg, B.I., and Wohlfahrt-Mehrens, M. (2015). Interaction of cyclic ageing at high-rate and low temperatures and safety in lithium-ion batteries. J. Power Sources *274*, 432–439.

Foulds, C., and Christensen, T.H. (2016). Funding pathways to a low-carbon transition. Nat. Energy *1*, 1–4.

Galushkin, N.E., Yazvinskaya, N.N., and Galushkin, D.N. (2018). Mechanism of thermal runaway in lithium-ion cells. J. Electrochem. Soc. *165*, A1303–A1308.

Gao, T., and Lu, W. (2021). Machine learning toward advanced energy storage devices and systems. iScience *24*, 101936.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature *521*, 452–459.

Harari, Y.N. (2017). Reboot for the AI revolution. Nature *550*, 324–327.

He, H., Han, D., and Dezert, J. (2020). Disagreement based semi-supervised learning approaches with belief functions. Knowl. Based Syst. *193*, 105426.

Krewer, U., Röder, F., Harinath, E., Braatz, R.D., Bedürftig, B., and Findeisen, R. (2018). Dynamic models of Li-Ion batteries for diagnosis and operation: a review and perspective. J. Electrochem. Soc. *165*, A3656.

Li, W., Zhu, J., Xia, Y., Gorji, M.B., and Wierzbicki, T. (2019a). Data-driven safety envelope of lithium-ion batteries for electric vehicles. Joule *3*, 2703–2715.

Li, Y., Feng, X., Ren, D., Ouyang, M., and Han, X. (2019b). Thermal runaway triggered by plated lithium on the anode after fast charging. ACS Appl. Mater. Interfaces *11*, 46839–46850.

Li, W., Rao, S., Xiao, Y., Gao, Z., Chen, Y., Wang, H., and Ouyang, M. (2021a). Fire boundaries of lithium-ion cell eruption gases caused by thermal runaway. iScience *24*, 102401.

Li, X., Gao, P., and Lai, Y.Y. (2021b). Symmetry-breaking design of an organic iron complex catholyte for a long cyclability aqueous organic redox flow battery. Nat. Energy *6*, 873–881.

Liu, B., Jia, Y., Li, J., Sha, Y., and Xu, J. (2018a). Safety issues caused by internal short circuits in lithium-ion batteries. J. Mater. Chem. A *6*, 21475–21484.

Liu, K., Liu, Y., Lin, D., Pei, A., and Cui, Y. (2018b). Materials for lithium-ion battery safety. Sci. Adv. *4*, eaas9820.

Lombardo, T., Duquesnoy, M., and El-Bouysidy, H. (2021). Artificial intelligence applied to battery research: hype or reality? Chem. Rev. https://doi.org/10.1021/acs.chemrev.1c00108.

Manthiram, A., and Goodenough, J.B. (2021). Layered lithium cobalt oxide cathodes. Nat. Energy *6*, 323.

Ng, M.F., Zhao, J., Yan, Q., Conduit, G.J., and Seh, Z.W. (2020). Predicting the state of charge and health of batteries using data-driven machine learning. Nat. Mach. Intell. *2*, 161–170.

Nitta, N., Wu, F., Lee, J.T., and Yushin, G. (2015). Li-ion battery materials: present and future. Mater. Toady *18*, 252–264.

Noori, A., El-Kady, M.F., Rahmanifar, M.S., Kaner, R.B., and Mousavi, M.F. (2019). Towards establishing standard performance metrics for batteries, supercapacitors and beyond. Chem. Soc. Rev. *48*, 1272–1341.

Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: a review. Neural Netw. *113*, 54–71.

Pastor-Fernandez, C., Uddin, K., Chouchelamane, G.H., Widanage, W.D., and Marco, J. (2017). A comparison between electrochemical impedance spectroscopy and incremental capacity-differential voltage as Li-ion diagnostic techniques to identify and quantify the effects of degradation modes within battery management systems. J. Power Sources *360*, 301–318.

Plötz, P., Axsen, J., Funke, S., and Gnann, T. (2019). Designing car bans for sustainable transportation. Nat. Sustain. *2*, 534–536.

Ren, D., Kandler, S., Guo, D., Han, X., Feng, X., and Lu, L. (2018). Investigation of lithium plating-stripping process in li-ion batteries at low temperature using an electrochemical model. J. Electrochem. Soc. *165*, A2167–A2178.

Roberts, A.D., Li, X., and Zhang, H. (2014). Porous carbon spheres and monoliths: morphology control, pore size tuning and their applications as li-ion battery anode materials. Chem. Soc. Rev. *45*, 4341–4356.

Roman, D., Saxena, S., Robu, V., Pecht, M., and Flynn, D. (2021). Machine learning pipeline for battery state-of-health estimation. Nat. Mach. Intell. *3*, 447–456.

Russell, S., and Norvig, P. (2020). In Artificial Intelligence: a Modern Approach, Fourth edition (Pearson).

Sendek, A.D., Cubuk, E.D., Antoniuk, E.R., Cheon, G., Cui, Y., and Reed, E.J. (2019). Machine learning-assisted discovery of many new solid li-ion conducting materials. Chem. Mater. *31*, 342–352.

Severson, K.A., Attia, P.M., Jin, N., Perkins, N., Jiang, B., Yang, Z., Chen, M.H., Aykol, M., Herring, P.K., Fraggedakis, D., et al. (2019). Data-driven prediction of battery cycle life before capacity degradation. Nat. Energy *4*, 383–391.

Sun, P., Bisschop, R., Niu, H.C., and Huang, X. (2020). A review of battery fires in electric vehicles. Fire Technol. *56*, 1–50.

Thackeray, M.M., and Amine, K. (2021). Layered Li–Ni–Mn–Co oxide cathodes. Nat. Energy *6*, 933.

Wang, Q., Mao, B., Stoliarov, S.I., and Sun, J. (2019). A review of lithium-ion battery failure mechanisms and fire prevention strategies. Prog. Energy Combust. Sci. *73*, 95–131.

Wu, H., Zhuo, D., and Kong, D. (2014). Improving battery safety by early detection of internal shorting with a bifunctional separator. Nat. Commun. *5*, 5193.

Wu, D., Zhou, D., Zhang, J., and Chen, M. (2020). Multimode process monitoring based on fault dependent variable selection and moving window-negative log likelihood probability. Comput. Chem. Eng. *136*, 106787.

Xiong, R., Ma, S., Li, H., Sun, F., and Li, J. (2020). Toward a safer battery management system: a critical review on diagnosis and prognosis of battery short circuit. iScience *23*, 101010.

Yang, X.G., Ge, S., Liu, T., Leng, Y., and Wang, C.Y. (2018). A look into the voltage plateau signal for detection and quantification of lithium plating in lithium-ion cells. J. Power Sources *395*, 251–261.

Zeng, X., Li, M., El-Hady, D.A., Alshitari, W., and Amine, K. (2019). Commercialization of lithium battery technologies for electric vehicles. Adv. Energy Mater. *9*, 1900161.

Zhang, S., and Pan, N. (2015). Supercapacitors performance evaluation. Adv. Energy Mater. *5*, 1401401.

Zhao, J., and Burke, A.F. (2020). Electrochemical capacitors: performance metrics and evaluation by testing and analysis. Adv. Energy Mater. *11*, 2002192.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
NMC 622	BYD Auto Industry Co., Ltd.	provided by BYD Auto Industry Co., Ltd.
Software and algorithms		
Python 3.8	Python Software Foundation	https://www.python.org
Matlab 2020a	Matlab - MathWorks	https://matlab.mathworks.com

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be handled by the lead contact, Jingyuan Zhao (jyzhao_scholar@yeah.net).

Materials availability

This study did not generate new materials.

Data and code availability

The data generated in this study could not be shared due to confidentiality (proprietary from BYD company).

The original code is available for academic purposes from the lead contact upon reasonable request.

Any additional information in this paper is available from the lead contact upon request.

METHOD DETAILS

Electrochemical feature engineering at the cell level

Internal short circuit

Online ISC detection can be a valuable indicator of possible battery failure in EV use (Figure S1). In this regard, Ouyang and coworkers provide an effective online ISC detection method for both the battery pack and cells (Feng et al., 2016, 2018). Four indicators were used for the ISC characterization: (i) SOC imbalance, (ii) abnormal heat generation, (iii) cell temperature deviation and (iv) cell voltage deviation.

Lithium plating

Differential voltage analysis using small charge or discharge currents is one common method for detection of lithium plating (Figure S2). Voltage plateau during the rest period after a charge stage is considered as an appropriate metric for the evaluation of lithium stripping process. Methods related to differential analysis of the voltage plateau have been proposed and widely used for the estimation of the amount of plated lithium (Ren et al., 2018; Yang et al., 2018; Campbell et al., 2019). Study has shown that it takes a longer time to reach dV/dt peak when larger amount of lithium is plating. It is also suggested that the duration of voltage plateau (i.e. the time to reach the dV/dt peak) can be an important performance metric for quantifying the amount of plated lithium under different conditions.

Capacity fade

Up to now, a variety of techniques have been developed for state of health estimation of lithium-ion battery (Figure S3). The feature-based data-driven model is a promising method for the prediction of battery SOH, which has recently been demonstrated in a study by Severson and coworkers. In addition, incremental capacity analysis (ICA) is another effective method in providing detailed information to identify battery fade

(Pastor-Fernandez et al., 2017; Feng et al., 2019). In EV use, ICA has a particular advantage for on-board prediction using partial charge/discharge voltage segments.

Statistical feature engineering at the pack level

In practice, the voltage values of the cells for each battery pack in a charging process can be treated as signal information, for which the voltage curves equivalent to the waveforms in the time domain. Therefore, the voltage values can be evaluated using the same method as used in the signal analysis and then used for the unsupervised learning modelling. The formula for the statistical feature engineering can be obtained as follows:

Categories	Statistics	Equation
Dimensionless Statistic	Maximum Value	$\Delta V_{\max} = \max(\{\Delta V_i\})$
	Mean Value	$\overline{\Delta V} = \frac{1}{N}\sum_{i=1}^{N}\Delta V_i, N = \|\|\{\Delta V_i\}\|\|$
	Root Mean Squared	$\mathrm{RMS}(\Delta V) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\Delta V_i\|^2}$
	Variance (unbiased)	$\mathrm{Var}(\Delta V) = \frac{1}{N-1}\sum_{i=1}^{N}(\Delta V_i - \overline{\Delta V})^2$
Dimensionless Statistic	Skewness Factor	$\mathrm{Skew}(\Delta V) = \frac{1}{(N-1)\cdot Var(\Delta V)^{3/2}}\sum_{i=1}^{N}(\Delta V_i - \overline{\Delta V})^3$
	Kurtosis Factor	$\mathrm{Kur}(\Delta V) = \frac{1}{(N-1)\cdot \mathrm{Var}(\Delta V)^2}\sum_{i=1}^{N}(\Delta V_i - \overline{\Delta V})^4 - 3$
	Peak Factor	$\mathrm{Peak}(V) = \frac{\Delta V_{\max}}{\mathrm{RMS}(\Delta V)}$
	Pulse Factor	$\mathrm{Pul}(V) = \frac{\Delta V_{\max}}{\overline{\Delta V}}$
	Margin Factor	$\mathrm{Mar}(V) = \frac{\Delta V_{max}}{\left(\frac{1}{n}\sum_{i=1}^{n}\sqrt{\|\Delta V_i\|}\right)^2}$
	Waveform Factor	$\mathrm{Wav}(V) = \frac{\mathrm{RMS}(\Delta V)}{\overline{\Delta V}}$

Supervised learning methodology

For a dataset D and a feature f, there are V dividing methods that could split dataset D by feature f. The information gain is defined as

$$\mathrm{Gain}(D, f) = \mathrm{Ent}(D) - \sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v) \quad \text{(Equation 1)}$$

The information entropy (*Ent*) of a dataset with K kind of labels is defined as.

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|K|}p_k \log_2 p_k \quad \text{(Equation 2)}$$

where p_k is the frequency of data with a label of k_{th}.

Besides being the basis of estimating threshold, the information gain is also the decisive factor of determining which feature would be the best feature to divide the dataset. In knowledge-based decision tree model, the feature selection is accomplished by domain knowledge. Machine learning task in this model is only to calculate the threshold.

Unsupervised learning methodology

Principal components analysis

Principal Component Analysis (PCA) is an effective method to reduce the complexity of high-dimensional data. To project points in original space into a vector, the centroids of the original space should be given by

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \text{(Equation 3)}$$

Assume that u_1 is the projection vector and that the variance after projection is

$$\frac{1}{N}\sum_{n=1}^{N}\left\{u_1^{\mathrm{T}} x_n - u_1^{\mathrm{T}}\bar{x}\right\}^2 = u_1^{\mathrm{T}} S u_1 \quad \text{(Equation 4)}$$

Using the Lagrange multiplier method:

$$\mathcal{L} = u_1^{\mathrm{T}} S u_1 + \lambda_1\left(1 - u_1^{\mathrm{T}} u_1\right) \quad \text{(Equation 5)}$$

Taking the derivative of the above equation and let the equation be equal to zero,

$$S u_1 = \lambda_1 u_1 \quad \text{(Equation 6)}$$

which is the standard eigenvalue expression, where λ corresponds to the eigenvalue and u corresponds to the eigenvector. The condition for obtaining the maximum value on the left side of the above equation is when λ1 is maximum, i.e., when the maximum eigenvalue is obtained. Suppose we want to project a D-dimensional data space into an M-dimensional data space (M<D), then we take the first M eigenvectors to form the projection matrix.

K-means clustering

As a clustering technique, the K-Means randomly selects k points as the initial cluster centers, and then calculates the distances from each data point to each cluster center and categorizes the data object into the class of the nearest cluster center. The center of adjusted clusters is computed with new cluster data. In each iteration, the classification of each sample is checked to see if it is correct, and if it is not, it is adjusted. After adjusting all the data, modify the cluster centers to enter the next iteration. If all data objects are correctly classified in once iteration, there will be no adjustment on the cluster center, which indicates that clustering process has converged.

The loss function of K-Means is:

$$J = \sum_{i=1}^{C}\sum_{j=1}^{N} r_{ij} v\left(x_j, \mu_i\right) \quad \text{(Equation 7)}$$

where

$$v\left(x_j, \mu_i\right) = \|x_j - \mu_i^2\|,\ r_{nk} = \begin{cases} 1 \text{ if } x_n \in k \\ 0 \text{ if } x_n \notin k \end{cases} \quad \text{(Equation 8)}$$

To get the maximum value, let the partial derivative of the loss function equal to 0:

$$\frac{\partial J}{\partial \mu_k} = 2\sum_{j=1}^{N} r_{jk}\left(x_j - \mu_k\right) = 0 \quad \text{(Equation 9)}$$

$$\mu_k = \frac{\sum_{j=1}^{N} r_{jk} x_j}{\sum_{j=1}^{N} r_{jk}} \quad \text{(Equation 10)}$$

where k is the k_{th} central point.

Semi-supervised learning methodology

Disagreement-based learning

Disagreement-based methods exploit unlabeled sample data through disagreement/diversity among multiple learners, which consists of gradient boosting decision tree, Gaussian naive Bayesian regressor, support vector machine and elastic net (Figure S5). The principle of this method is to train several initial

models based on the same labeled sample data, then let each models select the samples with the highest classification confidence and assign labels to the unlabeled samples. The pseudo-labeled sample data is fed into another classifier to learn.

Gradient boosting decision tree

Gradient Boosting Decision Tree (GBDT) is a common machine learning technique used in regression and classification tasks by establishing an ensemble model (i.e., a linear combination of basic functions) and continuously reducing the residual values generated by the training process.

The model can be described as

$$F_M(x) = \sum_{m=1}^{M} T(x;\theta_m) \qquad \text{(Equation 11)}$$

The model is trained for a total of M rounds, and each round produces a weak learner $T(x;\theta_m)$. The target function of the weak learner is:

$$\theta_m = \operatorname{argmin}\theta_m \sum_{i=1}^{N} L(yi, F_{m-1}(xi) + T(xi;\theta_m)) \qquad \text{(Equation 12)}$$

$F_{m-1}(x)$ is the current model, and GBDT determines the parameters of the next weak learner by minimizing the empirical risk.

Gaussian naive bayesian regressor

There are two steps in Gaussian Naïve Bayesian Regressor Learning Process, which are the inference step and the predicting step.

In inference step, model parameters would be inferred by given data. In linear regression, weight matrix W is learning by using least square method, while in Gaussian Naive Bayesian Learning W is learned by Bayesian equation.

In probability perspective, predicting result is a function of input data and Gaussian noise.

$$Y = \mathbf{W}^T X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \qquad \text{(Equation 13)}$$

$\mathbf{W}$ can be given by

$$\begin{aligned} p(\mathbf{W}|\text{Data}) &= p(\mathbf{W}|\text{X},\text{Y}) = \frac{p(\mathbf{W},\text{Y}|\text{X})}{p(\text{Y}|\text{X})} \\ &= \frac{p(\text{Y}|\mathbf{W},\text{X})\bullet p(\mathbf{W}\text{X})}{\int p(\text{Y}|\mathbf{W},\text{X})\bullet p(\mathbf{W}|\text{X})dW} \end{aligned} \qquad \text{(Equation 14)}$$

Due to independence of $\mathbf{W}$ and X,

$$p(\mathbf{W}|\text{Data}) = \frac{p(\text{Y}|\mathbf{W},\text{X})\bullet p(\mathbf{W})}{\int p(\text{Y}|\mathbf{W},\text{X})\bullet p(\mathbf{W})dW}, \; p(\mathbf{W}) \sim N(0,\Sigma) \qquad \text{(Equation 15)}$$

According to the "naive" assumption and Gaussian assumption, X is independent between its components, where each component confirmed to the Gaussian distribution. As a result, $p(\mathbf{W}|\text{Data})$ should also confirmed to a Gaussian distribution. The likelihood of Y is

$$\begin{aligned} p(\text{Y}|\mathbf{W},\text{X}) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}\left(y_i - W^T x_i\right)^2 \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - W^T x_i\right)^2 \right\} \\ &= \frac{1}{2\pi^{\frac{n}{2}}\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2}(Y - XW)^T \sigma^{-2} I (Y - XW) \right\} \\ & p(\text{Y}|\mathbf{W},\text{X}) \propto N(XW, \sigma^2 I) \end{aligned} \qquad \text{(Equation 16)}$$

where I is identity matrix and N is the number of components of X.

Using Gaussian subsection multiplication method, we achieve the distribution of **W** and the mean value μ_W and variance Σ_W of this distribution can be estimated by:

$$\begin{cases} p(\mathbf{W}|\text{Data}) = N(\mu_W, \Sigma_W) \\ \mu_W = \sigma^{-2} A^{-1} X^T Y \\ \Sigma_W = A^{-1} \end{cases} \quad \text{(Equation 17)}$$

where $A = \sigma^{-2} X^T X + \sum^{-1}$

In predicting step, for a new input data x^*, the prediction value y^* is sampled by

$$y^* = f(x^*) + \varepsilon = N\left(x^{*T}\mu_W,\ x^{*T}\Sigma_W x^* + \sigma^2\right) \quad \text{(Equation 18)}$$

Support vector machine

Support vector machine (SVM) provides a flexible method for both linear and nonlinear predictions. The basic idea of SVM is to maximize the margin between different kinds of data, which is the minimal distance from all the points to the classifying hyperplane. For a hyperplane $y = w^T x + b$, the distance from a point to this hyperplane is $w^T x + b$. Maximizing the margin can be expressed as

$$\begin{aligned} &\max\ margin = \max\ \min\ distance = \max\ \min \frac{1}{\|W\|}\left|W^T X + b\right| \\ &\text{s.t.}\ \ y_i\left(W^T X + b\right) > 0 \end{aligned} \quad \text{(Equation 19)}$$

With scaling of the problem, $y_i(W^T X + b)$ can be larger than an arbitrary nu mber. Set $y_i(W^T X + b) > 1$, the equation goes to

$$\begin{aligned} &\min \frac{1}{2} W^T W \\ &\text{s.t.}\ \ 1 - y_i\left(W^T x_i + b\right) \le 0 \end{aligned} \quad \text{(Equation 20)}$$

Applying Lagrange multiplier method,

$$\begin{aligned} &\min_{w,b} \max_{\lambda} \mathcal{L}\,(w, b, \lambda) \\ &\text{s.t.}\ \ \lambda_i \ge 0 \end{aligned} \quad \text{(Equation 21)}$$

$$\mathcal{L}\,(w, b, \lambda) = \frac{1}{2} W^T W + \sum \left\{ 1 - y_i\left(W^T x_i + b\right) \right\} \quad \text{(Equation 22)}$$

The solution of this equation is untraceable, as a result we turn to solve the dual problem of Lagrange problem.

$$\begin{aligned} &\max_{\lambda} \min_{w,b} \mathcal{L}\,(w, b, \lambda) \\ &\text{s.t.}\ \ \lambda_i \ge 0 \end{aligned} \quad \text{(Equation 23)}$$

It has been proved that the original problem and the dual problem shares the same solution. Using the KTT method, the dual problem can be solved as the following equation.

$$\begin{aligned} &\max_{\lambda} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum \lambda_i \\ &\text{s.t.}\ \ \lambda_i \ge 0, \quad \sum \lambda_i y_i = 0 \end{aligned} \quad \text{(Equation 24)}$$

Elastic net

Elastic Net is a general form of regularized linear regression with an improvement of elastic regularization ratio. Assuming our labeled data input is $\{X, Y\}$. X is the input feature and Y is the target value of each learning module. The number of data is N.

$$X = \{x_1, x_2, \ldots, x_N\}^T \quad \text{(Equation 25)}$$

$$Y = \{y_1, y_2, \ldots, y_N\}^T \quad \text{(Equation 26)}$$

The target function of basic linear regression is

$$\theta = \underset{\theta}{\operatorname{argmin}} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta^T X^{(i)} - Y^{(i)}\right)^2 \quad \text{(Equation 27)}$$

There are two kinds of regularized linear regression, Lasso Regression and Ridge Regression. Lasso Regression add L1 penalty term on the target function, while Ridge Regression add L2 penalty term on the target function.

$$\theta_{\text{Lasso}} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta^T X^{(i)} - Y^{(i)}\right)^2 + \lambda \sum_{j=1}^{N} \|\theta\|^1 \quad \text{(Equation 28)}$$

$$\theta_{\text{Ridge}} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta^T X^{(i)} - Y^{(i)}\right)^2 + \lambda \sum_{j=1}^{N} \|\theta\|^2 \quad \text{(Equation 29)}$$

However, in some cases it is not able to decide which kind of penalty terms contribute more on regularization effect. Elastic Net combines the advantages of both regularized linear regressions, using a ratio to balance the effect of two penalty terms.

$$\theta_{\text{Elastic}} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta^T X^{(i)} - Y^{(i)}\right)^2 + \lambda_1 \sum_{j=1}^{N} \|\theta\|^1 + \lambda_2 \sum_{j=1}^{N} \|\theta\|^2 \quad \text{(Equation 30)}$$

QUANTIFICATION AND STATISTICAL ANALYSIS

Figures represent averaged or representative results of multiple independent machine learning models. Analyses and plots were performed with python.