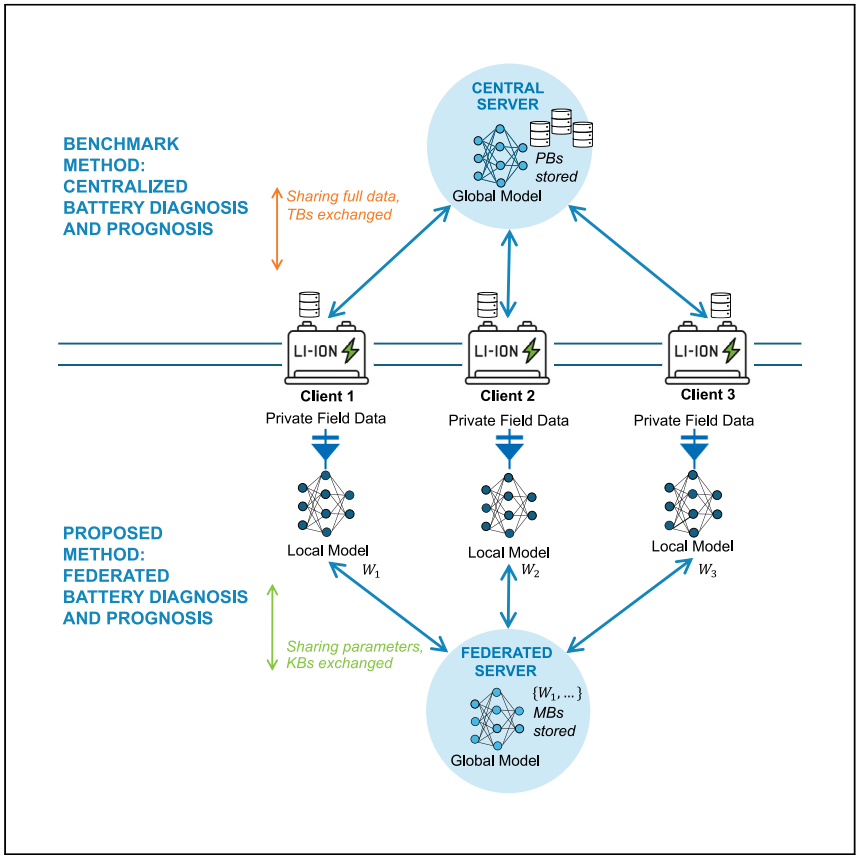
Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Catalyzing deep decarbonization with federated battery diagnosis and prognosis for better data management in energy storage systems



Industrial data analytics and effective asset management are key for catalyzing widespread deployment of energy storage for electrified transportation and renewable energy. Altinpulluk et al. propose a federated battery diagnosis and prognosis model that processes data locally, reduces communication load, and enhances privacy, enabling scalable and secure battery management systems.

Nur Banu Altinpulluk, Deniz Altinpulluk, Paritosh Ramanan, Noah H. Paulson, Feng Qiu, Susan J. Babinec, Murat Yildirim

murat@wayne.edu

Highlights

Data analytics tools in battery management face deployment challenges

Proposed federated model processes battery data locally, enhancing privacy and reducing load

Privacy is enabled by distributed data processing, avoiding centralized data collection

Experiments demonstrate similar predictive and cost performance for the federated model

Altinpulluk et al., Cell Reports Physical Science 5, 102215
October 16, 2024 Published by Elsevier Inc.
https://doi.org/10.1016/j.xcrp.2024.102215

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Catalyzing deep decarbonization with federated battery diagnosis and prognosis for better data management in energy storage systems

Nur Banu Altinpulluk,[1] Deniz Altinpulluk,[1] Paritosh Ramanan,[2] Noah H. Paulson,[3] Feng Qiu,[4] Susan J. Babinec,[3] and Murat Yildirim[1,5,*]

SUMMARY

Industrial data analytics methods play a central role in improving energy storage performance and efficiency, impacting the future of electrified transportation and renewable electricity generation. However, significant challenges hinder the large-scale deployment of batteries. Conventional methods rely on centralized collection and processing of fleet-level data, leading to database size issues and privacy concerns due to potential data breaches. To enable scalable deployment of battery management systems, this article proposes a federated battery diagnosis and prognosis model, which distributes the processing of battery standard current-voltage-time-usage data in a privacy-preserving manner. Instead of transferring the raw data, this approach communicates only the locally processed parameters, thus reducing communication load and preserving data confidentiality. The federated model offers a paradigm shift in battery health management through privacy-preserving distributed methods for battery data processing and lifetime prediction, ensuring the reliable and sustainable deployment of lithium-ion batteries in a rapidly evolving world.

INTRODUCTION

Deeply decarbonized energy sources for transportation and electricity generation are key components in the fight against climate change. Electrified transportation has transitioned from a promising candidate to a globally embraced solution. Wind and solar electricity generation can similarly be cornerstone technologies if their intermittency issues can be adequately addressed. Lithium-ion batteries are the key enablers for both markets, with the grid presenting especially unique challenges as it continues to evolve its design, and thus its storage requirements. In fact, large-scale deployment of lithium-ion batteries has become paramount in forging a sustainable modern grid.[1,2] The magnitude of energy storage investments required for the global green transformation requires a more sophisticated level of performance prediction and monitoring than is presently available to minimize performance and the financial risks associated with their increasing use. The primary challenge for energy storage in general, and especially for lithium-ion, is predicting performance throughout life, since the loss of capacity and growth of resistance is highly path dependent, and the grid deployment scenarios (paths) are quite variable. Investable large-scale lithium-ion deployment is ultimately linked to effective battery health predictions and management strategies. Recently, artificial intelligence (AI) and machine learning (ML) tools have been shown to offer a solution to these incredibly complex problems,[3–5] but continued advancements require

[1]Industrial and Systems Engineering, Wayne State University, Detroit, MI, USA

[2]Industrial Engineering & Management, Oklahoma State University, Stillwater, OK, USA

[3]Argonne Collaborative Center for Energy Storage Science (ACCESS), Argonne National Laboratory, Lemont, IL, USA

[4]Energy Systems Department, Argonne National Laboratory, Lemont, IL, USA

[5]Lead contact

*Correspondence: murat@wayne.edu

https://doi.org/10.1016/j.xcrp.2024.102215

significantly more data than are broadly available. While energy storage performance data are abundant, their availability for data science efforts is scarce due to security concerns and infrastructure demands for transmitting large amounts of data, specifically within the context of field deployments. Removing these two roadblocks will release the critically needed data for the next generation of data science tools to enable large-scale lithium-ion deployments for the modern grid.

Practical solutions to these stringent data reliability, security, and management challenges will unleash the availability of existing data and enable improved tools to catalyze energy storage deployments by reducing performance and investment risks.[6] Traditional battery diagnosis and prognosis approaches use continuous collection and processing of comprehensive current-voltage-time-usage data in a central server. This approach, called centralized diagnosis and prognosis (see related surveys[3–5,7,8]), requires a vast amount of data transfer between batteries in the field and a central database, which congests computation and communication channels, and subjects the operators to significant vulnerabilities in terms of data residency and privacy.

Current methods face three distinct challenges:

(1) Data privacy requirements: Typically, each battery operator, also referred to as a client, collects data locally at the battery site within a local time series database. To support the discovery of fleet-level battery degradation trends, local data from each client need to be collected in a centralized data warehouse. These data warehouses are usually managed and maintained by the original equipment manufacturers (OEMs) of the batteries. Data sharing between clients and the central database raises significant security concerns about and vulnerabilities to data breach incidents.[9] IBM reports that in 2023, the average cost of a data breach in the energy sector reached $4.8 million per incident. Furthermore, when there are competing business interests, even minor data leaks can threaten an industrial organization's competitive edge while unleashing devastating legal and financial consequences.[10]
(2) Input/output (I/O) limitations: Centralized collection of high data volumes from a fleet of batteries increases the need for an ever-increasing capacity for I/O channels, on the order of terabytes (TBs) for regular applications. Trends suggest that there will be a need for an exponentially increasing data size acquired from batteries. In the research and development sector, a leading chemical producer generates over 70 million data points daily for battery characterization, while an academic consortium called Energie R2SE produces an annual output of 1 petabyte (PB) of battery data.[11]
(3) Storage and processing limitations: Centralized processing and storage of data places all the data management burden on a single unit, which significantly increases the load on central servers, requiring PBs of data stored and processed by a single entity, which extrapolates linearly with each individual processor. These fundamental challenges amplify with increasing scale of battery deployments and constitute significant barriers for the industrial implementation of battery diagnosis and prognosis applications.

Another important ecosystem consideration is that these challenges disproportionately impact small to medium enterprises that have limited risk tolerances and significantly lower capacities for data storage, communication, and processing. The present resource paradigm favors the large and discriminates against the small organizations that can often be the source of highly innovative technology solutions.

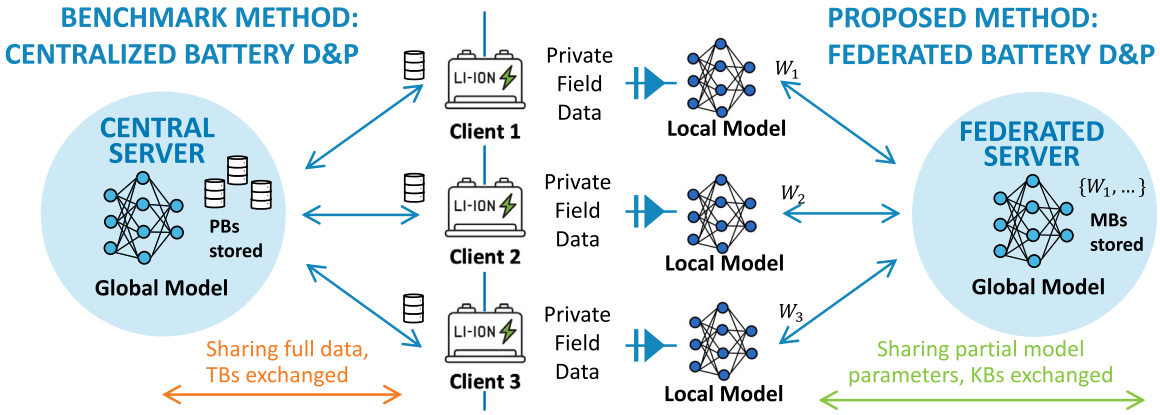


Figure 1. Comparative analysis for the conventional centralized battery diagnosis and prognosis models and the proposed federated battery diagnosis and prognosis models

Centralized models collect private field data from various sources, aggregating and storing it on a central server, which typically amounts to PBs of stored data. Battery diagnosis and prognosis models are also executed within this central server. This mode of operations poses higher risks of data leaks, as clients share their entire datasets for storage and processing in a central server. In contrast, the proposed approach offers an alternative that keeps private field data at their source. Local models for diagnosis and prognosis are executed at the source, sharing only local model parameters with the server. This federated structure mitigates the risk of data breaches and reduces communication load and storage burden on the server.

Federated learning (FL) is a new class of privacy-preserving ML method that has recently appeared in other data science markets with similar challenges. FL methods have gained momentum in recent years[12] with emerging applications such as Google GBoard[13] and have found applications across many domains, including health care,[14,15] cybersecurity,[16] manufacturing,[17,18] and energy.[19,20] FL methods show great promise for both dimensionality reduction[21–26] and prediction stages[27–29] of classical ML problems. For the prediction problem, there has been interesting emerging applications of FL for predicting battery health in recent years,[30–34] where the focus has been on the use of different neural network structures for addressing prediction problems ranging from battery heterogeneity and clustering to enhancing battery recycling applications. In the current approaches, we foresee two main challenges: (1) a lack of integration across federated data dimensionality reduction and federated prediction of remaining life, which would ensure an end-to-end federation, and (2) a need for quantification of the costs associated with adopting different predictive models, specifically while comparing the privacy-preserving approaches with the centralized counterparts. There are additional problems related to different client configurations, which may introduce additional difficulties for analysis. In industrial applications that use large-scale sensor data, both of the dimensionality reduction and prediction stages are required, and hence, they should be jointly federated to ensure data privacy and prediction accuracy of the resulting models.

Federated battery diagnosis and prognosis, therefore, is a potential solution to the central data-scarcity AI/ML roadblock since it builds a framework that focuses on collecting data-derived insights rather than collecting the data itself. In other words, FL data-driven insights can become the new critical asset that offers a unique and broadly useful solution profile, as opposed to standard current-voltage-time-usage data, which is cumbersome and can easily be compromised. A comparison of centralized and federated battery models is shown in Figure 1. The terms KBs, MBs, and TBs refer to kilobytes, megabytes, and terabytes, respectively. A second key feature is that centralized battery diagnosis and prognosis have PB data storage requirements and demand high bandwidth communication channels capable of

streaming data on the order of TBs. In contrast, the federated framework focuses on distributing data processing in a privacy-preserving manner to collect insights, rather than the data itself. Federated diagnosis and prognosis use local standard current-voltage-time-usage data to update local models for each battery (or client) and only collect locally updated model parameters (i.e., insights) within a federated server. This structure in our framework is called an information diode.

This proposed federated diagnosis and prognosis framework has three implementation challenges and potential solutions:

(1) Ensuring data privacy: information diodes are situated between the local data and the local model. As with diodes in electrical applications, these information diodes trap the raw data transfer and only enable the insights to pass from local models that are trained using local raw data, which inherently ensures privacy.
(2) Reducing I/O requirements: I/O requirements are reduced via a two-step process: –(1) processing high-fidelity data locally and building local ML models and (2) communicating only ML model parameters across channels across batteries and the central database. This operational change enables harnessing of the full information contained in the data while reducing the size of the communicated data by orders of magnitude.
(3) Distributing data storage and processing load: the processing and storage of standard current-voltage-time-usage data are distributed, and updating of the central model occurs exclusively via structured model updates, which is computationally efficient and requires minimal storage on the federated server.

Thus, an end-to-end FL framework is proposed for solving the data-scarcity issue, which impedes further progress in the prediction of battery performance before deployment and of remaining life once the battery is in use.

This holistic FL strategy integrates two steps: (1) a federated autoencoder (FA) stage for dimensionality reduction and (2) a federated deep neural network (DNN) step for remaining useful life prediction. During the first step, an FA model is deployed to reduce the dimensionality of feature set to efficiently capture the essential information while minimizing the data transmission overhead. In the second step, the transformed features are fed into a federated DNN model for predicting the remaining useful life of lithium-ion batteries. Collectively, this framework is referred to as the federated holistic battery prognostics and evaluation framework (HOPE-FED). It offers a holistic solution that addresses the unique needs of battery prognosis in an FL setting, providing an effective and practical approach for industry applications that require security and efficiency. The experimental procedures offer details on the development and implementation strategy.

Extensive experiments were conducted to evaluate the performance of the proposed model using two classes of metrics. Prediction accuracy evaluates the success of prediction as a function of deviation between predicted and realized remaining life. Cost quantification metrics demonstrate the economic impact of different prediction models as a function of battery replacement and failure costs. The proposed models are evaluated using extensive battery datasets that incorporate accelerated life testing data for lithium-ion batteries, with different chemistries, including lithium iron phosphate, $0.5Li_2MnO_3.0.5LiNi_{0.375}Mn_{0.375}Co_{0.25}O_2$ (HE5050) and $LiNi_{0.5}Mn_{0.3}Co_{0.2}O_2$ (NMC532).[35,36]

RESULTS AND DISCUSSION

Framework evaluation

The performance of the proposed HOPE-FED framework was evaluated with detailed case study experiments using two public databases: a *Nature Energy* database[35] and a previously published Argonne database[36] (details of both are in Note S3).

In the first case study, the prediction accuracy of the proposed model at different stages of battery life for a range of lithium-ion chemistries was demonstrated. The remainder of the case studies compare the performance of HOPE-FED with state-of-the-art benchmark models. In the first set of benchmarking experiments, the performance of both age-based periodic replacement policy (APRP) and HOPE-FED-based predictive replacement policy was assessed. APRP is a common approach in the maintenance and reliability literature, where the assets are replaced when they reach a predefined age. This age is typically optimized by minimizing the expected maintenance costs (see Note S1 for more details). The HOPE-FED-based predictive replacement policy leverages the prediction information obtained through the fully federated model.

In the second case study, a set of benchmarking experiments were conducted that compared the performance of the proposed HOPE-FED approach with two benchmark models. The first benchmark model executes all dimensionality reduction and remaining useful lifetime (RUL) prediction tasks using a centralized approach, following the traditional paradigm of ML referred to as the fully centralized model. In the second benchmark model, the computational experiments are performed without dimensionality reduction via autoencoder to demonstrate the effect of the autoencoder in the performance. This particular benchmark model is called the pure FeRUL model.

Lastly, in the final benchmark model, a partial application of the FL framework was implemented by partitioning batteries into groups and preserving the privacy of each group, thereby representing the batch-federated approach. The batch-FL scenario represents multiple companies that collaborate with a single analytics provider. In this model, the analytics provider must ensure the privacy of a subset of client data, distinct from others, while processing and aggregating the data. This approach strikes a balance between fully federated and centralized models, allowing for efficient data sharing and analysis with tailored privacy considerations for different client groups. The computational results of batch-federated approach are shared in Note S5.

The performance of HOPE-FED was compared to the benchmark models using two essential success criteria: prediction accuracy and cost quantification (see experimental procedures). Prediction accuracy evaluates the percentage of deviation between the predicted and the actual remaining life. Cost quantification metrics consider the long-run average cost of batteries as a function of replacement decisions. This metric demonstrates the trade-off between early (i.e., premature) replacement and late replacement actions.

In these experiments, the batteries were continuously monitored and the responses were used in a hypothetical use scenario context—when the remaining life predictions fall below a certain threshold (e.g., 2 weeks to failure), a replacement order is placed that takes a certain time to be resolved (e.g., 1 week). Evaluation of the metric is demonstrated in Algorithm S2 (see Note S1).

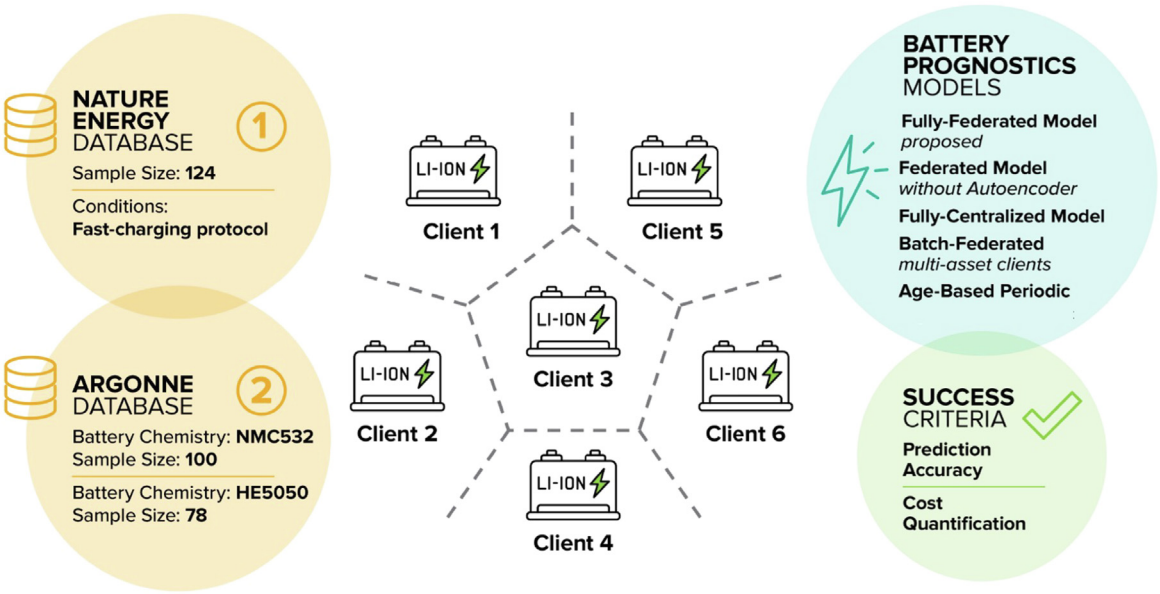


Figure 2. Synopses of computational experiments

The computational experiments were conducted utilizing two publicly available databases: the *Nature Energy* database and the Argonne database. To evaluate the effectiveness of the proposed approach, comparisons were made against different benchmark models, a federated model without autoencoder, a fully centralized model, a batch-federated model, and an age-based periodic approach. Evaluation criteria were based on two key metrics, prediction accuracy and cost quantification, which were used to assess the performance of each approach.

Additionally, insights into the number of early replacements and failures were provided, as well as the average unused life and average number of unavailable days. The average unused life refers to the extent to which battery replacement occurs before failure, indicative of the wasted potential useful life of batteries. Conversely, the average number of unavailable days represents the average duration during which a battery remains inactive due to replacement operations. This calculation considers both the time required for battery replacement and the waiting time for the crew. For a comprehensive understanding of the algorithms used to calculate the average unused life and average number of unavailable days, please refer to Algorithms S4 and S5 in Note S4.

The overview of the computational experiments is depicted in Figure 2. The ability to implement scenario adjustments is implicit in our construct.

Predictive performance of the proposed approach in different chemistries

The HOPE-FED approach was evaluated using datasets generated from the two databases from *Nature Energy* and Argonne. To showcase the effectiveness of the HOPE-FED approach, three distinct sets of results were presented, using the *Nature Energy* database, Argonne database's HE5050 chemistry, and Argonne database's NMC532 chemistry. The prediction error was assessed across three distinct datasets and plotted against different percentiles of battery lifetimes; prediction error vs. the stage of battery lifetime is crucial because remaining lifetime prediction typically is performed continuously. Lifetime percentile refers to the percentage of the battery's total lifetime that has elapsed by a specific point in time. For instance, 50% lifetime means the battery is halfway in its lifetime, 90% lifetime indicates that 10% of the battery's lifetime is remaining until failure. Evidently, lower lifetime percentiles indicate that the battery is in its early stages, similar to a brand-new condition. Conversely, higher lifetime percentiles signify that the battery is approaching its end of life,

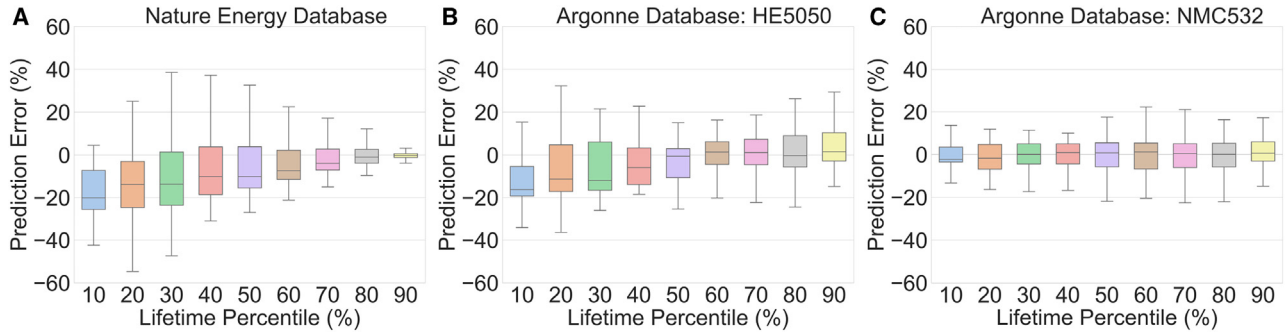


Figure 3. Prediction error values across different lifetime percentiles for the HOPE-FED approach

The plots (A), (B), and (C) illustrate the prediction error percentages across different lifetime percentiles of lithium-ion batteries from the *Nature Energy* and Argonne databases for HE5050 and NMC532 chemistries, respectively. The lifetime percentile of a battery indicates the proportion of its total expected lifespan that has passed at a given point in time. The prediction accuracy is defined as percentage deviation across the predicted and actual lifetimes. In the *Nature Energy* database, prediction accuracy improves as batteries age, which is beneficial for optimizing replacement strategies and balancing costs. However, no such trend is observed in the Argonne databases.

making accurate predictions and alerts more critical for effective replacement scheduling. However, this comes with a caveat. The alert should be close to the failure time to ensure we can use the battery lifetime effectively, but also give ample time for proactive planning of replacement decisions. Evidently, an alert that comes too late is also no longer useful. Therefore, unsuccessful predictions at higher lifetime percentiles can result in inefficiencies in replacement decisions. Figure 3 displays the prediction error values across lifetime percentiles ranging from 10% to 90%. The absolute average prediction errors for the *Nature Energy*, Argonne: HE5050, and Argonne: NMC532 test sets are 11.3%, 13.0%, and 6.9%, respectively, indicating the performance of the HOPE-FED approach.

In the *Nature Energy* database, the absolute average prediction error values consistently decrease as the lifetime percentile increases. This trend indicates that since the battery predictions use a history of observations, the resulting predictions improve as batteries age. Such a circumstance is advantageous for conducting replacement activities efficiently, as it allows for balancing the trade-off between battery replacement cost and the opportunity cost associated with early battery replacements. By accurately predicting the remaining lifetime of batteries nearing failure, decision makers can optimize their replacement strategies and make informed choices regarding replacement. In the case of the Argonne database, no specific trend is observed between the absolute average prediction error and the lifetime percentile. The HOPE-FED approach effectively mitigates the risk of costly failures and supports more informed decision making in replacement operations.

Comparative analysis with APRP

This section compares an optimal replacement policy based on the proposed HOPE-FED approach with a corresponding replacement model based on age-based replacement.

(1) APRP performs the replacement activities for all of the batteries following a pre-specified frequency, without considering any prediction information to determine the replacement trigger time. To establish the threshold age t^* for triggering replacement, the long-run average cost for training set is minimized by following the steps of Algorithm S3 (see Note S1). For each potential

Table 1. Comparison of periodic replacement vs. fully federated prediction-based replacement approaches

	Nature Energy		Argonne: HE5050		Argonne: NMC532	
	APRP	HOPE-FED	APRP	HOPE-FED	APRP	HOPE-FED
No. of preventive	29	30	10	10	18	19
No. of corrective	2	1	9	9	7	6
Unused life	444.5	20.3	262.9	114.2	551.8	95.8
Unavailable days	1.3	1.2	3.4	3.4	2.4	2.2
Average cost, $/day	20.3	12.6	32.5	25.7	26.5	19.1

replacement trigger time t^*, the long-run average costs over all training batteries are calculated. The long-run average costs are computed in the same manner as the long-run average cost calculation in the predictive replacement policy, as described in Algorithm S2 (see Note S1), which include both failure and early replacement costs. The t^* value that minimizes the long-run average cost is selected as the optimal replacement trigger time, and all battery replacement activities for the test datasets are scheduled to occur at time t^*.

(2) In contrast, HOPE-FED-based replacement policy schedules maintenance as a function of the remaining lifetime predictions. More specifically, while the threshold for the APRP policy is based on the age of the component (and hence, not dependent on predictions), the HOPE-FED policy threshold is based on the remaining life predictions from the model. HOPE-FED policy triggers maintenance when the remaining lifetime prediction for a battery reaches a certain threshold. It ensures proactive replacement strategies that find an optimal balance between mitigating the risk of failures and prolonging the lifespan of batteries.

The performance of age-based periodic replacement and predictive replacement policies were compared in Table 1 for the *Nature Energy* database and the Argonne database for chemistries HE5050 and NMC532, respectively. The best-performing thresholds are reported in Table 1, while more detailed results across various threshold levels are provided in Tables S3–S5 in Note S5.

Table 1 presents the replacement trigger time, the number of preventive and corrective replacements, the average unused life, the average number of days unavailable due to replacement activities, and the long-run average costs across various policies and datasets. The best-performing HOPE-FED-based predictive replacement policy demonstrated a long-run average cost improvement of 38%, 21%, and 22% compared to the APRP for the *Nature Energy* database, Argonne database with HE5050 chemistry, and Argonne database with NMC532 chemistries, respectively. For the *Nature Energy* database, the number of failures, average unused life, and average number of unavailable days were significantly reduced compared to the APRP. For the HE5050 chemistry in the Argonne database, the HOPE-FED-based predictive replacement policy exhibited the same number of failures, number of early replacements, and average number of unavailable days as the periodic replacement policy. However, the average unused life improved drastically, indicating that the customized predictions per battery helped in preventing premature replacements and enabled the effective use of equipment lifetime. For NMC532 chemistry, the number of failures increased by 1, whereas average unused life improved significantly, which paves the way for improving the long-run average cost and the associated environmental impact of the batteries.

Table 2. Comparison of performance measures for benchmark policies and HOPE-FED on *Nature Energy* and Argonne databases

	Fully centralized	Pure FeRUL	Fully federated
Nature Energy database			
Optimal threshold	25	25	25
No. of preventive	30	25	30
No. of corrective	1	6	1
Unused life	22.3	71.1	20.3
Unavailable days	1.03	2.0	1.2
Average absolute % error	11.80	13.73	11.31
Average cost, $/day	12.3	16.5	12.6
Argonne Database: HE5050			
Optimal threshold	25	50	25
No. of preventive	11	8	10
No. of corrective	8	11	9
Unused life	96.9	288.0	114.2
Unavailable days	3.1	3.9	3.4
Average absolute % error	14.20	17.36	13.00
Average cost, $/day	24.5	31.9	25.7
Argonne Database: NMC532			
Optimal threshold	25	100	50
No. of preventive	19	14	19
No. of corrective	6	11	6
Unused life	59.8	136.4	95.8
Unavailable days	2.2	3.0	2.2
Average absolute % error	6.25	13.87	6.89
Average cost, $/day	18.8	22.3	19.1

Overall, the HOPE-FED-based predictive replacement outperformed the APRP in all datasets. By leveraging predictions to schedule replacement activities in our hypothetical use scenarios, enhanced long-run average costs and improved performance measures were observed. This approach allowed for the creation of customized replacement plans for each battery by monitoring standard current, voltage, time, and usage information effectively.

Comparative analysis with state-of-the-art predictive models

This section offers a comparative analysis of the proposed approach, HOPE-FED, with two benchmark models, the fully centralized approach and the pure FeRUL approach. Detailed results can be found in Table 2 for the *Nature Energy* database and the Argonne database (HE5050 and NMC532 chemistries). A summary of the prediction error plots across different lifetime percentiles are shown in Figures 4, 5, and 6.

The first benchmark study focused on a fully centralized approach, where both the autoencoder and remaining lifetime estimation were performed in a centralized fashion using traditional ML techniques. In this approach, battery features were aggregated before applying the autoencoder, and the transformed features were used together to train the remaining lifetime estimation DNN. This approach showcases the best-case scenario for prediction purposes, since the predictive model has access to all the raw data from the batteries. However, by collecting the raw data, the approach also risks data privacy and stresses communication and computational capacities. When comparing the average absolute error results, we observe that there is no significant difference between the performance of fully federated HOPE-FED and the fully centralized models. However, there is an

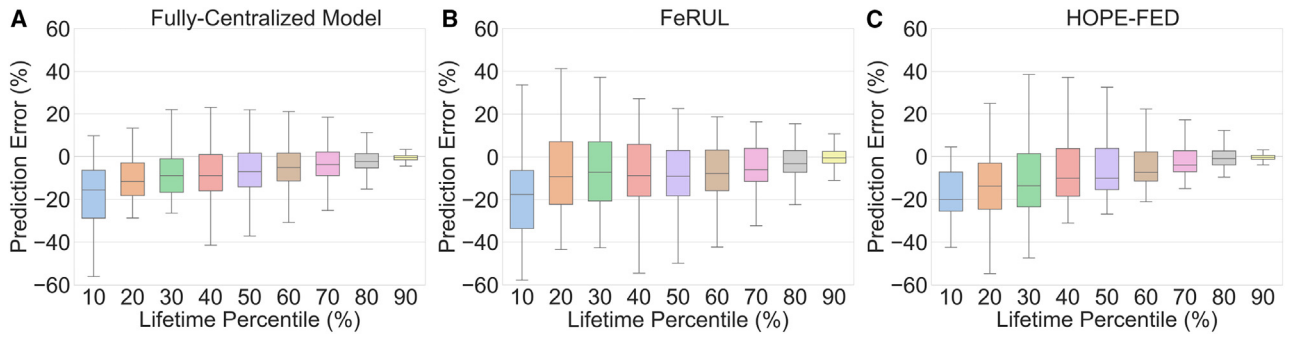


Figure 4. Comparison of benchmark models for the *Nature Energy* database

The performances of fully centralized (A), FeRUL (B), and HOPE-FED (C) approaches are compared using percentage prediction error plots across different lifetime percentiles for the *Nature Energy* database.

inevitable but slight difference between the average cost between fully federated HOPE-FED and the fully centralized models, which constitute the cost of privacy. The results illustrate that this cost is minimal, resulting in only a 2.4% increase in the long-run average cost for the *Nature Energy* database and 4.7% and 1.6% increases for the Argonne database (HE5050 and NMC532 chemistries, respectively).

In the second benchmark study, the pure FeRUL approach, experiments were performed without implementing the autoencoder to assess the effectiveness of the FA. This involved directly applying the FeRUL prediction tasks by embedding all battery features. The results showed that leveraging FAs improved the long-run average cost by 24%, 19%, and 14% for the *Nature Energy* database and the Argonne database (HE5050 and NMC532 chemistries), respectively, while keeping hyperparameters constant in both experimental settings.

The FeRUL approach performed the worst for all the databases, indicating that autoencoders are powerful tools for obtaining high-quality features to improve the prediction performance of ML algorithms. Fully centralized models slightly outperformed fully federated models in terms of long-run average costs for all three datasets, which constitute the cost of privacy. A slight increase in the long-run average costs demonstrates the trade-off for achieving fully preserved privacy and highlights the strength of federated models.

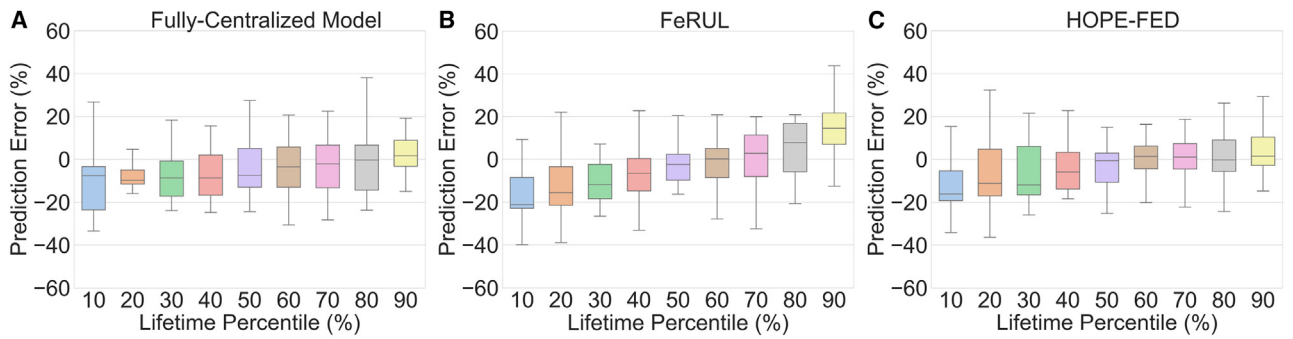


Figure 5. Comparison of benchmark models for the Argonne database: HE5050

The effectiveness of fully centralized (A), FeRUL (B), and HOPE-FED (C) methods are evaluated using percentage prediction error plots across different lifetime percentiles for the Argonne database, HE5050 chemistry.

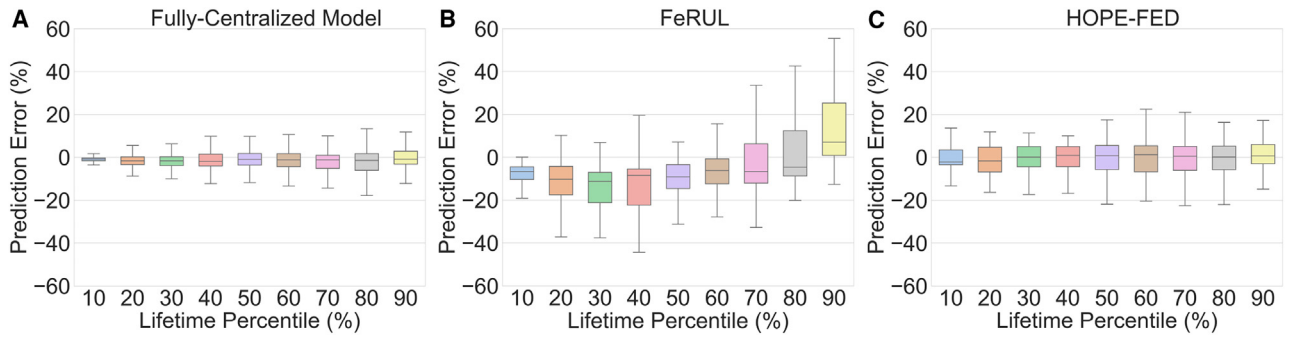


Figure 6. Comparison of benchmark models for the Argonne database: NMC532

The effectiveness of fully centralized (A), FeRUL (B), and HOPE-FED (C) methods are evaluated using percentage prediction error plots across different lifetime percentiles for the Argonne database, NMC532 chemistry.

Lithium-ion batteries are a cornerstone technology for electrified transportation and for fully decarbonized wind and solar energy systems, but they face significant risks associated with an inadequate ability to precisely predict lifetime performance. This is especially critical as the magnitude of investments increases during the "green transition." A more in-depth understanding and effective management of performance can be catalyzed by the broad use of data science tools, especially with respect to the future grid, but lack of data remains a key roadblock. This paper offers a new approach to estimating battery health by considering real-time information, such as standard current, voltage, time, and usage statistics across a fleet of batteries owned and operated by diverse stakeholders.

The proposed approach referred to as HOPE-FED presents a holistic solution for the battery prognosis domain. The novel contributions of this work stem from the use of an FL paradigm to train complex ML models for estimating battery health that eliminates the need to move prognosis-oriented battery data collected by individual stakeholders and operators. The novelty of the proposed approach is further augmented by the development of a federated mechanism for training autoencoders that are ultimately responsible for dimensionality reduction. The use of a federated technique for training autoencoders circumvents the need for centralized data aggregation to carry out the dimensionality reduction step, which is a common attribute in most state-of-the-art FL frameworks. The proposed approach demonstrates an end-to-end, holistic FL mechanism that enables stakeholders to retain full ownership of their data both in terms of learning complex trends influencing battery health and feature dimensionality reduction methods that are essential for robust learning outcomes. The computational results conclusively show that HOPE-FED outperforms conventional periodic maintenance policies in terms of improved costs and reduced downtime. However, the results also demonstrate that HOPE-FED provides costs and downtime performance that are similar to those of the centralized ML versions. These aspects of the proposed approach show that battery stakeholders can expect similar or better prognostic quality as compared to the state of-the-art methods with the added advantage of data privacy and lower data storage and processing latency. Executing the proposed framework with different DNN models could be an interesting future direction that will be investigated by the authors.

From a practical standpoint, HOPE-FED allows for compelling business use cases in the context of lithium-ion batteries enabling collaboration between battery analytics consultants and OEMs. Battery analytics consultants can leverage HOPE-FED to

train a powerful ML model designed to optimize battery performance, efficiency, and longevity. Using the end-to-end, holistic federated approach, HOPE-FED can allow the sharing of insights and models with multiple OEM stakeholders with the promise of retaining complete ownership over their proprietary raw battery data. Consequently, HOPE-FED can ensure that OEMs collectively benefit from a continuously improving, industry-standard model without revealing sensitive data. The practical impact of HOPE-FED is underpinned by real-world use cases that encourage the seamless exchange of knowledge and expertise while promoting cutting-edge battery technologies that respect the autonomy and security of each client's valuable data assets.

In conclusion, the adoption of FL of lithium-ion batteries, especially in the field, represents a paradigm shift, by addressing the fundamental challenges of battery health management: the security and efficiency of data use on a large scale. This game-changing approach ultimately promotes sustainability and efficiency in the electrification era. By harnessing the power of FL, the battery industry can unlock new frontiers of innovation, ensuring the reliable and environmentally responsible deployment of lithium-ion batteries in a rapidly evolving world.

EXPERIMENTAL PROCEDURES

Overview of FL framework for battery prognosis

This section develops the methodology for HOPE-FED, a holistic federated battery prognosis framework geared toward resolving privacy and data acquisition "pain points" for large-scale battery deployments. A significant benefit of the HOPE-FED approach is that it eliminates the need to collect sample datasets for learning the low-dimensional embeddings of the underlying distributed performance data. As a result, HOPE-FED enables the entire data analysis pipeline to be federated, resulting in providing clients with complete ownership of their data. In the following subsections, the two distinct stages of HOPE-FED are described. The first subsection focuses on dimensionality reduction and the second shifts the focus to remaining lifetime prediction. The proposed two-stage architecture offers a modular approach that enables operators to execute these stages at different frequencies, if desired. Owing to their sequential nature, both stages leverage a similar algorithmic structure, along with a shared computational architecture to carry out federation among a diverse set of clients.

FA for dimensionality reduction

The first stage of HOPE-FED involves the design and development of an FA for dimensionality reduction. Compared to traditional techniques such as principal-component analysis, autoencoders present an enticing alternative for capturing nonlinear feature relationships owing to their reliance on feedforward DNNs.[37] Autoencoders consist of an encoder and a decoder, which convert high-dimensional data into low-dimensional encodings and reconstruct the original data using the encodings. By minimizing the reconstruction error, the autoencoder generates accurate encodings that represent the input data with lower dimensions. The ultimate goal of the FA dimensionality reduction is an effort to culminate in a more effective feature extraction. While the proposed federated algorithm significantly reduces the use of communication and computer resources, FAs still improve feature extraction performance and enable better predictions.

In the case of HOPE-FED, the federation of the autoencoder is the initial step that is needed for successfully realizing a federated battery diagnosis and prognosis framework. The training for the FA must be executed asynchronously and in parallel across

a subset of the client pool in an iterative fashion. Each client trains a local autoencoder model by leveraging locally available battery data. Post-completion of local training, the distinct FA models are communicated to federated servers. The federated server aggregates the received FA models to update the globally maintained FA model estimate. The updated copy of the FA model is shared by the central server with the clients to enable the next training round.

The FA component of HOPE-FED empowers clients to collaboratively learn low-dimensional embeddings of their local battery degradation states accurately without moving their local data. As a result, the FA component becomes a vital piece of the overall prognosis framework described by HOPE-FED.

FeRUL prediction

Conventionally, the remaining lifetime of industrial assets can be predicted using a regression-based approach. However, in the case of lithium-ion batteries, regression-based approaches fall severely short owing to several sources of nonlinearities inherently present in battery performance data. Therefore, a more sophisticated ML modeling paradigm is needed to cater to nonlinearities between I/O variables, independence and homoscedasticity of errors, and constant relationships over time.

In the FeRUL step, the training of feedforward DNNs is federated to effectively predict the remaining lifetime for batteries across heterogeneous clients. The reduced set of features obtained from the autoencoder are used as inputs to the DNN, which can be designed with a specific architecture that includes the number of hidden layers, neurons in each layer, and activation functions. The model is then trained using a dataset that includes compressed features and their corresponding remaining lifetime values.

The FeRUL step uses the same computational architecture as FA. A subset of clients is delegated to train a DNN locally on the low-dimensional embeddings that are obtained by leveraging FA on local battery data. Following the local training, the DNN parameters are again communicated to the monitoring service provider (MSP), which is in charge of aggregating the model parameters obtained from each client in the subset. The aggregation step results in the updates to the globally maintained model for predicting remaining lifetime. The MSP concludes the iteration by sending the updated model parameters to each client. It is important to note that the FA and the FeRUL steps are individually critical to the success of HOPE-FED. Each client benefits from a collaboratively trained, low-dimensional, standardized representation on account of the FA step. The output of the FA step is critical for training the FeRUL component accurately.

Federated prognosis algorithm

The federated prognosis algorithm consists of several steps that involve applying the FA and FeRUL prediction sequentially in the FL framework. First, the battery dataset $\{D_1,\ldots,D_{\mathcal{M}}\}$ is randomly split into train $\{D_1,\ldots,D_{\mathcal{K}}\}$ and test sets $\{D_1,\ldots,D_{\mathcal{L}}\}$ and normalize the raw input datasets before initiating the FL framework, and the corresponding target RUL values, $\mathcal{P}_i$, $i = 1,\ldots,\mathcal{M}$, are prepared. Second, the hyperparameters of the FL algorithm, such as the number of rounds for FA and FeRUL prediction, are set. Based on the preliminary experiments, the network parameters for the encoder, decoder, and DNN for RUL prediction, f_θ, f_β, and f_γ, respectively, are determined.

The FA operates in rounds, where each round t involves training a separate model for a random subset of $\mathcal{S}$ batteries from the training set. For each battery s in the subset, the local computation of FA is accomplished by applying the Autoencoder function (see Note S1) to a randomly sampled R% portion of its own data. The trained weights of the encoder and decoder, w_θ^{ts} and w_β^{ts}, respectively, are acquired and subsequently shared with the central node, which aggregates them using the Federated Averaging (FedAvg) function.[38] It is assumed that all participating batteries carry equal weight. The aggregated model weights are then sent back to the batteries, and a new federation round begins. This process continues until all federation rounds are completed. Once finished, the encoder and decoder weights are frozen and are used to transform all training data into a compressed feature representation. With the compressed feature set, FL for remaining lifetime prediction is then initiated using the training data.

The remaining lifetime prediction task begins by sampling batteries and their associated datasets that will contribute to the current federation round. Then, a DNN is trained separately for each sampled battery s by calling the *RUL* function (see Note S1) that locally trains DNN weights for battery s at round t. Once all sampled batteries complete the execution of the RUL function, DNN weights are aggregated using the FedAvg function until the convergence criteria are satisfied. The federated prognosis weight aggregation process is summarized in Figure 7. After completing the training for FeRUL prediction, the proposed approach is evaluated on the test data. First, the test data are fed into the encoder to obtain transformed features $\mathcal{C}$ for every test battery. Second, the DNN, f_γ, is called for every battery l to achieve RUL predictions $\widehat{\mathcal{P}}_l$. Finally, the performance measures that are shared in the next section are reported. Algorithm S1 (see Note S1) outlines the steps of the federated prognosis algorithm used in the FL process.

In industrial deployments, the HOPE-FED model may need to be retrained periodically to ensure that the proposed framework adapts to the changing trends in the data. To address this important aspect, deployment strategies, designed to sustain the performance and stability of the HOPE-FED framework (see Note S2) are proposed.

Performance measures

The evaluation metric mechanism evaluates the performance in terms of both the prediction accuracy and cost quantification. In this regard, two fundamental evaluation metrics are established: prediction error and long-run average cost. The predictive error depends on the analysis of the difference between the predicted and actual remaining lifetime.[39] The second metric, the long-run average cost,[40] incorporates the average cost incurred per battery per period (in US dollars per day), including both the battery's initial cost and any associated replacement expenses. In the following sections, the details of the calculations for the prediction error and long-run average cost are shared.

Evaluating the predictive accuracy

To quantify the prediction error, the remaining lifetime of each battery is predicted at multiple times throughout its lifespan.[39] Then, the prediction error, E_i^k, for battery i at time period k is calculated using the following formula:

$$E_i^k = \frac{\left(p_i^k + t_i^k\right) - t_{fi}}{t_{fi}} \qquad \text{(Equation 1)}$$

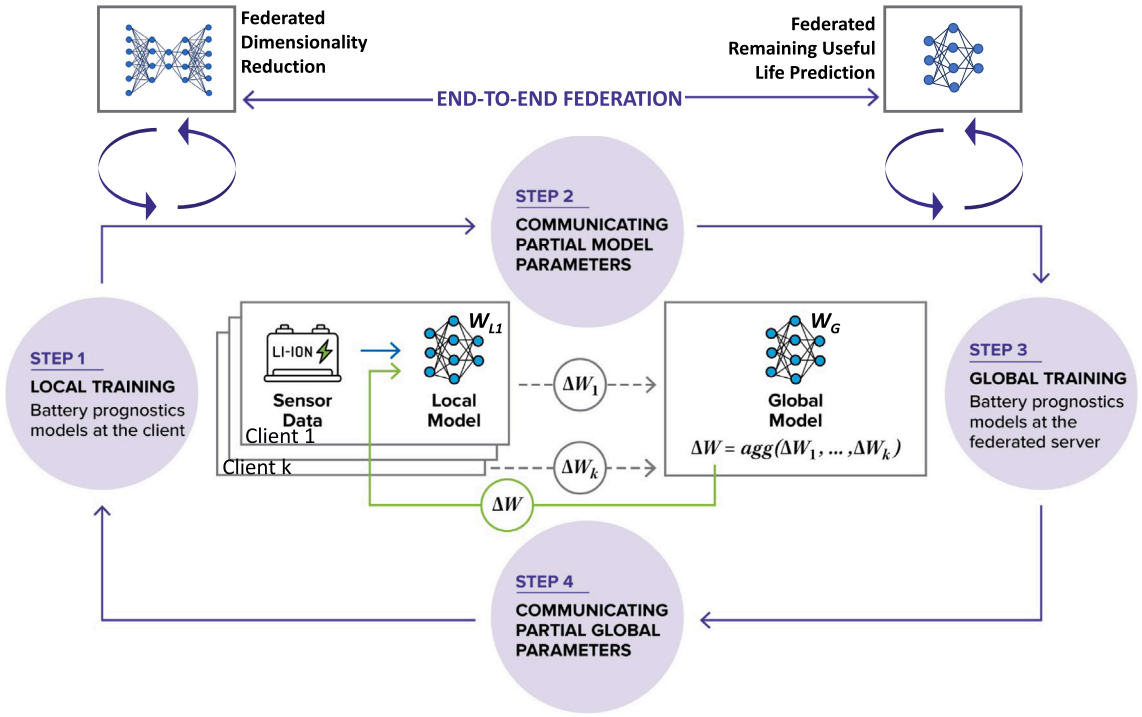


Figure 7. Procedural stages of FL for battery prognosis

Battery prognostic models are constructed on local servers. Following local model training, the neural network weights specific to each local server are transmitted to the central server. These collected weights are then aggregated and processed at the central server before being redistributed to the local servers, thus initiating another round of FL. This cycle is executed for the federated dimensionality reduction and FeRUL prediction stages, to enable end-to-end federation for battery diagnostics and prognostics.

where p_i^k, t_i^k, and t_{fi} refer to the remaining lifetime prediction for battery i at time period k, its current age at time period k, and its failure time, respectively. The E_i^k measure indicates the deviation of the remaining lifetime prediction from the true failure time, while also considering the current age of the battery. This measure encompasses both positive and negative values, indicating instances of overestimation or underestimation of the failure time, respectively. In an industrial context, particularly in applications like electric vehicles, overestimating the battery's lifespan can have undesired consequences, leading to unexpected failures without warning, and resulting in costly battery replacements. Conversely, premature battery replacements can burden the client with unnecessary expenses, which is unsustainable from an economic standpoint. Hence, the following cost quantification analysis incorporates the calculation of long-run average costs associated with both overestimation and underestimation of lifetime.

Quantifying the cost

The cost quantification is evaluated using the long-run average cost metric, which provides a comprehensive assessment of the overall expected cost accrued per battery during each time period. This assessment considers not only the initial battery cost but also the significant impact of replacement expenses. In evaluating the effectiveness of the prediction policies, it is important to recognize the pivotal role played by replacement costs. Furthermore, the magnitude of these replacement costs is intricately tied to the timing of replacements (e.g., replacement before or after the battery failure). Therefore, the prediction of the

remaining lifetime presents a valuable opportunity for devising efficient battery replacement strategies, optimizing resource allocation, and enhancing overall system reliability.

The long-run average cost metric is computed by considering distinct cost values for the early replacement case c_r and corrective replacement case c_f. In the computational experiments, a structured replacement policy is considered that replaces the battery when the remaining lifetime prediction reaches a prespecified threshold value, δ. Let us assume this event occurs at time t^*, and the service crew arrives after t_c time periods. If the battery is still functional upon its arrival, then early replacement is performed, effectively preventing failures. However, in scenarios where the battery fails before the remaining lifetime prediction reaches the threshold value, δ, or while the service crew is en route, failures arise, resulting in significantly higher replacement costs.

In all the benchmarking studies, premature battery replacement is considered undesirable as it undermines the optimal utilization time of the batteries. Hence, the calculation of the long-run average cost per battery, considering its lifespan and replacement time, is employed to ensure optimal utilization of battery life while minimizing replacement costs. Algorithm S2 (see Note S1) provides a detailed outline of the steps involved in computing the long-run average cost per battery when implementing predictive replacement policies. This approach enables effective decision making in selecting appropriate battery replacement policy.

RESOURCE AVAILABILITY

Lead contact

Additional details and inquiries regarding data and code should be directed to the lead contact, Murat Yildirim (murat@wayne.edu).

Materials availability

This study did not generate new materials.

Data and code availability

The datasets employed in this study, namely the *Nature Energy* database and the Argonne database, are publicly available. The *Nature Energy* database can be obtained from https://data.matr.io/1/projects/5c48dd2bc625d700019f3204, and comprehensive data explanations can be found in Severson et al.[35] Similarly, the Argonne database can be downloaded from https://acdc.alcf.anl.gov/mdf/detail/camp_2023_v3.5/, with data details provided in Paulson et al.[36]

ACKNOWLEDGMENTS

M.Y., N.B.A., and D.A. disclose support for the research in this work from the National Science Foundation, award number 2104455. The contribution of the Argonne National Laboratory is based upon work supported by Laboratory Directed Research and Development funding from Argonne National Laboratory, provided by the Director, Office of Science, of the US Department of Energy under Contract No. DE-AC02-06CH11357.

AUTHOR CONTRIBUTIONS

Methodology, N.B.A., D.A., P.R., N.H.P., F.Q., S.J.B., and M.Y. Computational experiments, N.B.A. and D.A. Writing, N.B.A. and D.A. Writing & revising, P.R., N.H.P., F.Q., S.J.B., and M.Y.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrp.2024.102215.

Received: February 21, 2024
Revised: June 25, 2024
Accepted: August 29, 2024
Published: September 19, 2024

REFERENCES

1. Kawamura, H., LaFleur, M., Iversen, K., and Cheng, H.W.J. (2023). United Nations Department of Economic and Social Affairs, Frontier Technology Issues: Lithium-ion Batteries - A Pillar for a Fossil Fuel-Free Economy. United Nations. Accessed on October 2, 2023. https://www.un.org/development/desa/dpad/publication/frontier-technology-issues-lithium-ion-batteries-a-pillar-for-a-fossil-fuel-free-economy/.

2. Ward, L., Babinec, S., Dufek, E.J., Howey, D.A., Viswanathan, V., Aykol, M., Beck, D.A., Blaiszik, B., Chen, B.-R., Crabtree, G., et al. (2022). Principles of the battery data genome. Joule *6*, 2253–2271.

3. Hu, X., Xu, L., Lin, X., and Pecht, M. (2020). Battery lifetime prognostics. Joule *4*, 310–346.

4. Li, Y., Liu, K., Foley, A.M., Zülke, A., Berecibar, M., Nanini-Maury, E., Van Mierlo, J., and Hoster, H.E. (2019). Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review. Renew. Sustain. Energy Rev. *113*, 109254.

5. Lipu, M.H., Hannan, M., Hussain, A., Hoque, M., Ker, P.J., Saad, M., and Ayob, A. (2018). A review of state of health and remaining useful life estimation methods for lithium-ion battery in electric vehicles: Challenges and recommendations. J. Clean. Prod. *205*, 115–133.

6. U.S. Department of Energy (2019). Spotlight: Solving Energy Challenges in Energy Storage. Accessed on October 2, 2023. https://www.energy.gov/sites/prod/files/2019/07/f64/2018-OTT-Energy-Storage-Spotlight.pdf.

7. Meng, H., and Li, Y.-F. (2019). A review on prognostics and health management (PHM) methods of lithium-ion batteries. Renew. Sustain. Energy Rev. *116*, 109405.

8. 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.

9. Kaissis, G.A., Makowski, M.R., Rückert, D., and Braren, R.F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nat. Mach. Intell. *2*, 305–311.

10. Cheng, L., Liu, F., and Yao, D.D. (2017). Enterprise data breach: causes, challenges, prevention, and future directions. WIREs Data Min. &. Knowl. *7*, e1211.

11. Lombardo, T., Duquesnoy, M., El-Bouysidy, H., Årén, F., Gallo-Bueno, A., Jørgensen, P.B., Bhowmik, A., Demortière, A., Ayerbe, E., Alcaide, F., et al. (2022). Artificial intelligence applied to battery research: hype or reality? Chem. Rev. *122*, 10899–10969.

12. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Nitin Bhagoji, A., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. FNT. Mach. Learn. *14*, 1–210.

13. Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. (2018). Federated learning for mobile keyboard prediction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1811.03604.

14. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. NPJ Digit. Med. *3*, 119.

15. Pfitzner, B., Steckhan, N., and Arnrich, B. (2021). Federated learning in a medical context: a systematic literature review. ACM Trans. Internet Technol. *21*, 1–31.

16. Huong, T.T., Bac, T.P., Long, D.M., Luong, T.D., Dan, N.M., Quang, L.A., Cong, L.T., Thang, B.D., and Tran, K.P. (2021). Detecting cyberattacks using anomaly detection in industrial control systems: A Federated Learning approach. Comput. Ind. *132*, 103509.

17. Ge, N., Li, G., Zhang, L., and Liu, Y. (2021). Failure prediction in production line based on federated learning: an empirical study. J. Intell. Manuf. *33*, 2277–2294.

18. Mehta, M., and Shao, C. (2022). Federated learning-based semantic segmentation for pixel-wise defect detection in additive manufacturing. J. Manuf. Syst. *64*, 197–210.

19. Saputra, Y.M., Hoang, D.T., Nguyen, D.N., Dutkiewicz, E., Mueck, M.D., and Srikanteswara, S. (2019). Energy demand prediction with federated learning for electric vehicle networks. In 2019 IEEE Global Communications Conference (GLOBECOM) (IEEE), pp. 1–6.

20. Savi, M., and Olivadese, F. (2021). Short-term energy consumption forecasting at the edge: A federated learning approach. IEEE Access *9*, 95949–95969.

21. Grammenos, A., Mendoza Smith, R., Crowcroft, J., and Mascolo, C. (2020). Federated principal component analysis. Adv. Neural Inf. Process. Syst. *33*, 6453–6464.

22. Novoa-Paradela, D., Romero-Fontenla, O., and Guijarro-Berdiñas, B. (2022). Fast Deep Autoencoder for Federated learning. Patter. Recog. *143*, 109805.

23. Banerjee, S., Elmroth, E., and Bhuyan, M. (2021). Fed-FiS: a Novel Information-Theoretic Federated Feature Selection for Learning Stability. In International Conference on Neural Information Processing (Springer), pp. 480–487.

24. Wang, L., Pang, Q., Wang, S., and Song, D.. Secure Federated Feature Selection. https://aaai-ppai22.github.io/files/3.pdf. n.d.

25. Hu, Y., Zhang, Y., Gong, D., and Sun, X. (2022). Multi-Participant Federated Feature Selection Algorithm with Particle Swarm Optimizaiton for Imbalanced Data under Privacy Protection. IEEE Trans. Artif. Intell. *4*, 1002–1016.

26. Gao, Y., Zhang, G., Zhang, C., Wang, J., Yang, L.T., and Zhao, Y. (2021). Federated tensor decomposition-based feature extraction approach for industrial IoT. IEEE Trans. Industr. Inform. *17*, 8541–8549.

27. Wang, Y., Bennani, I.L., Liu, X., Sun, M., and Zhou, Y. (2021). Electricity consumer characteristics identification: A federated learning approach. IEEE Trans. Smart Grid *12*, 3637–3647.

28. Cassará, P., Gotta, A., and Valerio, L. (2022). Federated feature selection for cyber-physical systems of systems. IEEE Trans. Veh. Technol. *71*, 9937–9950.

29. Xiao, Z., Xu, X., Xing, H., Song, F., Wang, X., and Zhao, B. (2021). A federated learning system with enhanced feature extraction for human activity recognition. Knowl. Base Syst. *229*, 107338.

30. Arunan, A., Qin, Y., Li, X., and Yuen, C. (2023). A federated learning-based industrial health prognostics for heterogeneous edge devices using matched feature extraction. IEEE Trans. Autom. Sci. Eng. *21*, 3065–3079.

31. Wong, K.L., Tse, R., Tang, S.-K., and Pau, G. (2024). Decentralized Deep Learning Approach for Lithium-Ion Batteries State of Health Forecasting Using Federated Learning. IEEE Trans. Transp. Electrif. *PP*, 1.

32. Xiao, F., and Wu, L. (2023). Personalized Federated Lithium-ion Battery Capacity Prediction via Cluster and Fusion Modules. IEEE Trans. Transp. Electrif. https://doi.org/10.1109/TTE.2023.3334826.

33. Kröger, T., Belnarsch, A., Bilfinger, P., Ratzke, W., and Lienkamp, M. (2023). Collaborative training of deep neural networks for the lithium-ion battery aging prediction with federated learning. eTransportation *18*, 100294.

34. Tao, S., Liu, H., Sun, C., Ji, H., Ji, G., Han, Z., Gao, R., Ma, J., Ma, R., Chen, Y., et al. (2023).

Collaborative and privacy-preserving retired battery sorting for profitable direct recycling via federated machine learning. Nat. Commun. *14*, 8032.

35. 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.

36. Paulson, N.H., Kubal, J., Ward, L., Saxena, S., Lu, W., and Babinec, S.J. (2022). Feature engineering for machine learning enabled early prediction of battery lifetime. J. Power Sources *527*, 231127.

37. Sakurada, M., and Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. In Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, pp. 4–11.

38. McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (PMLR), pp. 1273–1282.

39. Gebraeel, N. (2006). Sensory-updated residual life distributions for components with exponential degradation patterns. IEEE Trans. Autom. Sci. Eng. *3*, 382–393.

40. Elwany, A.H., and Gebraeel, N.Z. (2008). Sensor-driven prognostic models for equipment replacement and spare parts inventory. IIE Trans. *40*, 629–639.

Cell Reports Physical Science, Volume *5*

Supplemental information

Catalyzing deep decarbonization with federated battery diagnosis and prognosis for better data management in energy storage systems

Nur Banu Altinpulluk, Deniz Altinpulluk, Paritosh Ramanan, Noah H. Paulson, Feng Qiu, Susan J. Babinec, and Murat Yildirim

SUPPLEMENTARY EXPERIMENTAL INFORMATION

Note S1: HOPE-FED Algorithmic Structure

The HOPE-FED framework consists of two main stages: federated autoencoder (FA) and federated RUL (FeRUL) estimation. To achieve these tasks, the *Autoencoder* and *RUL* functions were developed which are integral components of the federated battery prognosis framework. The *Autoencoder* function summarizes the steps involved in the dimensionality reduction task and is employed within the Federated Battery Prognosis Algorithm. Each sampled battery executes its own autoencoder function, ensuring the preservation of the federated learning structure. It is important to note that the autoencoder structure and hyperparameters remain consistent across all sampled batteries. Similarly, the *RUL function* is individually invoked for each sampled battery within the federated battery prognosis framework to train the deep neural network for predicting the remaining lifetime of lithium-ion batteries. In the next sections, details of federated autoencoder, federated remaining lifetime prediction, and main algorithm of HOPE-FED are shared.

Federated Autoencoder (FA) Mechanism

The autoencoder is trained to reduce the feature dimension of the standard current-voltage-time-usage data, aiming to alleviate the computational burden in the downstream FeRUL task. During training, the autoencoder learns to encode the input data in a way that minimizes the reconstruction error. After training, the encoder can be used to obtain the lower-dimensional representation of new input data, which can be utilized for various downstream tasks such as clustering, classification, or visualization. To evaluate the effectiveness of the autoencoder for dimensionality reduction, the reconstruction error of the decoder is assessed, which provides a measure of how accurately the autoencoder can reconstruct the original input from the lower-dimensional encoding.

The autoencoder is trained using the Adam optimizer with a mean-squared error (MSE) loss function, with the batch size and the number of epochs set. As the autoencoder learns, it encodes the input data in a way that minimizes the MSE loss between the reconstructed input and the original input. To optimize the weights of the network, the error is backpropagated through the network and the weights are updated using the gradients computed by the optimizer. The *Autoencoder* function summarizes the steps performed during autoencoder training and takes five different inputs, including the network parameters for the encoder f_θ and decoder f_β, the network weights of encoder w_θ and decoder w_β, and dataset for autoencoder training represented with D. The network parameters for the encoder f_θ and decoder f_β include the structure of the respective networks, such as the number of layers, types of activation function, and the number of neural units in each layer. Additionally, the target feature size, which becomes the neuron size of the last layer of the encoder for reducing the input dimension to the desired feature size, is also included in the network parameters. In each epoch and batch of the training, the reconstruction loss value is calculated, and the encoder and decoder network weights, w_θ and w_β, are optimized using the Adam optimizer to improve the model's performance.

function AUTOENCODER($f_\theta, f_\beta, w_\theta, w_\beta, \mathcal{D}$)

Set number of epochs to $\mathcal{E}$

for epoch=1,2,..., $\mathcal{E}$

apply ENCODER:

Compress $\mathcal{D}$ to $\mathcal{C}$ by leveraging f_θ

$\mathcal{C} \leftarrow f_\theta(\mathcal{D}; w_\theta)$

apply DECODER:

Reconstruct $\mathcal{C}$ by leveraging f_β to obtain $\hat{\mathcal{D}}$

$\hat{\mathcal{D}} \leftarrow f_\beta(\mathcal{C}; w_\beta)$

Compute loss function:

$$L(w_\beta) = \frac{1}{m}\sum_{i=1}^{m}(\mathcal{D}_i - f_\beta(\mathcal{C}_i; w_\beta))^2$$

Update the weights w_θ and w_β

end for

return w_θ, w_β

Federated RUL (FeRUL) Prediction Mechanism

The *RUL* function is designed to train the DNN for predicting the remaining useful lifetime of lithium-ion batteries. This function takes four different inputs: network parameters f_γ, network weights w_γ, compressed battery features, $\mathcal{C}$, and target variables for prediction (actual RUL values), $\mathcal{P}$. Initially, the number of epochs and batch size are set initially, and for each epoch and batch, the input data is fed to the DNN. The loss values are then calculated between predicted RUL values, $\hat{\mathcal{P}}$, and actual RUL values, $\mathcal{P}$. Based on the loss values, the DNN weights, w_γ, are optimized.

function RUL($f_\gamma, w_\gamma, \mathcal{C}, \mathcal{P}$)

Set number of epochs to $\mathcal{E}$

for epoch=1,2,..., $\mathcal{E}$

Feed input data C to the DNN

$\hat{\mathcal{P}} \leftarrow f_\gamma(\mathcal{C}; w_\gamma)$

Compute loss function:

$$L(w_\gamma) = \frac{1}{m}\sum_{i=1}^{m}(\mathcal{P}_i - f_\gamma(\mathcal{C}_i; w_\gamma))^2$$

Update the network weights w_γ by calling Adam optimizer

end for

return w_γ

Network Structures of HOPE-FED Framework

Autoencoder is a type of neural network designed to learn efficient codings of input data. The structure begins with defining the input shape of the network which takes an input feature vector. The encoder part of the autoencoder then processes this input through two connected dense layers. The first dense layer reduces the dimensionality to 64 units with a hyperbolic tangent (tanh) activation function, and the second dense layer further reduces it to target feature size (30 for Nature Energy database and 40 for Argonne database), also using the tanh activation function. This bottleneck layer represents the compressed encoding of the input data. Following this, the decoder part of the network reconstructs the input by passing the encoded data through another dense layer with 64 units and tanh activation, and finally, through a dense layer with original feature size units and a rectified linear unit (ReLU) activation function. The output of this layer aims to replicate the original input data. The complete autoencoder model is constructed by connecting the input layer to the final decoded output.

Following this, the federated RUL framework takes the encoded features produced by an autoencoder as its input. The neural network structure for RUL predictions begins with a dense layer consisting of 512 neurons with a ReLU activation function, which takes input features of dimension target feature size of associated database. This is followed by a series of dense layers with decreasing neuron counts: 256, 128, 64, 32, and 16, all using the ReLU activation function. Finally, the network concludes with a single neuron in the output layer, also with a ReLU activation function.

HOPE-FED Algorithm

The *Autoencoder* and *RUL*, as introduced in the preceding sections play a pivotal role in HOPE-FED framework. They are responsible for executing local training for the Autoencoder and DNN. In the structural framework of HOPE-FED, the process unfolds in two distinct stages. Initially, the FA undergoes distributed training, utilizing federated averaging to update local weights over a predetermined number of federation rounds. Following this, in the second stage, the FeRUL component is trained using the compressed representations of the current-voltage-time-usage data derived from FA during the predetermined number of federation rounds. To provide a comprehensive overview, Algorithm S1, *Federated Prognosis Algorithm*, outlines the federated prognosis steps within the HOPE-FED framework. This algorithm sequentially applies FA and FeRUL.

Algorithm S1: Federated Prognosis Algorithm

Notation:
$\mathcal{M}$: set of batteries
$\mathcal{K}$: set of training batteries
$\mathcal{L}$: set of test batteries
$\mathcal{S}$: number of sampled batteries at each round
$\mathcal{R}$: sampling ratio of data points at each round
$\mathcal{D}_{\mathsf{m}}$: input dataset for battery m, $\mathsf{m} \in \{1, 2, .., \mathcal{M}\}$
$\mathcal{P}_{\mathsf{m}}$: target dataset for battery m, $\mathsf{m} \in \{1, 2, .., \mathcal{M}\}$ which consist of actual values of RUL
$\mathcal{T}_{\mathsf{auteoencoder}}$: number of rounds for federated autoencoder
$\mathcal{T}_{\mathsf{RUL}}$: number of rounds for federated RUL prediction
f_θ: neural network function for encoder
f_β: neural network function for decoder
f_γ: neural network function for RUL prediction
w_θ: network weights of encoder
w_β: network weights of decoder
w_γ: network weights of RUL prediction
for each round t = $1, 2, .., \mathcal{T}_{\mathsf{autoencoder}}$ **do**:
  Randomly select $\mathcal{S}$ batteries from training set $\mathcal{K}$
  for each selected battery s = $1, 2, .., \mathcal{S}$, in parallel **do**:
    Sample from dataset $\mathcal{D}_{\mathsf{s}}$ with $\mathcal{R}\%$ and obtain $\mathcal{D}_{\mathsf{s}'}$
      $\mathsf{w}_\theta^{\mathsf{ts}}, \mathsf{w}_\beta^{\mathsf{ts}} \leftarrow$ **AUTOENCODER**$(\mathsf{f}_\theta, \mathsf{f}_\beta, \mathsf{w}_\theta^{\mathsf{ts}}, \mathsf{w}_\beta^{\mathsf{ts}}, \mathcal{D}_{\mathsf{s}'})$
  end for
  Apply federated averaging:
      $\mathsf{w}_\theta^{\mathsf{t+1}} \leftarrow \frac{1}{\|\mathcal{S}\|} \sum_{\mathsf{s}=1}^{\mathcal{S}} \mathsf{w}_\theta^{\mathsf{ts}}$
      $\mathsf{w}_\beta^{\mathsf{t+1}} \leftarrow \frac{1}{\|\mathcal{S}\|} \sum_{\mathsf{s}=1}^{\mathcal{S}} \mathsf{w}_\beta^{\mathsf{ts}}$
end for
Freeze encoder and decoder weights, w_θ and w_β
Transform train inputs using encoder:
for each train battery k = $1, 2, .., \mathcal{K}$ in parallel **do**:
  $\mathcal{C}_{\mathsf{k}} \leftarrow \mathsf{f}_\theta(\mathcal{D}_{\mathsf{k}}; \mathsf{w}_\theta)$
end for
for each round t = $1, 2, .., \mathcal{T}_{\mathsf{RUL}}$ in parallel **do**:
  Randomly select $\mathcal{S}$ batteries from training set $\mathcal{K}$
  Sample from dataset $\mathcal{C}_{\mathsf{s}}$ with $\mathcal{R}\%$ and obtain $\mathcal{C}_{\mathsf{s}'}$
  for each selected battery s = $1, 2, .., \mathcal{S}$, in parallel **do**:
      $\mathsf{w}_\gamma^{\mathsf{ts}} \leftarrow$ **RUL**$(\mathsf{f}_\gamma, \mathsf{w}_\gamma^{\mathsf{ts}}, \mathcal{C}_{\mathsf{s}'}, \mathcal{P}_{\mathsf{s}'})$
  end for

Apply federated averaging:

$$w_{\gamma}^{t+1} \leftarrow \frac{1}{\|\mathcal{S}\|} \sum_{s=1}^{\mathcal{S}} w_{\gamma}^{ts}$$

end for

Freeze RUL prediction weights, w_{γ}

for each test battery $l = 1, 2, .., \mathcal{L}$ **do**:

Transform inputs $\mathcal{D}_l$ using encoder:

$\mathcal{C}_l \leftarrow f_{\theta}(\mathcal{D}_l; w_{\theta})$

Make RUL predictions:

$\hat{\mathcal{P}}_l \leftarrow f_{\gamma}(\mathcal{C}_l, w_{\gamma})$

Report performance measures for the test set D_l

end for

Long-Run Average Cost Calculation Framework

Algorithm S2 provides a comprehensive overview of the long-run average cost calculation for each battery, enabling the cost quantification for prediction efficacy to facilitate comparisons between benchmark studies and the HOPE-FED approach.

Algorithm S2: Cost Calculation Algorithm

Notation:

$\mathbf{t_{fi}}$: failure time of battery i, $i \in 1, 2, .., \mathcal{M}$

$\mathbf{t_c}$: time for service crew to initiate a battery replacement operation after a replacement is triggered

δ: threshold RUL level for triggering a replacement operation before a failure happens

$\mathbf{t_i^*}$: age of battery i, $i \in 1, 2, .., \mathcal{M}$, when RUL prediction hits under δ for the first time

$\mathbf{c_r}$: the cost of replacing a battery before it experiences a failure

$\mathbf{c_f}$: the cost of replacing a battery after it experiences a failure

$\mathbf{C_i}$: long-run average cost per battery i

for each battery $i = 1, 2, ..., \mathcal{M}$ **do**:

Calculate long-run average cost of prediction C_i:

if $t_i^* + t_c < t_{fi}$:

$$C_i = \frac{c_r}{t_i^* + t_c}$$

else:

$$C_i = \frac{c_f}{t_{fi}}$$

end for

Long-Run Average Cost Calculation for Age-based Periodic Replacement Policy (APRP)

Algorithm S3 introduces the long-run average cost calculation for APRP, which plays a crucial role in determining the optimal time period to initiate replacement activities. By minimizing the long-run average cost among the available candidate time periods, this algorithm aids in the selection of the most effective replacement triggering point for APRP.

Algorithm S3: Benchmark Cost Calculation

for each time period $t^* = 1, 2, ..., \mathcal{T}$ **do**:
  for each battery $i^* = 1, 2, ..., \mathcal{K}$ in the training set **do**:
    Calculate long-run average cost of prediction C_{it^*}:
      if $t^* + t_c < t_{fi}$:
        $C_{it^*} = \frac{c_r}{t^* + t_c}$
      else:
        $C_{it^*} = \frac{c_f}{t_{fi}}$
  end for
end for
Select t^* value that minimizes $\sum_i C_{it^*}$

Note S2: Deployment Strategies for HOPE-FED

Deployment of federated learning algorithms is a critical aspect for ensuring their continued performance and effectiveness. As data changes over time, federated learning models may become outdated and lose accuracy, which necessitates regular updates to the algorithms. This process typically involves monitoring the model's performance, identifying, and addressing any issues that may arise, and retraining the model with updated datasets.

To maintain the efficacy of HOPE-FED, it is essential to establish a set of rules for triggering retraining. To this end, the metric called long-run average cost per battery measure (as explained in Algorithm S2 in Note S1, is used as a key indicator of performance to monitor for retraining. An increase in long-run average costs signifies potential issues with the algorithm's performance. A threshold value, α, can be set as a limit for deviation in the long-run average cost; whereby, when the long-run average cost exceeds α, it serves as an alert for the decision-maker, prompting the retraining of the model to enhance its performance. By implementing this approach, it is ensured that HOPE-FED remains a reliable tool for accurately predicting the remaining lifetime of lithium-ion batteries, consistently delivering valuable insights for replacement strategies.

Note S3: Data Preparation

To implement the HOPE-FED framework, two datasets, namely the Nature Energy and Argonne databases, were utilized. The feature generation and data preprocessing procedures utilized for these datasets is explained under this section.

Nature Energy Database

In the computational experiments, one of the most extensive publicly available datasets for lithium-ion batteries, which was introduced by Severson et al. [1], was employed. This dataset comprises 124 commercial lithium-ion phosphate (LFP) / graphite cells with a nominal capacity of 1.1 Ah and a nominal voltage of 3.3 V. The cycle lives of these cells range from 150 to 2300 cycles, with cycle life defined as the number of cycles completed corresponding to 80% of the nominal capacity. The dataset generated by Severson et al. contains approximately 96,700 cycle data points from 124 commercial lithium-ion batteries. The average cycle life is 806 cycles with a standard deviation of 377 cycles.

During the data collection process, the batteries were subjected to 72 distinct fast-charging conditions while maintaining identical discharging conditions (4.0 C / 2.0 V). The fast-charging rates varied between 3.6 C and 6.0 C for a duration of 10 minutes, and the batteries were charged until reaching 80% state-of-charge (SOC) conditions using one or two different fast-charging rates. After the completion of fast-charging, the batteries were further charged until reaching 100% SOC using a 1C CC-CV charge, ramping up to 3.6 V with a C/50 charge cutoff. Throughout the data generation process, the voltage, current, internal resistance, and cell temperature were recorded. It is important to note that although the cell temperature was initially set to 30°C, charging and discharging operations could cause the cell temperature to fluctuate by up to 10°C. For more comprehensive details regarding the data collection process, the reader is referred to Severson et al.'s work [1].

Argonne Database

The second battery dataset is obtained from the Argonne Cell Analysis, Modeling, and Prototyping (CAMP) facility [2]. This dataset comprises 300 batteries with six distinct metal oxide cathode chemistries, namely NMC111, NMC532, NMC622, NMC811, HE5050, and 5Vspinel. The selection of batteries for this dataset was based on specific criteria: they utilize graphite as the active material, have charging rates equal to or less than 1C, and have undergone performance testing for a minimum of 100 cycles. It is worth noting that batteries belonging to different chemistries exhibit varying cycle life values, and even within the same chemistry, properties such as porosity, loadings, and materials can vary. In the analysis, the NMC532 and HE5050 chemistries, which are the largest chemistries within the dataset, are used and the experimental results for these two chemistries are presented. For further in-depth information on the CAMP dataset, please refer to [2].

Feature Generation and Data Preprocessing

To generate the input data for the FL framework, various relevant features, including charge and discharge capacity, temperature, and charging time, were incorporated as an input. Furthermore, additional features from the raw data of each cycle, such as the minimum discharge capacity observed, were derived.

The Nature Energy database, as presented in [1], encompasses a wide range of features that are categorized into summary data and cycle data. The summary data provides per-cycle information regarding charge and discharge capacity, internal resistance, charging time, cycle

number, and temperature statistics. On the other hand, the cycle data captures detailed information within each cycle, including the data stream of various features such as charge and discharge capacity, temperature, voltage, and current.

By leveraging these diverse time-series cycle features, it becomes possible to generate new features that contribute to a better understanding of the underlying causes of degradation in lithium-ion batteries. These additional features ultimately enhance the prediction task by providing valuable insights into the degradation behavior. Through the utilization of different time-series cycle features, new features can be specifically designed to improve the understanding of the underlying reasons behind the degradation behavior exhibited by lithium-ion batteries.

Capacity is a pivotal health indicator for lithium-ion batteries [3], providing a time frame for the operation of a fully charged battery under current environmental conditions. Capacity fade curves are frequently generated to observe the degradation behavior of batteries [4, 5]. The capacity itself, along with its temporal changes, serves as a powerful feature for estimating the remaining useful life of batteries. Furthermore, capacity values can be evaluated in conjunction with voltage values, and analyzing their time-series behavior within cycles can offer insights into degradation mechanisms. Notably, Severson et al. [1] predicted the lifetime classification of different cells by leveraging features derived from observed changes in the capacity curve. In a more intricate approach, they incorporated additional features such as temperature, internal resistance, and charge time to enhance their prediction task.

In this study, a set of 68 diverse features were carefully engineered to predict the remaining lifetime of lithium-ion batteries. These features encompass information extracted from both summary and cycle data sources. The summary features provide valuable insights into the internal resistance, charge and discharge capacities, average, minimum, and maximum temperature, as well as charging times observed for each cycle. Additionally, features representing the charging policy have been incorporated into the feature set. To further enrich the feature set, statistical measures such as mean, variance, skewness, and kurtosis were calculated for comparing the cycle data from previous cycles with the current cycle. This comprehensive set of features enables to capture various aspects of battery behavior and maximize the predictive capabilities for remaining lifetime estimation. The list of features generated for Nature database are presented in Table S1.

In the case of the Argonne database [2], a similar feature generation procedure was applied, resulting in the creation of a total of 74 features that are embedded to the dimensionality reduction task. The dataset includes discharge energy and capacity, as well as charge energy and capacity values for each cycle. Additionally, raw feature values are available, which can be leveraged to generate additional features. By utilizing the voltage and current values from the raw data statistical measures such as mean and standard deviation were calculated resulting in the creation of supplementary features. Furthermore, temporal features and statistical measures were derived from the discharge and charge capacities and energies within the database. This comprehensive feature set captures various aspects of the battery behavior and enables effective dimensionality reduction techniques to be employed for further analysis and modeling. The list of features generated for Argonne database are presented in Table S2.

After creating the feature set, the datasets were preprocessed for the computational experiments. For both datasets, similar experimental settings were applied. Since the cycle lives of the batteries in both datasets exceeded 100, the remaining lifetime prediction during the first

Table S1: Feature set for Nature Energy database

Summary Features:
Charge time
Temperature (T)
Min temperature
Max temperature
Internal resistance (IR)
Charge capacity (Q_c)
Discharge capacity (Q_d)
Charging Policy:
Charge rate 1
Charge percentage
Charge rate 2
Cycle Features
$Q_d$lin(t) – $Q_d$lin(10) (a) :
min(a), mean(a), var(a), skewness(a), kurtosis(a)
$Q_d$lin(t) – $Q_d$lin(100) (b) :
min(b), mean(b), var(b), skewness(b), kurtosis(b)
Temporal Summary Features
10 buckets from history of Q_d, Q_c, and IR -bucket means are reported as features
Other Features
Levels and ratios of Q_d, Q_c, and IR at different cycles
Average charge time over first 5 cycles
Maximum and minimum T and IR observed until cycle 100
Maximum and minimum T and IR observed between cycles 100 and t
Maximum(Q_d until t)-Q_d(2)

Table S2: Feature set for Argonne database

Summary Features:
Charge capacity (Q_c)
Discharge capacity (Q_d)
Charge energy (E_c)
Discharge energy (E_d)
Cycle Features:
Average and standard deviation of voltage level at cycle t
Average and standard deviation of current level at cycle t
Temporal Features:
10 buckets from history of Q_d, Q_c, E_d, and E_c -bucket means are reported as features
Other Features:
State of health ratios based on discharge capacity referenced to cycles 20 and 100
Levels and ratios of Q_d, Q_c, E_d, and E_c at different cycles
Maximum Q_d, Q_c, E_d, and E_c observed until cycle 100
Maximum and minimum Q_d, Q_c, E_d, and E_c observed till cycle t

100 cycles was not activated. Instead, information from these initial cycles were collected and used to create various features for the dimensionality reduction task. However, it is important to note that this threshold can be adjusted for different datasets, as remaining lifetime prediction can be done even in the early stages of a battery's lifetime, as indicated by Severson et al. Additionally, both datasets were normalized prior to leveraging dimensionality reduction. The both datasets were split into train and test sets, with the train set consisting of 75% of the data and the test set consisting of 25%.

Note S4: Algorithms for Performance Metric Calculations

Algorithm S4 outlines the step-by-step process for computing the average unused life across all batteries in the test data.

Algorithm S4: Calculation of average unused life

Notation:

$\mathbf{e_i}$: number of unavailable days for lithium-ion battery i

$\mathbf{t_m}$: duration of replacement activities in terms of days

$\mathcal{R}$: set of replaced batteries

for each battery i = 1, 2, ..., $\mathcal{L}$

  if $t^* + t_c < t_{fi}$:

    Extend battery i to set R

    $e_i = t_{fi} - (t^* + t_c + t_m)$

Calculate average unused periods:

$$\bar{e} = \frac{\sum_{i \in \mathcal{R}} e_i}{|\mathcal{R}|}$$

Algorithm S5 presents the sequential steps involved in calculating the average number of days that batteries are unavailable due to replacement operations across the entire test data set.

Algorithm S5: Calculation of average number of unavailable days

Notation:

$\mathbf{u_i}$: number of unavailable periods for lithium-ion battery i

$\mathbf{t_m}$: duration of replacement activities in terms of days

for each battery i = 1, 2, ..., $\mathcal{L}$ **do**:

  if $t^* + t_c < t_{fi}$:

    $u_i = t_m$

  else if: $t^* < t_{fi}$:

    $u_i = (t_c + t_m) - (t_{fi} - t^*)$

  else:

    $u_i = t_c + t_m$

end for

Calculate average unavailable periods:

$$\bar{u} = \frac{\sum_i u_i}{|\mathcal{L}|}$$

Note S5: Experimental Setup and Additional Computational Results

In this section, an overview of the settings for distributed computing and hyperparameter tuning for HOPE-FED framework are provided. Additional computational results for age-based periodic replacement policy (APRP) are presented. In addition, the comparative findings between HOPE-FED and batch-federated approaches are summarized.

Experimental Setup

To establish a robust FL architecture, a high-performance computing (HPC) cluster, where each client's server is represented by separate nodes, was utilized. This design allows an HPC node to access the data of a specific client and train its corresponding model. By employing a distributed computing framework using the Message Passing Interface (MPI), different nodes in the HPC cluster were allocated to represent each client. MPI serves as a standardized means of information transfer between multiple devices, facilitating synchronization and communication among parallel nodes in a constrained setting. For the implementation, OpenMPI was employed along with mpi4py to initiate and manage multiple distributed memory client processes, accurately simulating user devices in real-world field scenarios.

Within the FL framework, the Keras and TensorFlow libraries were used to construct the autoencoder and DNN models for remaining lifetime prediction. The autoencoder structure includes two layers for both the encoder and decoder, excluding the input layers. For remaining lifetime prediction, the DNN consists of seven layers. Hyperparameter tuning was performed for both the federated autoencoder and remaining lifetime prediction tasks to improve their performances. Furthermore, the hyperparameters of the FL algorithm were tailored by exploring variations in the number of federation rounds, the ratio of sampled batteries, and the ratio of sampled data points from each battery per round. Throughout both phases, the Adam optimizer

was employed and the mean squared error loss function was utilized. The initial experiments guided to set the target feature size for the encoder as 30 for the Nature Energy database and 40 for the Argonne database.

Comparative Analysis: HOPE-FED Approach vs. Age-based Periodic Replacement Policy (APRP)

Tables S3, S4, and S5 provide a detailed comparison of the age-based periodic replacement policy and the HOPE-FED framework under different threshold values for triggering replacement activities, specifically for the Nature Energy and Argonne databases with HE5050 and NMC532 chemistries, respectively.

Table S3: Comparison of age-based periodic replacement vs. fully-federated prediction-based replacement approaches on Nature Energy database

	Benchmark Policy	**Threshold=10**	**Threshold=25**	**Threshold=50**	**Threshold=100**
Trigger Time	451	Prediction-based	Prediction-based	Prediction-based	Prediction-based
# Preventive	29	25	30	31	31
# Corrective	2	6	1	0	0
Unused Life	444.5	5.4	20.3	46	114.3
Unavailable Days	1.3	1.6	1.2	1	1
Avg. Cost (in $/day)	**20.3**	**13.4**	**12.6**	**12.9**	**16.1**

Table S4: Comparison of age-based periodic replacement vs. fully-federated prediction-based replacement approaches on Argonne database: HE5050

	Benchmark Policy	**Threshold=10**	**Threshold=25**	**Threshold=50**	**Threshold=100**
Trigger Time	1013	Prediction-based	Prediction-based	Prediction-based	Prediction-based
# Preventive	10	7	10	10	13
# Corrective	9	12	9	9	6
Unused Life	262.9	148.9	114.2	143.8	156
Unavailable Days	3.4	3.9	3.4	3.4	2.6
Avg. Cost (in $/day)	**32.5**	**30.4**	**25.7**	**27.5**	**28.8**

Table S5: Comparison of age-based periodic replacement vs. fully-federated prediction-based replacement approaches on Argonne database: NMC532

	Benchmark Policy	**Threshold=10**	**Threshold=25**	**Threshold=50**	**Threshold=100**
Trigger Time	593	Prediction-based	Prediction-based	Prediction-based	Prediction-based
# Preventive	18	9	17	19	22
# Corrective	7	16	8	6	3
Unused Life	551.8	78.2	73.9	95.8	124.6
Unavailable Days	2.4	3.9	2.5	2.2	1.6
Avg. Cost (in $/day)	**26.5**	**25.7**	**20.6**	**19.1**	**21.2**

Comparative Analysis: HOPE-FED Approach vs. Batch-Federated Approach

Batch-federated learning is a collaborative training approach that proves beneficial for multi-asset clients aiming to leverage the advantages of FL while addressing specific requirements. In batch-federated learning, a subset of the clients within the multi-asset environment participates in aggregating their data during the model training process. Rather than involving all clients, this approach allows for the consolidation of data from a specific group. By aggregating data from these selected clients, batch-federated learning enables the creation of a unified and representative training dataset for model updates. This method strikes a balance between

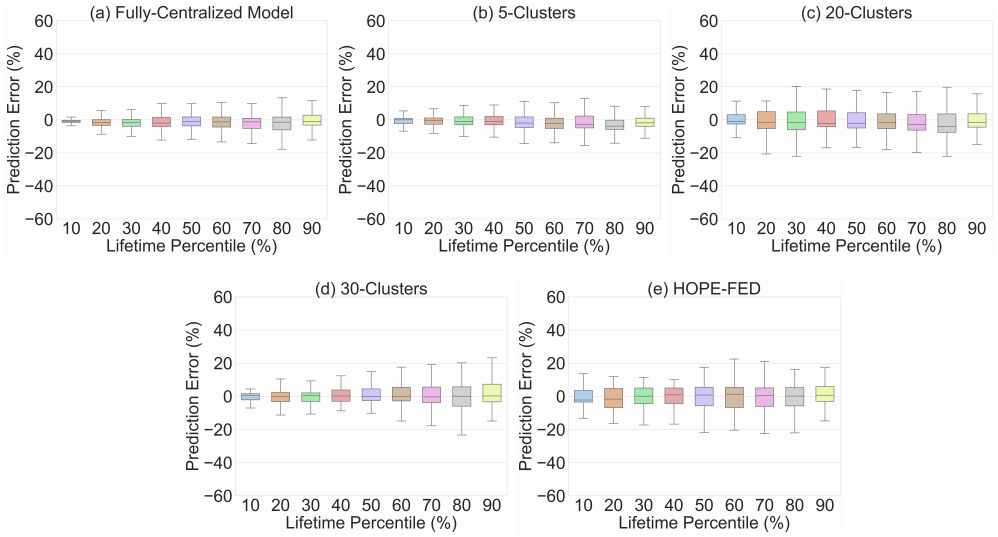


Fig S1: Batch-federated computational results for Argonne database: NMC532.

sharing data for collaborative learning to preserve privacy, and improving operational efficiency specific to multi-asset environments.

Batch-federated learning can be highly advantageous for multi-asset clients, such as vehicle fleet operators, seeking to leverage the benefits of collaborative model training while ensuring data privacy and operational efficiency. By aggregating data from a subset of vehicles within a fleet, batch-federated learning enables the creation of a more diverse and representative training dataset, leading to improved model accuracy and performance. Additionally, the reduced communication overhead in transmitting aggregated data enhances the overall efficiency of the training process, especially in scenarios where bandwidth is limited. With faster convergence enabled by a larger combined dataset, vehicle fleets can derive valuable insights and optimize their operations, and maintenance, across their entire fleet of vehicles.

To implement the batch-federated learning approach, the batteries were randomly divided into 5, 20, and 30 clusters. Within each cluster, the datasets of the associated batteries were aggregated. Increasing the number of clusters brings the approach closer to the proposed fully-federated approach, HOPE-FED. Conversely, fewer clusters align the approach more closely with the fully-centralized approach. The long-run average costs and other relevant metrics for the Argonne database, specifically for the NMC532 chemistry, are documented in Table S6. Additionally, Fig. S1 presents a box plot illustrating the prediction error across different lifetime percentiles.

Fig. S1 displays the box plots depicting the results of batch-federated experiments conducted on the NMC532 chemistry from the Argonne database. The x-axis represents lifetime percentiles, while the y-axis represents prediction errors. Notably, the fully-centralized approach exhibits lower levels of error compared to other approaches. As the number of clusters increases, indicating a decrease in data aggregation, the model characteristics begin to converge to the fully-federated approach. Consequently, the prediction error generally increases as the model moves closer to the fully-federated approach, aligning with expectations.

Table S6: Comparison of batch-federated vs. fully-federated approaches for the Argonne database: NMC532

	Fully-Centralized	5-Clusters	20-Clusters	30-Clusters	Fully-Federated
Optimal Threshold	25	25	50	25	50
# Preventive	19	19	21	18	19
# Corrective	6	6	4	7	6
Unused Life	59.8	85.6	101.3	61.5	95.8
Unavailable Days	2.2	2.0	1.8	2.4	2.2
Avg. Cost (in $/day)	**18.8**	**19.0**	**18.9**	**18.5**	**19.1**

The findings presented in Table S6 demonstrate that utilizing 5 and 20 clusters yields long-run average costs that fall between those of the fully-centralized and fully-federated approaches. Conversely, employing 30 clusters slightly enhances the cost rate performance of both the fully-centralized and fully-federated approaches. These results affirm the competitive capability of the HOPE-FED approach in terms of performance. Furthermore, the batch-federated approach becomes beneficial in reducing computational load, particularly in scenarios where clients have the flexibility to aggregate datasets from multiple assets they possess.

Note S6: Hyperparameter Selection

The hyperparameters for both the federated autoencoder and federated RUL prediction were determined through preliminary experiments. This section presents the results of these initial tests. Table S7 displays the computational experiment results conducted to select hyperparameters for the federated autoencoder using the Nature dataset. The table varies two key hyperparameters: the number of federation rounds and the number of target features, which represent the reduced dimension size resulting from the autoencoder transformation. This reduction determines the input size for the subsequent prediction task, federated RUL prediction. Both the mean square error (MSE) of the training and test sets are reported. The choice of 2000 federation rounds and 30 target features was based on the observation that increasing either measure raises computation times without proportionally reducing MSE levels. Similar computational experiments were conducted for the Argonne database with NMC532 and HE5050 chemistries, as detailed in Table S8. In this dataset, 2000 federation rounds and 40 target features were selected.

Hyperparameters for federated RUL prediction were determined based on monitoring both cost rates and average absolute percentage errors. Table S9 presents the results of hyperparameter tuning using the Argonne database HE5050 chemistry training data. Increasing the number of federation rounds generally reduces the average absolute error. To mitigate overfitting, the number of federation rounds was capped at 4000, as beyond this point, the cost rate stabilizes with minimal improvement. Similar experiments were conducted for other databases, setting the number of federation rounds to 7500 for the Nature database and 6000 for the Argonne database NMC532 chemistry.

Table S7: Autoencoder hyperparameter selection for Nature database

Federation Rounds	Number of Target Features	MSE of Train Data	MSE of Test Data
500	30	0.0027	0.0054
500	40	0.0027	0.0057
500	50	0.0027	0.0055
1000	30	0.0018	0.0045
1000	40	0.0015	0.0041
1000	50	0.0015	0.0039
2000	30	0.0009	0.0033
2000	40	0.0009	0.0032
2000	50	0.0009	0.0033
3000	30	0.0006	0.0033
3000	40	0.0007	0.0031
3000	50	0.0007	0.0029
4000	30	0.0006	0.0029
4000	40	0.0006	0.0029
4000	50	0.0006	0.0029
5000	30	0.0006	0.0031
5000	40	0.0005	0.0029
5000	50	0.0006	0.0029

Table S8: Autoencoder hyperparameter selection for Argonne database

		Chemistry: NMC532		Chemistry: HE5050	
Federation Rounds	Number of Target Features	MSE of Train Data	MSE of Test Data	MSE of Train Data	MSE of Test Data
500	40	0.0016	0.0022	0.0011	0.0018
500	50	0.0015	0.0020	0.0011	0.0016
500	60	0.0013	0.0017	0.0010	0.0017
1000	40	0.0007	0.0011	0.0006	0.0010
1000	50	0.0008	0.0012	0.0007	0.0011
1000	60	0.0007	0.0012	0.0006	0.0010
2000	40	0.0003	0.0004	0.0004	0.0008
2000	50	0.0004	0.0005	0.0003	0.0007
2000	60	0.0004	0.0007	0.0005	0.0010
3000	40	0.0003	0.0005	0.0003	0.0008
3000	50	0.0002	0.0004	0.0003	0.0007
3000	60	0.0002	0.0003	0.0003	0.0007
4000	40	0.0002	0.0004	0.0003	0.0007
4000	50	0.0002	0.0004	0.0003	0.0008
4000	60	0.0002	0.0003	0.0003	0.0007

Table S9: Cost rates and average absolute percentage errors across varying numbers of federation rounds for RUL prediction in HE5050 chemistry training data

	Number of Federation Rounds				
Threshold	**1000**	**2000**	**3000**	**4000**	**5000**
10	17.65	19.14	16.61	14.97	17.36
25	17.04	15.45	15.39	14.20	14.17
50	16.89	15.86	15.79	15.12	14.98
100	19.13	18.50	18.18	18.08	18.14
Best Cost Rate (in $/day)	16.89	15.45	15.39	14.20	14.17
Average Absolute % Error	10.06%	8.18%	6.25%	4.64%	2.97%

References

1. 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. Nature Energy *4*, 383–391.

2. Paulson, N. H., Kubal, J., Ward, L., Saxena, S., Lu, W., and Babinec, S. J. (2022). Feature engineering for machine learning enabled early prediction of battery lifetime. Journal of Power Sources *527*, 231127.

3. Diao, W., Saxena, S., Han, B., and Pecht, M. (2019). Algorithm to determine the knee point on capacity fade curves of lithium-ion cells. Energies *12*, 2910.

4. Saxena, S., Ward, L., Kubal, J., Lu, W., Babinec, S., and Paulson, N. (2022). A convolutional neural network model for battery capacity fade curve prediction using early life data. Journal of Power Sources *542*, 231736.

5. Honkura, K., Takahashi, K., and Horiba, T. (2011). Capacity-fading prediction of lithium-ion batteries based on discharge curves analysis. Journal of power sources *196*, 10141–10147.