Data-Driven SoC Forecasting in Electric Vehicles: A Federated Learning Perspective

Tanish Patel[0009−0001−1381−2986] and
Harshvardhan Gaikwad[0000−0002−1787−1211]

Abstract Electric vehicle (EV) adoption stands as a pivotal step in curbing carbon emissions and combating climate change. However, the persistent specter of range anxiety continues to impede widespread acceptance. Traditional State of Charge (SoC) estimation methods, coupled with basic machine learning models, grapple with accuracy and adaptability limitations. Yet, the advent of Federated Learning heralds a transformative era in the EV landscape, placing a premium on data security and privacy. This approach involves training models across a network of distributed sources, such as individual EVs, harnessing the wealth of diverse data streams. It continually refines models, ensuring SoC calculations maintain precision as the EV fleet evolves. Notably, it fortifies against cyberattacks by obviating centralized data storage. Embracing Federated Learning within the EV industry not only alleviates range anxiety but also fosters sustainable transportation, underscoring the role of EVs in shaping an ecologically conscious future. This paper concentrates on SoC estimation through the prism of Federated Learning. Techniques including Federated Averaging, Differential Privacy Techniques, and Ensemble Learning will be employed. To facilitate this research, a comprehensive vehicle model, encompassing the powertrain and heating circuit, was validated through real driving trips with a BMW i3 (60 Ah), yielding the requisite dataset. This dataset underwent meticulous extraction, cleansing, and exploratory data analysis involving numerous parameters. These preparatory steps lay the groundwork for the subsequent application of Federated Averaging, Differential Privacy Techniques, and Ensemble Learning to the dataset, aiming for precise SoC estimation.

Tanish Patel
Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar e-mail: tanishpatel0106@gmail.com

Harshvardhan Gaikwad
Department of Electrical Engineering, School of Energy Technology, Pandit Deendayal Energy University, Gandhinagar e-mail: harshishere.me@gmail.com

1 Introduction

The evolution of electric vehicles (EVs) has brought forth a myriad of technological advancements and challenges. Among the most crucial aspects of EVs is the ”State of Charge” (SoC), which refers to the current battery capacity as a percentage of its maximum capacity. Understanding the SoC is pivotal, not only for users who need to know when to recharge their vehicles but also for manufacturers and service providers aiming to optimize battery lifespan and vehicle performance.

However, the mere recognition of SoC's importance isn't sufficient. The precision in its estimation is equally, if not more, crucial. Accurate estimation of the State of Charge ensures the safe operation of EVs, providing real-time data that can be pivotal for decisions related to charging, discharging, and overall vehicle operation.

In the quest for more accurate and efficient methods of data processing and analysis, the concept of Federated Learning emerges as a promising approach. Federated Learning is a machine learning technique where the model is trained across multiple decentralized devices or servers holding local data samples, without exchanging the data itself. This method stands in stark contrast to traditional centralized learning and offers significant advantages, especially in terms of data privacy and efficiency.

With the increasing integration of EVs into our daily lives, the datasets associated with them have become more comprehensive and valuable. This upsurge in data brings to light the indispensable need for robust cybersecurity measures. Any compromise in the EV dataset could lead to severe consequences, ranging from financial losses to potential threats to passenger safety.

This is where the beauty of Federated Learning truly shines. By allowing data to remain on local devices and only updating the model's parameters, Federated Learning inherently provides an added layer of cybersecurity. Not only does this approach enhance the protection of sensitive data, but it also paves the way for more collaborative and large-scale machine learning endeavors without the associated risks of data breaches.

2 Related Works

Lee et al. [1] explore the potential of a data-driven Gaussian process (GP) for estimating the state-of-charge (SOC) in batteries. While traditional methods such as Coulomb counting and voltage-based estimation have their limitations, the GP model uses historical charging and discharging data for more accurate SOC predictions. Notably, this model outperforms conventional methods, especially under varied operating conditions. This work underscores the significance of data-driven techniques in enhancing battery management systems for better longevity and performance.

Song et al. [2] present a detailed review of energy management strategies (EMS) for hybrid electric vehicles (HEVs) using machine learning (ML) techniques. They argue that traditional rule-based EMS face challenges in adaptability and optimization. On the other hand, ML techniques, including neural networks and decision trees, hold promise in improving both fuel efficiency and battery lifespan. The emphasis is on the capability of ML in predicting driving patterns, which leads to better HEV performance. This area remains ripe for further exploration.

Hu et al. [3] focus on the application of advanced ML techniques in the management and optimization of lithium-ion batteries. Highlighting the shortcomings of conventional approaches, the study reveals the advantages of ML in predicting battery degradation and refining SOC estimation. Experimental results particularly spotlight deep learning networks as effective tools for battery health forecasting. These findings advocate for a more significant ML role in battery management, emphasizing improved safety and efficiency.

Chandran [4] evaluates the accuracy of various ML algorithms, including SVM, Random Forest, and Neural Networks, in the SOC estimation of lithium-ion batteries for electric vehicles. The research highlights that ML-based methods generally exceed the accuracy of traditional techniques, especially in dynamic driving scenarios. The integration of these algorithms with real-time sensor data can lead to improved battery management systems and safer EV operations.

Harippriya et al. [5] introduce a novel approach to estimate battery aging using both deep learning (DL) and ML algorithms within battery management systems (BMS). By employing models like CNNs and Random Forest, the research indicates that DL methods can more precisely predict battery end-of-life than conventional techniques. Such integration can enhance battery lifespan and offer more reliable energy storage solutions across various sectors.

Manoharan et al. [6] tackle the intricate issue of SOC estimation in multi-cell battery packs for electric vehicles. They propose a parallel algorithmic approach that estimates the SOC of individual cells simultaneously. This parallel strategy, which merges data fusion techniques with ML models, proves more accurate and faster than sequential methods. The results highlight the advantages of parallel computing in BMS, ensuring precise SOC predictions crucial for EV safety.

How et al. [7] present an innovative methodology for SOC estimation in Li-ion batteries. Recognizing the vital role of accurate SOC predictions, the study introduces an algorithm that integrates sensor data with ML techniques. This method stands out in terms of accuracy and adaptability, especially in dynamic settings, reinforcing the benefits of modern computational techniques in battery management.

Zhang et al. [8] review the role of deep learning (DL) techniques in SOC estimation for Li-Ion batteries in EVs. They highlight the transformative power of DL models, such as CNNs, RNNs, and LSTMs, which show considerable improvements in SOC predictions over traditional methods. This review suggests that DL's adoption can lead to more efficient and reliable battery management in EVs.

Pang [9] emphasizes the importance of accurate SOC estimation in batteries. The study contrasts traditional methods with a newly proposed algorithm that combines sensor inputs with computational techniques, leading to better precision, especially under variable conditions. This research suggests that leveraging sophisticated computational methods can greatly enhance battery performance and lifespan.

Mashkov et al. [10] applied SVM and DNN models to the NASA Battery dataset, demonstrating effective online SoC estimation using input features like battery volt-

Table 1 Number of data records in each trip

Trip Number	Number of Datapoints
1	10090
2	14130
3	6760
4	6760
5	13668
6	31646
7	20934
8	28060
9	18345
10	14177

age and temperature, suggesting a potential for real-time applications in battery management systems. Meanwhile, Youssef et al. [11] tested ML algorithms including MLR, MLP, SVR, and RF on mixed driving cycles datasets, with the RF model showing the best performance, underscoring the effectiveness of ensemble methods in handling complex, nonlinear SoC estimation tasks

Sadykov et al. [12] explored the application of RNNs, including GRU and LSTM models, for EV battery SoC estimation. The models showed high accuracy but highlighted challenges in capturing mid-range SoC values, suggesting areas for further improvement. Ni and Yang [13] introduced physics-constrained neural networks to reduce abrupt errors in ML-based SoC estimations, improving accuracy by integrating physical laws directly into the learning process, showcasing a novel approach to hybrid model-data methods

Dang et al. [14] developed differential equation-informed neural networks (DENNs) that incorporate differential equations in training to enhance stability and accuracy in SoC estimation, offering a sophisticated method that blends deep learning with traditional engineering models. El Maliki et al. [15] combined the Thevenin model with an extended Kalman filter to refine SoC estimation, significantly enhancing accuracy, which is critical for real-time battery management applications

Li, Jin, and Yu [16] proposed an online estimation method using a dual time-scale technique that enhances the accuracy of estimating battery parameters and SoC without requiring open-circuit voltage experiments, demonstrating potential for adaptive, real-time applications

3 Proposed Methodology

The dataset is extracted from an open-source website [33] .It consisted of two categories namely A and B, but due to technical difficulties in processing power faced by the systems, only the first 10 files of Category A have been considered. The composition of data records for each file has been presented in the Table 1.

The anomalies and inconsistencies in the dataset had to be cleared by cleaning the entire dataset. Handling missing values, removing outliers and standardizing values to ensure data integrity and consistency throughout had to be done too.

After performing exploratory data analysis on the dataset, the following correlation heat map was developed. The map in the fig.(1) along with the knowledge that to estimate the state of charge of a high voltage battery, one would typically consider parameters like cell voltage, current, temperature and historical charging/discharging cycles helped to decide which columns should be considered for furthering the application of models on the dataset.

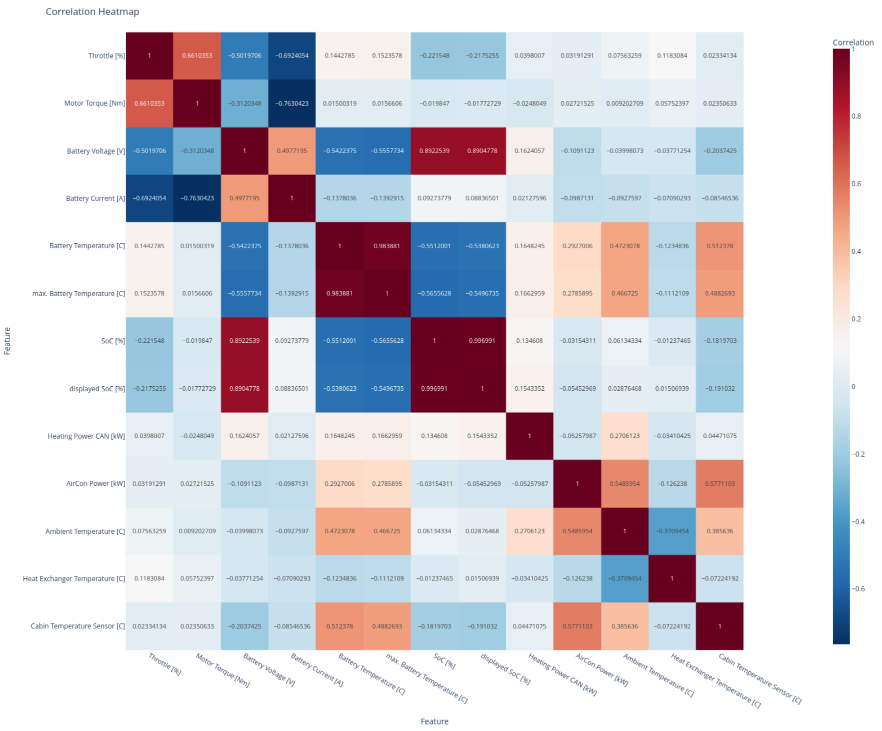

Fig. 1 Correlation Plot for the Data

The total number of parameters considered were 13 and the total number of data records are 1,61,881 that have been considered and thus the total datapoints have summed up to a staggering 21,04,453. Thus, it can be said that using federated learning on such a huge dataset will assist in acquiring considerably high efficiency of predictions for the state of charge.

The battery stands out as one of the most important components in the fast-changing world of electric vehicles (EVs). In addition to providing the vehicle's power and assuring its mobility, it also affects the EV's general performance, longevity, and operating safety. Therefore, predicting the State of Charge (SoC) of these batteries with accuracy becomes essential. The SoC functions similarly to the fuel gauge in conventional automobiles by displaying the current battery charge in relation to its capacity. When drivers know exactly how much range their EV has left, it makes it easier to plan routes, make charging decisions, and reduces the dreaded ”range anxiety.”

Additionally, EV batteries are considered possible energy storage options, enabling a more robust and adaptable energy system, as renewable energy sources become more and more integrated into the grid. An accurate SoC prediction under such circumstances can help with better grid management, demand-response tactics, and even monetary considerations for energy trading.

Traditional techniques of determining SoC, based on simple models or direct measurements, sometimes fall short in practical situations. This is caused by the complex chemistry of batteries, a wide range of outside factors that have an impact on battery performance, and the inherently variable nature of consumption patterns. Enter the world of data-driven methodologies and machine learning. When traditional methods are ineffective due to the lack of comprehensive battery chemical and physical data, these data-driven methodologies present a viable path for improving SoC estimation accuracy [17, 18].

In this section, we delve into various machine learning methodologies employed for SoC prediction, exploring their foundational principles, mathematical underpinnings, and potential advantages in a federated learning context. [19]

3.1 Federated Learning

Federated Learning (FL), a new strategy in machine learning, trains algorithms across several servers or devices without centralized data. This decentralized method of learning has important ramifications, particularly in settings where data security, privacy, and transmission costs are crucial considerations.

Fundamentally, FL works by delivering the model to the data, as opposed to the more conventional method of sending the data to the model. This inversion is essential in situations where data cannot be shared because of privacy concerns, legal restrictions, or sheer volume [20] . To address privacy concerns, for example, training models on consumer devices like smartphones guarantees that sensitive data (such private messages or health information) never leaves the device.

Mathematically, the essence of Federated Learning can be distilled to the process of local training and global aggregation. For a given number of K devices, the federated learning process is as follows as in algorithm 3.1

FL’s decentralized structure has advantages beyond merely privacy. It also makes it possible to use a variety of data sources, which might result in models that are stronger and more generalized[21]. Stragglers (devices that take a long time to train and update), device dropouts, and uneven data distributions among devices are some of the new difficulties FL presents[22].

Algorithm 1 Federated Learning

Initialization: A Global model with parameter θ is initiated.

Local Training: The Global model with the set parameter is sent to a subset of clients with each device Ki updating the model based on the local data for the client, hence resulting in a local model update as in eq. (1)

$$\delta\theta_k = LocalTraining(\theta, Data_k) \tag{1}$$

Global Aggregation: The Local Client Model Updates are sent back to the central server or the global server and are aggregated to update the global model. Specifically, Weighted Averaging is used here as given in eq. (2)

$$\theta_{\text{global}} = \theta + \sum_{i=1}^{i=K} w_i \cdot \Delta\theta_i \tag{2}$$

Where w_i is the weight that is set based on the number of samples or the size of the dataset on device Ki.

Iteration: Steps Local Training & Aggregation are iteratively performed till convergence or a set number of rounds

By using FL to implement SoC prediction, each vehicle (or charging station) will be able to develop local SoC prediction models based on its own usage patterns and battery characteristics. A more thorough and precise SoC estimation technique that can be applied to a variety of circumstances can then be provided by the global model, which has been updated with insights from numerous devices[23].

3.2 Lasso Regression

The common linear regression technique known as Lasso Regression, or ”Least Absolute Shrinkage and Selection Operator,” incorporates L1 regularization [24]. This regularization term’s inclusion not only promotes sparsity but also aids in preventing overfitting. Lasso Regression can effectively eliminate unimportant features as a result, producing a parsimonious and interpretable model. Due to its inherent feature selection capabilities, Lasso is very helpful when working with datasets that contain a lot of characteristics or features. Mathematically the objective function for LASSO Regression can be defined as in eq(3)

$$\min_{\beta}\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{3}$$

Here, β are the coefficients, λ is the regularization strength, and p represents the number of features. The L1 penalty $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ is responsible for the shrinking coefficients and try to limit it to zero, leading to a more interpretable model[25].

3.3 Ridge Regression

A linear regression extension that incorporates an L2 penalty component is called ridge regression, commonly referred to as Tikhonov regularization[26]. When there is multicollinearity or when there are more predictors than observations in the dataset, this type of regularization seeks to reduce overfitting. Ridge Regression ensures that the coefficients don't assume huge values by applying a penalty proportional to the square of the magnitude of the coefficients. A more reliable and stable model is produced as a result. Mathematically Objective Function for Ridge Regression is depicted as in eq(4)

$$\min_{\beta}\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right)+\lambda\sum_{j=1}^{p}{\beta_j}^2 \tag{4}$$

It is clearly observed that in comparison to Lasso Regression, Ridge Regression used the L2 Regularization

3.4 ElasticNet

ElasticNet is a powerful regression method that combines the benefits of Ridge Regression and Lasso Regression[27]. When working with datasets that have multicollinearity or a lot of features, it introduces a combined penalty term that makes it especially helpful. ElasticNet maintains a balance between coefficient shrinkage and feature selection, resulting in a model that is comprehensible and reliable. The Objective function for ElasticNet is similar to that of Ridge and Lasso Regression with the only difference that ElasticNet uses both the L1 and L2 Regularization Parameters as seen in eq(5)

$$\min_{\beta}\left(\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right)+\lambda\sum_{j=1}^{p}\left|\beta_j\right|+\lambda\sum_{j=1}^{p}\beta_j^2 \tag{5}$$

3.5 Support Vector Regression

The concept of support vector machines is extended to regression issues by support vector regression (SVR). SVR looks for a hyperplane that most closely resembles the data while making sure that deviations don't go above a certain margin ϵ. In order to deal with non-linear relationships in data, SVR typically employs the Radial Basis Function (RBF) kernel, which transforms the input data into a higher-dimensional space where a linear relationship may be identified2. The use of SVR with RBF kernel for State of Charge (SoC) prediction in electric vehicles is one of its enticing features. Recent research has shown that RBF kernel equipped SVR provides a reliable and accurate methodology for SoC forecasting, beating other traditional methods[17]. The optimization for SVR is given as in eq(6)

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}|w|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) - \epsilon\sum_{i=1}^{N}\left(y_i - w^T\phi\left(x_i\right) - b\right) \tag{6}$$

subject to the constraints as:

$$\begin{aligned} y_i - w^T\phi\left(x_i\right) - b &\leq \epsilon + \xi_i \\ w^T\phi\left(x_i\right) + b - y_i &\leq \epsilon + \xi_i^* \end{aligned} \tag{7}$$

3.6 Random Forest Regressor

When doing regression tasks, Random Forest is a potent ensemble learning technique that builds many decision trees during the training phase and delivers the average prediction of these individual trees[20]. This ensemble approach is chosen for complex regression issues since it naturally manages biases, variances, and overfitting.

Random Forest has lately become a viable option for predicting an electric vehicle's State of Charge. Random Forest has been demonstrated to perform better in terms of accuracy and resilience in SoC estimation than numerous other machine learning algorithms due to its capacity to process high-dimensional data and capture subtle patterns [29]. Its capacity to prioritize feature importance offers insights into which predictors have the greatest influence on the SoC, assisting in the development of improved battery management and diagnostic techniques. Mathematically, the criterion for splitting nodes in the trees within a Random Forest is typically the Mean Squared Error (MSE).

3.7 Gradient Boosting Machine

A sophisticated ensemble machine learning technique that expands on the boosting principle is the gradient boosting machine (GBM). In contrast to Random Forest, which develops trees separately, GBM builds trees sequentially, with each tree correcting the flaws of the one before it. Through an iterative process, decision trees, which are normally weak learners, can be strengthened into powerful predictive models.

The optimization of a loss function is the core of GBM. GBM effectively moves in the direction that minimizes the error by fitting a new tree to the loss function's negative gradient (or ”pseudo-residuals”) at each step. Due to its adaptability to different loss functions, GBM is able to be used for both regression and classification applications, allowing for its versatility. [30]

Due to its capacity to deal with non-linear correlations, manage missing data, and deliver excellent accuracy even with scant training data, GBM has gained popularity in the field of SoC prediction. Better battery management and diagnostics are made possible by its feature importance rating, which also provides insights into the crucial aspects affecting SoC. [28]

3.8 Applied Methodology

Based on the supervised algorithms described earlier in this section, a methodology is defined so as for implementing Federated Learning to inculcate the factor of privacy for training Machine Learning models for State of Charge prediction. Following approach as given in algorithm 2 has been used in this specific research article

4 Results and Insights

An increased emphasis has been placed on effective battery management systems, particularly precise State of Charge (SoC) prediction, as a result of the global shift to electric vehicles (EVs). In addition to reducing the "range anxiety" that many EV drivers experience, a precise SoC forecast also helps to optimize grid management, considering the expanding role of EV batteries as workable energy storage systems [31, 32].

While classic SoC estimate techniques have significant advantages, they frequently face difficulties brought on by complex battery chemistries and a variety of external factors3. In light of this, machine learning has been identified as a potential game-changer for SoC prediction. It is praised for its aptitude for understanding intricate patterns and processing massive information. This paper launches a thorough evaluation of a few machine learning approaches - namely, Lasso Regression, Ridge Regression, ElasticNet, SVR, Random Forest Regressor, and Gradient Boosting Machine which are used to gauge their applicability, performance, and nuances in the domain of SoC forecasting.

Given the complex structure of the data and the complex behavior of batteries under various conditions, predicting the State of Charge (SoC) in electric vehicles requires accuracy. The evaluation of our predictive models becomes crucial as we navigate this complex environment. We try to capture the core of model performance through rigorous metric selection, assuring both accuracy and reliability.

The Mean Squared Error (MSE) is a central metric in this evaluation, representing the average of the squared differences between the predicted values $\widehat{y_l}$ and the actual values y_i . It can be mathematically expressed as in eq(8)

Algorithm 2 Application of Federated Learning for State-of-Charge Prediction

```
Require: List of CSV files, Model Training Function, Model Aggregation Function, Number of
    Rounds, Number of Clients = Number of Files
Ensure: Evaluation Metrics over Rounds, Visualization Plots
    Initialization:
    Initialize dictionaries: clients_data and clients_metrics.
    Data Loading & Preprocessing:
    for each file in csv_files do
        Load data from file into a dataframe data.
        Check for Null Values and Drop the not required columns.
        Split data into input features X and target variable y.
        Assign tuple (X, y) to clients_data using the file name or an identifier as the key.
    end for
    Federated Training for Models:
    for each round_num in range of n_rounds do
        Initialize dictionary client_models.
        for each client's key in clients_data do
            Split the client's data into shards for the current round.
            Train the model using train_model_function on the shard for the current round.
            Assign the trained model to client_models using the client's key.
        end for
        Aggregate the models in client_models using aggregate_models_function to get a new
            global_model.
        Use the size of each client's training shard as weights for the aggregation.
        for each client's key in clients_data do
            Evaluate the global_model using the client's data for the current round.
            Store evaluation metrics for that client and the current round in clients_metrics using the
                client's key.
        end for
    end for
    Results:
    for each client's key in clients_metrics do
        Print final evaluation metrics after all rounds for that client.
    end for
```

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_l})^2 \tag{8}$$

where n being the number of observations. The strength of MSE lies in its sensitivity to outliers, as it penalizes larger errors in comparison to the smaller ones.

Now, Complementing the MSE is the Mean Absolute Error (MAE), which is an indictment of a direct measure of prediction accuracy. It is defined as the average of the absolute differences between the predicted and the actual values. MAE treats all the errors on common ground and offers a clear view on the measure of the overall prediction accuracy. It can be mathematically depicted as in eq(9)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_l}| \tag{9}$$

One another metric of importance is the Root Mean Squared Error (RMSE) which can be mathematically articulated as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_l})^2} \tag{10}$$

It is well observed that RMSE is essentially the square root of the Mean Squared Error, RMSE has an added advantage of being in the same unit as the target variable, hence providing an interpretable measure of the model's overall prediction accuracy.

The Coefficient of Determination or popularly the R^2 coefficient quantifies the proportion of variance in the dependent variable and as explained by the independent variables. It is represented as in eq(11)

$$R^2 = 1 - \frac{SS_{\text{Res}}}{SS_{\text{Tot}}} \tag{11}$$

where Sum of Residual Squares is given in eq(12)

$$SS_{\text{Res}} = \sum_{i=1}^{n}(y_i - \widehat{y_l})^2 \tag{12}$$

and the Total Sum of Residual Squares is given in eq(13)

$$SS_{\text{Tot}} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{13}$$

An R^2 value close to unity indicated that the model is able to explain a large proportion of variance in the dependent variable and a value close to zero indicated that the model is not able to capture most of the patterns in the data and is not able to explain the large proportion of variance in the data.

As previously discussed, that for federated learning, we would be using a strategy is what referenced as Weighted Global Aggregation, the ten rounds of training that is performed here is done on 10 different comma value separated files which constitute the dataset. Notably, not all of these files are uniform and hence the weighted averaging is used here. The fractional contributions of the data to a Global Model are shown as in fig(2).

Hereupon, the different machine learning models were explored and simultaneously all the models are trained. First, we explore Lasso Regression. For Lasso, the main hyperparameter of interest is the regularization strength. An exhaustive grid search was conducted to identify the optimal value for this hyperparameter, ensuring that the model struck a balance between bias and variance. The plot for the same is visualised in fig.(3).

Secondarily, Ridge Regression model is trained over the federated learning setup created here. To harness the full potential of Ridge Regression, it was essential to calibrate its regularization strength optimally. A comprehensive grid search was conducted to fine-tune this hyperparameter, ensuring that the model neither overfits nor underfits the training data. Ridge Regression, characterized by its L2 regularization, is a stalwart in the realm of linear models, especially when the dataset has features that are correlated. By adding a degree of bias, Ridge Regression reduces the variance of predictions, making the model more robust and less susceptible to overfitting. The Visualization plot for Ridge Regression is given as in fig(4).

As a profound fact as discussed, ElasticNet is a model that uses both the L1 and L2 regularization, the metric plot for ElasticNet is given in fig(5).

It is also established that Gradient Boosting is an ensemble technique that builds on weak learners, typically decision trees, in a sequential manner. Each tree aiming to correct the errors of its predecessor. Key hyperparameters like the learning rate, number of boosting rounds, and tree-specific parameters were meticulously tuned. The Visualization plot for the same is given in fig(6).

Support Vector Regression (SVR) operates by finding a hyperplane that best fits the data, aiming to have the maximum number of data points within a specified margin. It's known for its capability to handle non-linear relationships when equipped with appropriate kernels. The training involved selecting the most appropriate kernel (linear, polynomial, RBF) and tuning hyperparameters like the C (regularization parameter) and epsilon. The Visualization plot for the metrics of SVR is denoted in fig(7).

Random Forest, an ensemble of decision trees, is celebrated for its versatility and robustness. By aggregating predictions from multiple trees, it achieves high accuracy and is often resistant to overfitting. The training phase involved tuning key parameters like the number of trees, max depth, and min samples split. The plot for same is depicted in fig(8).

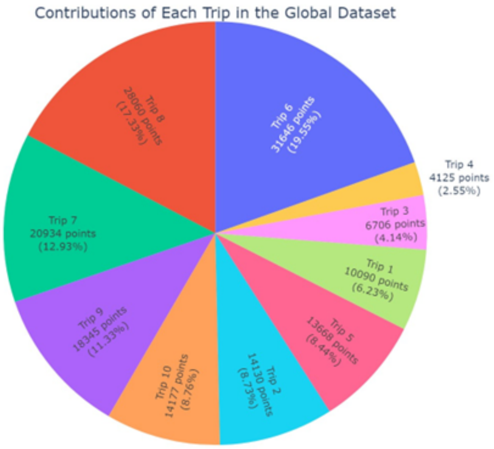

Fig. 2 Distribution of the Data points of the dataset

Table 2 Summary of the results

Model	MSE	MAE	RMSE	R2
ElasticNet	0.674915	0.660012	0.821309	0.995261
Ridge Regression	0.288179	0.401601	0.534815	0.987977
Lasso Regression	0.359379	0.466791	0.598882	0.987477
Random Forest Regressor	0.056105	0.161711	0.232401	0.999606
Gradient Boosting Machine	0.075907	0.211813	0.270880	0.998467
Support Vector Regressor	0.368143	0.288797	0.606587	0.997415

Table 3 Parameter settings for different models

Model	Alpha	L1	Random State	Kernel	N Estimators	Learning Rate	C	Epsilon
ElasticNet	0.1	0.5	42	-	-	-	-	-
Ridge Regression	0.1	-	42	-	-	-	-	-
Lasso Regression	0.1	-	42	-	-	-	-	-
Random Forest Regressor	-	-	42	-	100	-	-	-
Gradient Boosting Machine	-	-	42	-	100	0.1	-	-
Support Vector Regressor	-	-	-	rbf	-	-	1.0	0.1

The Overall metrics for all the six models that are applied here are tabulated in the following table(2) and the hyperparameters used in the machine learning models are summarised in table(3).

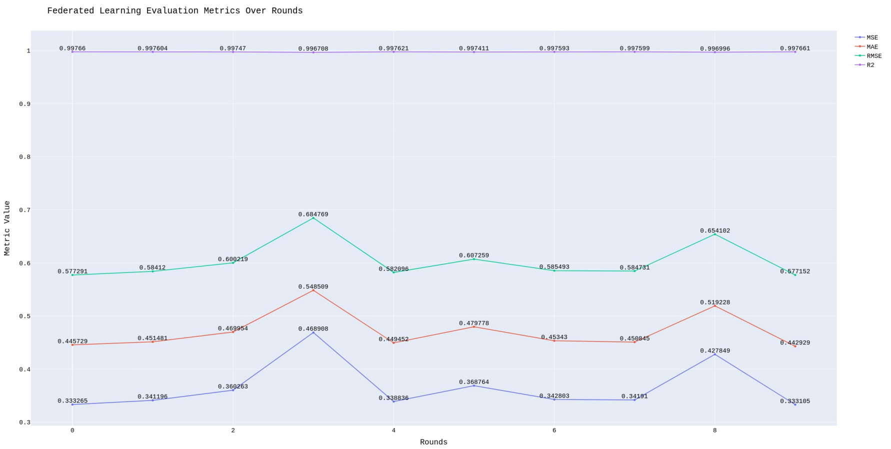

Fig. 3 The Visualization plot for Lasso Regression

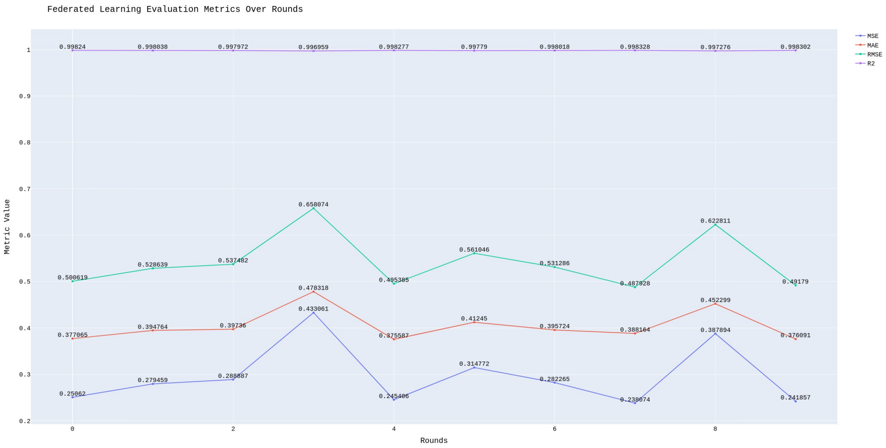

Fig. 4 The Visualization plot for Ridge Regression

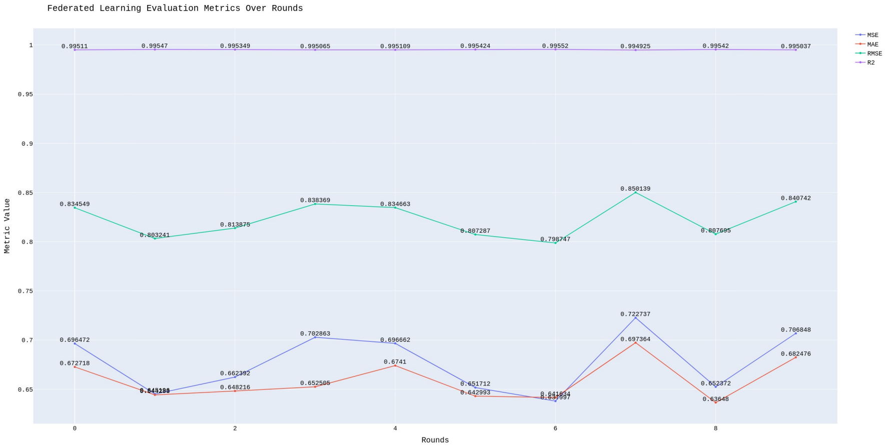

Fig. 5 The Visualization plot for ElasticNet

The accuracy of State of Charge (SoC) predictions has become a crucial topic considering the global movement towards electric vehicles (EVs). As the world embraces the electric revolution, it is crucial to allay EV consumers' ”range anxiety” and enable effective grid management. Our in-depth analysis of six different strategies highlights the significant potential that machine learning models have in this field. The Random Forest Regressor is the most effective model, followed closely by the Gradient Boosting Machine, according to our findings, which are supported by a thorough metric analysis. It's important to stress that the best model selection is inextricably linked to the properties of the underlying data.

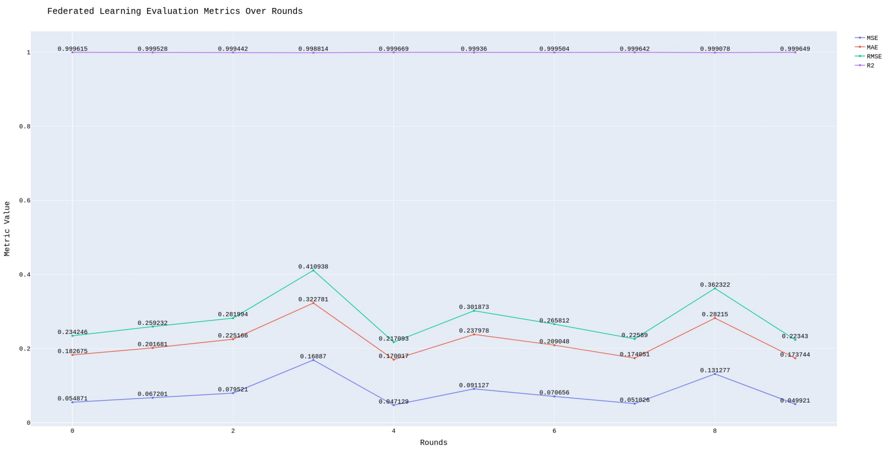

Fig. 6 The Visualization plot for Gradient Boosted Regressor

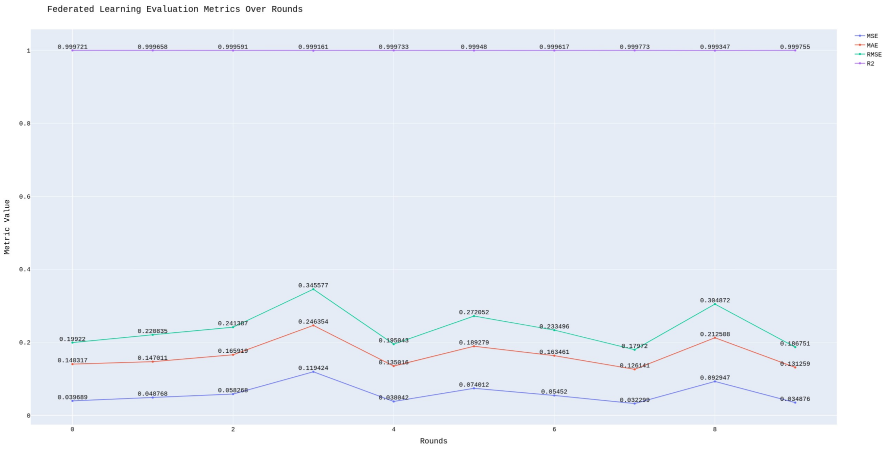

Fig. 7 The Visualization plot for Support Vector Regressor

5 Future Scope

The nascent field of Electric Vehicles (EVs) presents a broad field for more study in data driven SoC forecasting. Federated Learning's decentralized structure, while intriguing, also highlights a number of difficulties. A closer examination of solutions to problems including stragglers, device dropouts, and unequal data distributions among devices is necessary. The integration of EVs with smart grids and the Internet of Things (IoT) has exciting potential as well. Real-time SoC forecasts could be improved by dynamic charging suggestions depending on grid demand as urban

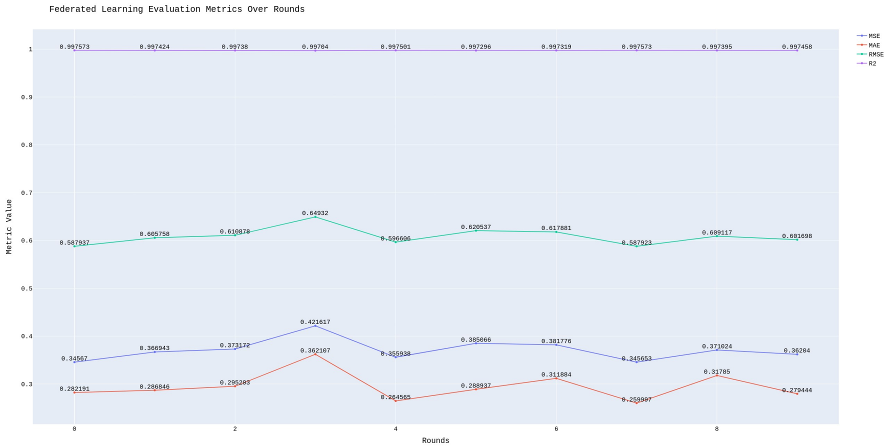

Fig. 8 The Visualization plot for Random Forest Regressor

landscapes develop into ”smart” entities, maximizing both user experience and grid efficiency.

SoC forecasting has not yet utilized all of deep learning models’ extensive capabilities. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are examples of architectures that may reveal models that may accommodate the numerous nuances of complicated datasets. Furthermore, because modern EVs are fitted with a variety of sensors, the combination of several data sources—covering driving habits, environmental circumstances, and battery health—can offer a thorough perspective on SoC, leading to improved forecast accuracy.

These models must be secure and resilient, especially in light of potential hostile attacks. Making sure machine learning is robust as EV systems grow more integrated with it is essential for their widespread adoption. Another intriguing option is tailored SoC forecasting. Models that adapt to individual user behaviors may be able to provide customized and accurate SoC estimates given the variations in personal driving habits, vehicle usage, and maintenance. Last but not least, as the world’s focus turns to environmental sustainability, models that go beyond simple SoC projections to provide insights about environmentally smart driving and charging practices may become crucial.

In essence, SoC forecasting in EVs, coupled with machine learning and Federated Learning, is still in its early stages. This convergence will surely be at the center of technology and environmental developments as we move toward a greener, more sustainable future.

6 Conclusion

The constant drive for the deployment of electric vehicles (EVs) highlights the critical importance of precise State of Charge (SoC) forecasts. These forecasts not only allay ”range anxiety” for potential EV buyers, but also open the door for efficient grid management in a world that is becoming more and more electrified.

This study emphasized the transformative potential of federated learning in SoC predictions, highlighting its decentralized structure’s benefits that go beyond privacy, such as utilizing a wide variety of data sources for reliable model creation. The strategy does face certain difficulties, though, such as device dropouts, stragglers, and uneven data delivery across devices.

In-depth analysis of machine learning models indicated that while conventional approaches like Lasso and Ridge Regression have its advantages, models like the Random Forest Regressor and Gradient Boosting Machine outperformed them in terms of SoC predictions. The evaluation measures used, such as MSE, RMSE, and R2, provided a thorough assessment of model accuracy with a focus on the results’ interpretability.

Despite the fact that there is still work to be done in order to perfect SoC forecasts in EVs, this study’s illumination of the existing research landscape shows potential directions. Continuous exploration and iteration are crucial because the selection of the best model is dependent on the properties of the underlying data. The development of data driven SoC forecasting, particularly with the aid of Federated Learning, will play a crucial role in defining this electrified horizon as the world moves toward a more sustainable future with EVs at its core.

References

1. Lee, K., Lee, W. & Kim, K. Battery state-of-charge estimation using data-driven Gaussian process Kalman filters. *Journal Of Energy Storage*. **72** pp. 108392 (2023)
2. Song, C., Kim, K., Sung, D., Kim, K., Yang, H., Lee, H., Cho, G. & Cha, S. A review of optimal energy management strategies using machine learning techniques for hybrid electric vehicles. *International Journal Of Automotive Technology*. **22** pp. 1437-1452 (2021)
3. Hu, X., Li, S. & Yang, Y. Advanced machine learning approach for lithium-ion battery state estimation in electric vehicles. *IEEE Transactions On Transportation Electrification*. **2**, 140-149 (2015)
4. Chandran, V., Patil, C., Karthick, A., Ganeshaperumal, D., Rahim, R. & Ghosh, A. State of charge estimation of lithium-ion battery for electric vehicles using machine learning algorithms. *World Electric Vehicle Journal*. **12**, 38 (2021)
5. Harippriya, S., Vigneswaran, E. & Jayanthy, S. Battery management system to estimate battery aging using deep learning and machine learning algorithms. *Journal Of Physics: Conference Series*. **2325**, 012004 (2022)
6. Manoharan, A., Sooriamoorthy, D., Begam, K. & Aparow, V. Electric vehicle battery pack state of charge estimation using parallel artificial neural networks. *Journal Of Energy Storage*. **72** pp. 108333 (2023)

7. How, D., Hannan, M., Lipu, M., Sahari, K., Ker, P. & Muttaqi, K. State-of-charge estimation of li-ion battery in electric vehicles: A deep neural network approach. *IEEE Transactions On Industry Applications*. **56**, 5565-5574 (2020)
8. Zhang, D., Zhong, C., Xu, P. & Tian, Y. Deep learning in the state of charge estimation for li-ion batteries of electric vehicles: A review. *Machines*. **10**, 912 (2022)
9. Pang, S., Farrell, J., Du, J. & Barth, M. Battery state-of-charge estimation. *Proceedings Of The 2001 American Control Conference.(Cat. No. 01CH37148)*. **2** pp. 1644-1649 (2001)
10. Mashkov, V., Karova, M. & Penev, I. State Of Charge Estimation in Lithium-Ion Batteries via Machine Learning. *2022 International Conference Automatics And Informatics (ICAI)*. pp. 95-99 (2022)
11. Youssef, H., Alkhaja, L., Almazrouei, H., Nassif, A., Ghenai, C. & AlShabi, M. A machine learning approach for state-of-charge estimation of Li-ion batteries. *Artificial Intelligence And Machine Learning For Multi-Domain Operations Applications IV*. **12113** pp. 674-682 (2022)
12. Sadykov, M., Haines, S., Broadmeadow, M., Walker, G. & Holmes, D. Practical Evaluation of Lithium-Ion Battery State-of-Charge Estimation Using Time-Series Machine Learning for Electric Vehicles. *Energies*. **16**, 1628 (2023)
13. Ni, Z. & Yang, Y. A combined data-model method for state-of-charge estimation of lithium-ion batteries. *IEEE Transactions On Instrumentation And Measurement*. **71** pp. 1-11 (2021)
14. Dang, L., Yang, J., Liu, M. & Chen, B. Differential Equation-Informed Neural Networks for State of Charge Estimation. *IEEE Transactions On Instrumentation And Measurement*. (2023)
15. Maliki, A., Anoune, K., Benlafkih, A. & Hadjoudja, A. Estimating the state of charge of lithium-ion batteries using different noise inputs. *International Journal Of Power Electronics And Drive Systems (IJPEDS)*. (2024)
16. Li, H., Jin, Y. & Yu, D. Online Estimation of Battery Model Parameters and State of Charge Using Dual Time-Scaled Technique Without Open Circuit Voltage Experiment. *IEEE Transactions On Instrumentation And Measurement*. **73** pp. 1-13 (2024)
17. Hannan, M., Lipu, M., Hussain, A., Ker, P., Mahlia, T., Mansor, M., Ayob, A., Saad, M. & Dong, Z. Toward enhanced state of charge estimation of lithium-ion batteries using optimized machine learning techniques. *Scientific Reports*. **10**, 4687 (2020)
18. Hou, J., Xu, J., Lin, C., Jiang, D. & Mei, X. State of charge estimation for lithium-ion batteries based on battery model and data-driven fusion method. *Energy*. **290** pp. 130056 (2024)
19. Chen, Y., Rezapour, A. & Tzeng, W. Privacy-preserving ridge regression on distributed data. *Information Sciences*. **451** pp. 34-49 (2018)
20. McMahan, B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence And Statistics*. pp. 1273-1282 (2017)
21. Smith, V., Chiang, C., Sanjabi, M. & Talwalkar, A. Federated multi-task learning. *Advances In Neural Information Processing Systems*. **30** (2017)
22. Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions On Intelligent Systems And Technology (TIST)*. **10**, 1-19 (2019)
23. Li, T., Sahu, A., Talwalkar, A. & Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*. **37**, 50-60 (2020)
24. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal Of The Royal Statistical Society Series B: Statistical Methodology*. **58**, 267-288 (1996)
25. Hastie, T., Tibshirani, R. & Wainwright, M. Statistical learning with sparsity. *Monographs On Statistics And Applied Probability*. **143**, 8 (2015)
26. Hoerl, A. & Kennard, R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. **12**, 55-67 (1970)
27. Tay, J., Narasimhan, B. & Hastie, T. Elastic net regularization paths for all generalized linear models. *Journal Of Statistical Software*. **106** (2023)
28. Li, Y., Garg, A., Shevya, S., Li, W., Gao, L. & Lee Lam, J. A hybrid convolutional neural network-long short term memory for discharge capacity estimation of lithium-ion batteries. *Journal Of Electrochemical Energy Conversion And Storage*. **19**, 030901 (2022)

29. Hussein, H., Esoofally, M., Donekal, A., Rafin, S. & Mohammed, O. Comparative Study-Based Data-Driven Models for Lithium-Ion Battery State-of-Charge Estimation. *Batteries*. **10**, 89 (2024)
30. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Frontiers In Neurorobotics*. **7** pp. 21 (2013)
31. Xia, B., Wang, H., Tian, Y., Wang, M., Sun, W. & Xu, Z. State of charge estimation of lithium-ion batteries using an adaptive cubature Kalman filter. *Energies*. **8**, 5916-5936 (2015)
32. Attanayaka, A., Karunadasa, J. & Hemapala, K. Estimation of state of charge for lithium-ion batteries-A Review. *Aims Energy*. **7**, 186-210 (2019)
33. Steinstraeter, M., Buberger, J. & Trifonov, D. Battery and heating data in real driving cycles. *IEEE Dataport*. **10** (2020)