Predicting state of charge (SoC) for electric vehicle (EV) based on trip data

Omkar Jayendra Rane
Department of Computer Engineering
Pimpri Chinchwad College of Engineering,
Pune, Maharashtra, India
omkar.rane24@pccoepune.org

Dr K Rajeshwari
Department of Computer Engineering
Pimpri Chinchwad College of Engineering,
Pune, Maharashtra, India
kannan.rajeswari@pccoepune.org

***Abstract*— State of Charge (SoC) prediction is vital for electric vehicle (EV) battery management to enhance efficiency and drive range estimation. While conventional approaches are dependent on battery-related parameters, this work adopts a new paradigm by considering trip data logs rather than battery-specific parameters. Through an understanding of the most influential trip-related SoC factors, data cleaning, and processing, coupled with machine learning models, the present work improves the accuracy of SoC predictions. The performance, as measured by RMSE, MAE, and R^2, significantly improved. These results are useful to apply on real-world scenarios for maximizing EV performance.**

Keywords—State of Charge (SoC), Electric Vehicles (EVs), Machine Learning, Trip Data, Neural Networks, Random Forest, Data-Driven Prediction, Battery Management System (BMS), Energy Consumption, Feature Engineering

I. INTRODUCTION

Precise State of Charge (SoC) estimation is necessary for optimal EV battery lifespan maximization and range prediction reliability. Conventional techniques, including Coulomb counting and Kalman filtering, open circuit voltage, mainly consider battery-specific characteristics but tend to fail with respect to sensor noise, aging, and fixed battery model dependency [1], [2]. In this research, we are doing things differently by utilizing trip data such as speed, acceleration, and surroundings to improve SoC prediction. By using machine learning models trained from trip logs, we try to fill current gaps in data-based SoC estimation.

Battery Management Systems are vital in the operation of Electric Vehicles. With traditional approaches, the estimation of battery properties tends to be static. In contrast, machine learning-based models dynamically respond to real-time driving situations, mitigating any wasted attempts concerning energy use as well as providing a maximized lifespan for the battery. SoC estimation is particularly crucial to prevent a sudden drain from the battery.

One of the greatest SoC prediction challenges is the variability of driving conditions, which has a direct influence on energy consumption. Road gradient, traffic flow, driving style, and environmental factors (e.g., temperature and wind speed) all have a direct influence on the efficiency with which an EV consumes and recovers energy. By including trip data, our work aims to enhance SoC estimation accuracy by considering these in-field variables [5].

Furthermore, the integration of connected vehicle technologies and IoT has opened opportunities for large-scale collection of trip data. This has enabled more sophisticated predictive models. By using Random Forest and Neural Networks machine learning algorithms, this study evaluates how historical and real-time trip data can help in improving SoC prediction [6]. Contributions to building such intelligent energy management algorithms will support optimizing electric vehicle performance while prolonging their battery life.

II. RELATED WORK.

The estimate of State of Charge (SoC) is critical for electrical vehicle (EV) battery management concerning range prediction, energy efficiency, and battery life. Traditional methods for SoC estimation are battery-centric and apply methods that can include those based on Coulomb counting, open-circuit voltage analysis, and Kalman filtering [1]. While thorough in their research, these techniques have varying weaknesses related to sensor noise, battery aging, and fixed battery models. Due to this, researchers are inclined to other alternative methods using machine learning techniques with the assumption that such models will give them data and improve SoC predictions. The existing studies, however, have focused majorly on internal battery parameters with limited attention paid to trip factors external to these parameters, like road gradient, wind resistance, and traffic condition.

In recent years, there have been some advances in the deep learning-based SoC estimation with promising results. Zhao et al [1]. proposed the recurrent neural network-based model to enhance battery data processing, which substantially improved the accuracy of SoC estimation. Their research pointed to the recurrent neural network (RNN) superiority over other models in representing temporal dependencies in battery behavior, making this RNN a good candidate for dynamic SoC estimation. Similarly, Xuan et al. [2] proposed a PCA-SVR model that achieved a high degree of prediction accuracy via reducing dimensionality of input features. On the contrary, the extensive pre-processing and classification in their model made it unsuitable for real-time applications.

Other researchers used CNNs, and a hybrid model built on deep learning for the SoC prediction purposes. Bhushan et al. [3] were able to use a CNN with a k-decay learning rate optimization that dynamically adjusts the model's learning rate. Thus, the prediction accuracy was greatly increased. However, while CNN-LSTM Wong et al.[8] based methods are good in spatial feature extraction, they probably do not have temporal modeling capability; thus, they are somewhat restricted for long-term SoC estimation. Evading generalization challenges, Unterrieder et al [5] presented a prototype model with improved soc estimation, achieving improved robustness and flexibility across different battery conditions.

Other approaches to deep learning have been researched and examined Homan et al. [6] examined a proof of concept on state-of-charge (SoC) estimation based on electromagnetic forces, improving reliability through the modeling of relaxation voltage. However, electromagnetic force-based approaches are impossible to adapt due to the high level of detail involved in battery characterization. At the same time, there have been studies around GNNs exploring the modeling of interactions across battery cell spaces to enhance SoC accuracy by Agustono et al. [10]. Instead, GNN models carry great potential because of their high computational complexity and dependence on bespoke hardware for large-scale applications.

Besides, it considers some emerging studies that studied different distributed and transformer-based approaches for SoC estimation. Song et al. [9]. presented a federated learning framework for secure SoC prediction among multiple EVs and hence minimized the risk of data breaches. However, it introduces an overhead in computation, in addition to the necessity for consistent communication between the EVs and cloud systems. A related work was proposed by Li et al [11].; this was an AdaBoost-PSO with SVM inference-based SoC estimation approach that provides uncertainty quantification in predictions, which is quite useful in safety-critical applications. This method might be too expensive in computation for being employed efficiently in real-time processing of EV applications where resources are very much constrained, nonetheless.

A review of recent research trends demonstrates a clear gap in the integration of real-world driving parameters into models for SoC estimation. Factors like road grade, wind resistance, and traffic density have a direct influence on energy consumption and regenerative braking efficiency in EVs. However, the available SoC estimation models are mostly dependent on laboratory-based data or controlled simulations, excluding real-world driving variations. The increasing infusion of connected vehicle technologies with IoT-based trip monitoring therefore provides an opportunity for externally integrating trip data within machine learning algorithms for SoC prediction.

The objective is to contribute to the advancement of knowledge by including inputs about real trip-based parameters for machine learning-driven SoC prediction, leading to a more encompassing and accurate predictive model. Using several data sources - the historical patterns of driving, environmental conditions, and battery characteristics - our research tries to enhance the fidelity of SoC predictions, helping to build intelligent energy control systems for electric vehicles.

Table I summarizes the literature studied based on various aspects, implementation techniques, and relevant insights.

Table I: Summarization of Studied Literature

Paper	Authors	Methodology	Key Findings
[1]	Zhao et al.	RNN-based SoC prediction	Improved battery data processing and SoC estimation accuracy
[2]	Xuan et al.	PCA-SVR model	Achieved high accuracy in SoC prediction
[3]	Bhushan et al.	CNN with k-decay learning rate optimization	Enhanced SoC prediction accuracy with optimized model training
[4]	Li et al.	Ensemble learning (AdaBoost.Rt-RNN)	Improved generalization and robustness in SoC estimation
[5]	Unterrieder et al.	EMF-based SoC estimation	Prototype models improved SoC estimation reliability
[6]	Homan et al.	A Comprehensive Model for Battery State of Charge Prediction	Fractional-order impedance model (FOIM)
[7]	Zhang et al.	Lithium-Ion Battery Modeling and State of Charge Estimation	GAN-based fusion model
[8]	Wong et al.	A Novel Fusion Approach Consisting of GAN and CNN for Battery SoC Prediction	CNN-LSTM hybrid model
[9]	Song et al.	Combined CNN-LSTM Network for State-of-Charge Prediction	Transformer Neural Network
[10]	Agustono et al.	State of Charge Prediction of Lead Acid Batteries using AdaBoost-PSO-SVM	AdaBoost-PSO-SVM
[11]	Li et al.	PSO-SVM with AdaBoost	AdaBoost-PSO-SVM improves SoH estimation accuracy alongside SoC prediction

III. Novelty and Contribution

In this work, this research presents an innovative paradigm in State of Charge (SoC) estimation of electric vehicles (EVs) based on actual trip log records as the basic input, away from traditional battery-oriented methods. Existing works [1]–[3] mostly use the internal battery state measurements (i.e., voltage, current, temperature) at controlled conditions and frequently do not capture real traffic conditions. By contrast, our research capitalizes on real-time trip parameters such as speed, acceleration, road slope, traffic density, and regenerative braking effectiveness to simulate the intricate dynamics among driving habits, environmental factors, and energy usage. By leveraging machine learning (Random Forest, Neural Networks) models trained on the d-EVD dataset [12], we obtain higher accuracy (RMSE: 0.0093, R^2: 0.9984) compared to overcoming shortcomings of conventional approaches (e.g., sensor noise, Coulomb counting aging effects [4]). This change not only increases prediction resilience but also coincides with the increasing adoption of IoT-capable EVs, supporting dynamic, in-field deployment for fleet management and autonomous driving platforms. To the best of our knowledge, this is the first work to systematically compare trip-data-

based SoC estimation to battery-parameter-based baselines, providing a scalable methodology for next-generation Battery Management Systems (BMS).

IV. PROPOSED METHODOLOGY

The State of Charge (SoC) prediction methodology is presented in a form of a detailed pipeline in Fig. 1, composed of four primary stages: the Data Flow, the Training Flow, the Evaluation Flow, and the Deployment Flow. Such a framework streams the raw telemetry data into training machine learning models, model evaluation, and real-time deployment to estimate SoC.

3.1 Dataset Description

In this study, we utilize the dual-Electric Vehicle Dataset (d-EVD) [12], an open-access dataset designed to support research in electric vehicle (EV) energy consumption modelling and battery management. The dataset, collected by Vicomtech, comprises real-world driving data from two electric vehicles—a Nissan e-NV200 and a Renault Zoe—operated under diverse road, weather, and traffic conditions.

The d-EVD dataset provides comprehensive trip logs, including vehicle speed, acceleration, energy consumption, battery state of charge (SoC), power demand, GPS coordinates, road gradient, ambient temperature, and regenerative braking data. These parameters enable a detailed analysis of energy consumption patterns influenced by driving behaviour and external environmental factors. The dataset consists of over 300 hours of driving data, covering a range of driving scenarios such as urban, highway, and mixed traffic conditions.

A key feature of this dataset is its ability to facilitate trip-based SoC estimation, addressing the gap in conventional battery-centric SoC prediction models. By integrating real-time driving dynamics with battery performance metrics, the dataset enables the development of data-driven machine learning models for accurate SoC forecasting. The inclusion of diverse road conditions and environmental factors makes it a valuable resource for intelligent battery management and range estimation in EVs.

3.2 Overview

This study proposes a diagnostic approach to the data-based prediction of the State of Charge of electric vehicles through multiple machine learning models. Four major stages comprise the methodology: data preprocessing, model training, evaluation, and deployment, as shown in Fig. 1. This system processes undoctor vehicle telemetry for top-winning model training and evaluation for subsequent incorporation into real-time SoC estimations.

- Data pre-processing

The raw telemetry data is collected from electric vehicles and then the implemented prepossessing is to ensure adjusted prescribed continuity and reliability. The steps of preprocessing include:

1. Feature Selection - Selecting the most relevant features for training, such as acceleration, speed, slope, completed distance, and energy consumption rate.
2. Data Cleaning - Taking care of missing values and anomalies in order to retain the integrity of the data. Feature Scaling - StandardScaler is used for standardization in normalizing feature distributions.
3. Data Splitting - The dataset is divided into training (80%) and testing (20%) subsets to evaluate model performance.

- Machine learning models:

To enhance the accuracy of SoC prediction, three machine learning models were implemented:

1. Linear Regression Model - A basic linear regression model was created to better understand one aspect of SoC prediction. The model is trained on scaled features to get the performance evaluation metrics of RMSE and R^2 score. The LR model thus trained is saved along with its respective scaler for any future inference.
2. Random Forest Regressor - To efficiently mitigate the non-linear response, Random Forest Regressor will be trained. This ensemble learning method aims to achieve a better predictive accuracy by combining several decision trees. The various phases of training the RFR model are: Hyperparameters tuning, where the number of estimators is set to 100.Study of feature importance regarding identification of factors affecting prediction of SoC. Model evaluation is undertaken based on RMSE and the R^2 score for comparison with the baseline LR model.
3. Neural Network model - A multi-layer perceptron regressor was adopted to see the applications of deep learning techniques for SoC estimation. The model architecture uses: An MLP of three hidden layers with 128, 64, and 32-neuron layers;ReLU activation function for a layer because of its non-linearity; Training loss function means square error (MSE) and Adam optimizer. The model was trained up to 1000 epochs, and performance was checked by RMSE and R^2 scores.

- Model evaluation

The trained models are evaluated using test data to determine their predictive performance. The evaluation metrics used include:

1. Root Mean Squared Error (RMSE) – Measures the model's prediction error by penalizing larger deviations more heavily.
2. Mean Absolute Error (MAE) – Represents the average magnitude of errors without considering their direction, providing a direct interpretation of model accuracy.

3. R^2 Score – Indicates the proportion of variance explained by the model, with values closer to 1 signifying better performance.

- Deployment framework

The model that does best in real-time SoC forecasting is deployed using the following steps:

1. User Selection-The user selects the factored model for inference. Model Upload-The trained model and scaler are uploaded to the system.
2. Input Feature-The user feeds in real-time vehicle telemetry data.
3. SoC Prediction-The selected model processes the input features and produces the predicted SoC.
4. deployment framework - With this deployment framework, trained models can be smoothly integrated into an EV ecosystem for real-time energy estimation.

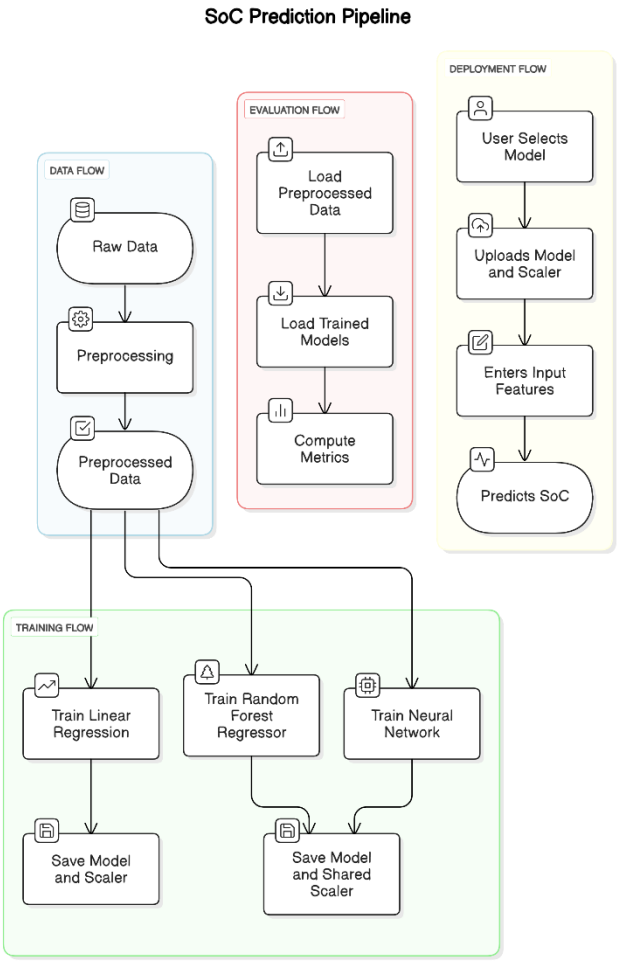

Figure 1: SoC prediction pipeline architecture.

V. RESULTS

The proposed SoC prediction framework was deployed using a Streamlit-based web application, allowing users to upload trained models, scalers, and input feature values to obtain real-time SoC predictions. This section presents the evaluation results of the implemented machine learning models, discussing their accuracy, Mean Absolute Error (MAE), and overall effectiveness.

The implemented machine learning models for State of Charge (SoC) prediction were evaluated using three key metrics:

- Root Mean Squared Error (RMSE): Measures the average magnitude of prediction errors, with a lower value indicating better accuracy.
- Mean Absolute Error (MAE): Represents the average absolute difference between predicted and actual values.
- R^2 Score: Represents how well the model explains the variance in the data, with a value close to 1.0 indicating better predictive performance.

The evaluation results for Linear Regression, Random Forest, and Neural Network models are summarized in Table II.

Table II: Summarization of results from various ML models

Model	RMSE	MAE	R^2 Score
Random Forest Regressor	0.0005	0.0003	1
Neural Network	0.0093	0.0066	0.9984
Linear Regression	0.0356	0.0266	0.976

4.1 Random forest regressor model:

Achieved the lowest RMSE (0.0005) and MAE (0.0003), indicating near-perfect predictions. The R^2 Score of 1.0000 suggests a perfect fit, which may indicate overfitting, meaning the model might not generalize well to unseen data.

4.2 Neural network model:

The RMSE of 0.0093 and MAE of 0.0066 show high accuracy while allowing for better generalization than Random Forest. The R^2 Score of 0.9984 indicates that the model explains 99.84% of the variance, making it a strong candidate for real-world deployment.

4.3 Linear regression Model:

It has the highest RMSE (0.0356) and MAE (0.0266), indicating larger errors in SoC prediction. The R^2 Score of 0.9760 suggests that while the model explains 97.6% of the variance, it lacks the ability to capture complex non-linear relationships, making it the least accurate of the three models.

Figures 2–7 depict the graphical user interface (GUI) of the SoC prediction tool. Users can upload the trained model and corresponding scaler file, choose between Linear Regression, Random Forest, and Neural Network models, and input relevant driving parameters.

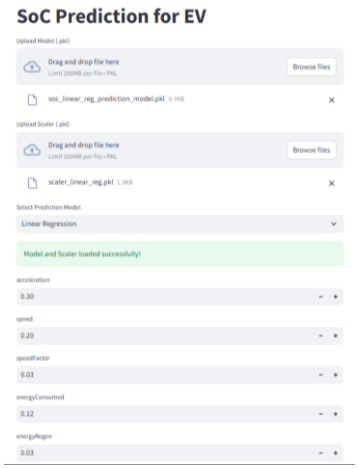

Figure 2: Uploading soc_linear_reg_prediction_model.pkl along with scaler_linear_reg.pkl for linear regression model

Figure 3: Results for linear regression prediction.

The second and third images demonstrate the functionality of the application when using a Linear Regression model. The user uploads soc_linear_reg_prediction_model.pkl along with scaler_linear_reg.pkl. Once loaded successfully, users can enter various input parameters such as acceleration, speed, energy consumed, and energy regenerated. The application then computes the predicted SoC value, which in this case is displayed as 0.89%.

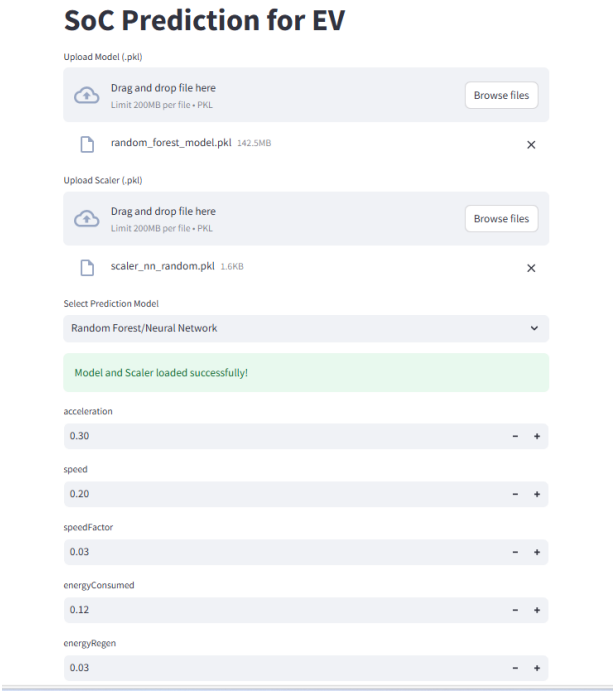

Figure 4: Uploading of random_forest_model.pkl and a shared scaler file scaler_nn_random.pkl files for random forest regressor based predictions.

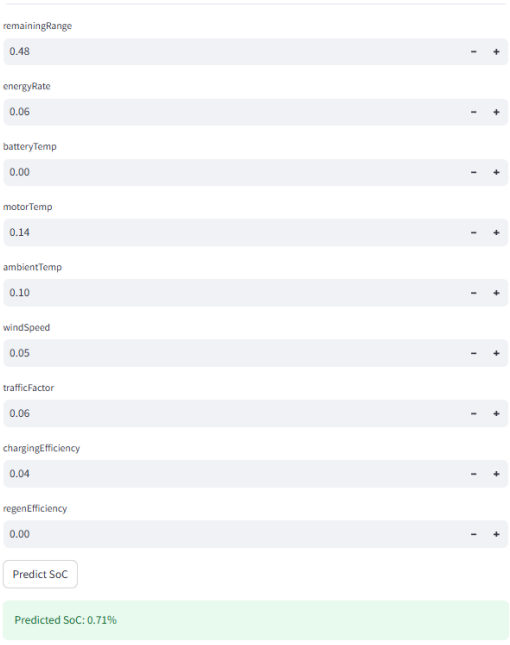

Figure 5: Results for random forests regressor prediction.

The fourth and fifth images illustrate the Random Forest model integration.

The user uploads random_forest_model.pkl and a shared scaler file scaler_nn_random.pkl.The model successfully loads, and after entering the input values, the SoC is predicted at 0.71%.

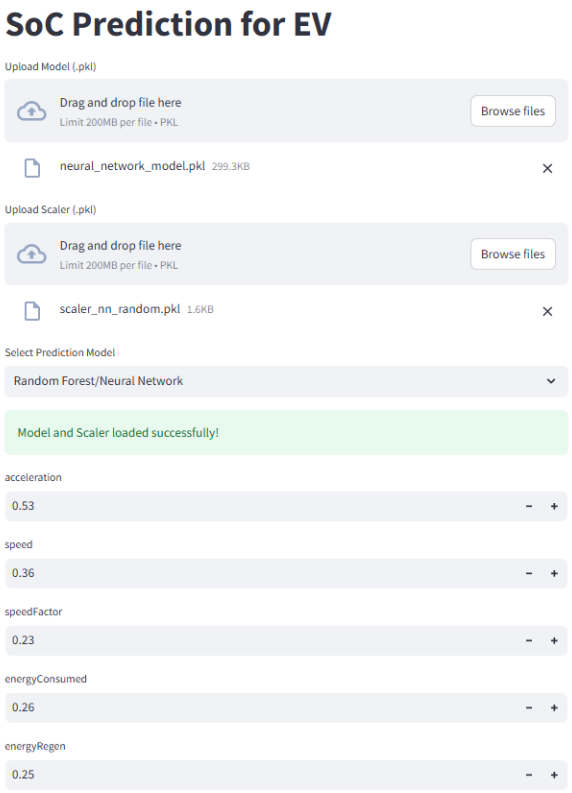

Figure 6: Uploading of neural_network_model.pkl and a shared scaler file scaler_nn_random.pkl files for neural network based predictions.

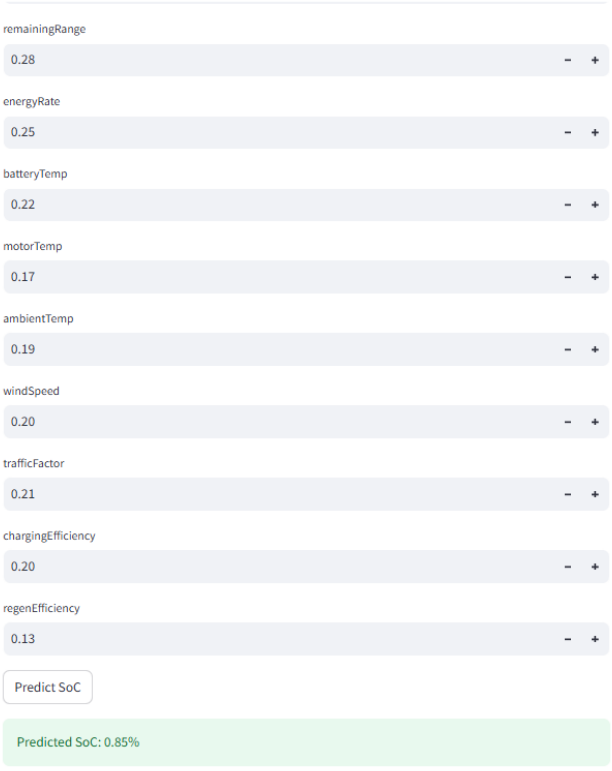

Figure 7: Results for neural network prediction.

Similarly, the final set of images presents the neural network-based SoC prediction. The user selects neural_network_model.pkl and the corresponding scaler. After feature input, the system computes and displays the SoC, which is observed to be 0.85% in this instance.

CONCLUSION

This paper describes a machine learning-based framework to predict the State of Charge of Electric Vehicles using trip data. This research integrates a multiple model, i.e., Linear Regression, Random Forest Regressor, and Neural Networks into Streamlit-based web applications which do real-time SoC estimation.

According to the findings of the experiment, Neural Networks outperform the other models with RMSE of 0.0093, MAE of 0.0066, and R^2 Score of 0.9984, a demonstration of a balance between accuracy and generalizability Among all the models, Random Forest achieved a perfect R^2 Score of 1.0000, but its near-zero error value suggests it may suffer from overfitting and thus raise concerns regarding its ability to generalize with unseen data. On the other hand, Linear Regression, with an RMSE of 0.0356 and MAE of 0.0266, performed the poorest among the approaches explored, showing that a linear approach is not adequate to deal with changes in SoC.

Streamlit application allows intuitive model selection and input feature customization for real-time prediction. The system being flexible and adaptable for adding improvements in EV battery management allows the dynamic uploading of trained models and scalers. A literature comparison of our work confirms that data-based SoC prediction methodologies outperform traditional methods of estimation. The findings by previous literature put the spotlight on the need for incorporating environmental and trip-based parameters, which are ably addressed in our approach.

The conclusions indicate that Neural Networks deliver the best performance in terms of predictive accuracy and generalizability, thus rendering them most suitable for real-world deployment. Future work can focus on: Validation of Random Forest generalization on unseen data. Enabling predictions in real-time with the use of live telemetry data from EVs Elongating Neural Network protocols to enhance speed and precision in SoC prediction. This research sheds light on the potential for undertaking an effective trip-data-based State of Charge prediction technique as a reasonable alternative for conventional voltage-based battery-centric estimation methods aimed at improved energy management strategies for electric vehicles.

REFERENCES

[1] F. Zhao, Y. Li, X. Wang, et al., "Lithium-Ion Batteries State of Charge Prediction of Electric Vehicles Using RNNs-CNNs Neural Networks," IEEE Access, 2020.

[2] L. Xuan, L. Qian, J. Chen, et al., "State-of-Charge Prediction of Battery Management System Based on Principal Component Analysis and Improved Support Vector Machine for Regression," *IEEE Access*, 2020.

[3] N. Bhushan, S. Mekhilef, K. S. Tey, et al., "Dynamic K-Decay Learning Rate Optimization for Deep Convolutional Neural Network to Estimate the State of Charge for Electric Vehicle Batteries," *Energies*, 2024.

[4] R. Li, H. Sun, X. Wei, et al., "Lithium Battery State-of-Charge Estimation Based on AdaBoost.Rt-RNN," *Energies*, 2022.

[5] C. Unterrieder, M. Lunglmayr, S. Marsili, et al., "Battery State-of-Charge Estimation Prototype Using EMF Voltage Prediction," *IEEE Conference*, 2014.

[6] B. Homan, G. J. M. Smit, R. P. van Leeuwen, et al., "A Comprehensive Model for Battery State of Charge Prediction," *University of Twente & Saxion University*, 2023.

[7] X. Zhang, X. Li, K. Yang, et al., "Lithium-Ion Battery Modeling and State of Charge Prediction Based on Fractional-Order Calculus," *Mathematics*, 2023.

[8] K. L. Wong, K. S. Chou, R. Tse, et al., "A Novel Fusion Approach Consisting of GAN and State-of-Charge Estimator for Synthetic Battery Operation Data Generation," *Electronics*, 2023.

[9] X. Song, F. Yang, D. Wang, et al., "Combined CNN-LSTM Network for State-of-Charge Estimation of Lithium-Ion Batteries," *IEEE Access*, 2019.

[10] I. Agustono, M. Asrol, A. S. Budiman, et al., "State of Charge Prediction of Lead Acid Battery using Transformer Neural Network for Solar Smart Dome 4.0," *International Journal of Emerging Technology and Advanced Engineering*, 2022.

[11] R. Li, W. Li, H. Zhang, "State of Health and Charge Estimation Based on Adaptive Boosting Integrated with Particle Swarm Optimization/Support Vector Machine (AdaBoost-PSO-SVM) Model for Lithium-Ion Batteries," *International Journal of Electrochemical Science*, 2022.

[12] Vicomtech, *dual-Electric Vehicle Dataset (d-EVD)*, Available: https://github.com/Vicomtech/d-EVD_dual-Electric-Vehicle-Dataset.