A Comparative Study of Electric Vehicles Battery State of Charge Estimation Based on Machine Learning and Real Driving Data

Salma Ariche, Zakaria Boulghasoul, Abdelhafid El Ouardi, Abdelhadi Elbacha, Abdelouahed Tajer, Stéphane Espié

To cite this version:

Salma Ariche, Zakaria Boulghasoul, Abdelhafid El Ouardi, Abdelhadi Elbacha, Abdelouahed Tajer, et al.. A Comparative Study of Electric Vehicles Battery State of Charge Estimation Based on Machine Learning and Real Driving Data. Journal of Low Power Electronics and Applications, 2024, 14 (4), pp.59. ⟨10.3390/jlpea14040059⟩. ⟨hal-05011417⟩

HAL Id: hal-05011417

https://hal.science/hal-05011417v1

Submitted on 25 Apr 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Journal of
Low Power Electronics and Applications

MDPI

Article

A Comparative Study of Electric Vehicles Battery State of Charge Estimation Based on Machine Learning and Real Driving Data

Salma Ariche [1,2,*], Zakaria Boulghasoul [2], Abdelhafid El Ouardi [1,*], Abdelhadi Elbacha [2], Abdelouahed Tajer [2] and Stéphane Espié [3]

[1] Université Paris-Saclay, ENS Paris-Saclay, CNRS, SATIE, 91110 Gif-sur-Yvette, France
[2] Engineering Systems and Applications (LISA), Cadi Ayyad University, Marrakech 40000, Morocco; z.boulghasoul@uca.ma (Z.B.); a.elbacha@uca.ac.ma (A.E.); a.tajer@uca.ac.ma (A.T.)
[3] Université Gustave Eiffel, SATIE, 91190 Gif-sur-Yvette, France; stephane.espie@univ-eiffel.fr
* Correspondence: salma.ariche@universite-paris-saclay.fr (S.A.); abdelhafid.elouardi@universite-paris-saclay.fr (A.E.O.)

Abstract: Electric vehicles (EVs) are rising in the automotive industry, replacing combustion engines and increasing their global market presence. These vehicles offer zero emissions during operation and more straightforward maintenance. However, for such systems that rely heavily on battery capacity, precisely determining the battery's state of charge (SOC) presents a significant challenge due to its essential role in lithium-ion batteries. This research introduces a dual-phase testing approach, considering factors like HVAC use and road topography, and evaluating machine learning models such as linear regression, support vector regression, random forest regression, and neural networks using datasets from real-world driving conditions in European (Germany) and African (Morocco) contexts. The results validate that the proposed neural networks model does not overfit when evaluated using the dual-phase test method compared to previous studies. The neural networks consistently show high predictive precision across different scenarios within the datasets, outperforming other models by achieving the lowest mean squared error (MSE) and the highest R^2 values.

Keywords: battery electric vehicles; state of charge; online estimation; real driving data; machine learning; neural networks

check for updates

Citation: Ariche, S.; Boulghasoul, Z.; El Ouardi, A.; Elbacha, A.; Tajer, A.; Espié, S. A Comparative Study of Electric Vehicles Battery State of Charge Estimation Based on Machine Learning and Real Driving Data. *J. Low Power Electron. Appl.* **2024**, *14*, 59. https://doi.org/10.3390/jlpea14040059

Academic Editor: Giorgos Dimitrakopoulos

Received: 29 October 2024
Revised: 26 November 2024
Accepted: 9 December 2024
Published: 11 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Today's automotive sector is moving towards intelligent transportation, which is emerging as a pivotal global industry, wielding significant influence not only on the economic front but also in the realm of research and development. A growing array of technological components are being integrated into vehicles to enhance safety and comfort.

In light of the critical imperative to combat climate change and lower air pollution levels [1], the environmental benefits of using electric vehicles (EV) are especially significant. As governments and policymakers worldwide intensify their efforts to meet these challenges, promoting EVs has become a transition towards decarbonization of the transportation system since manufacturers are electrifying their fleets to meet new legal requirements. The shift to EVs is seen as a technical advancement and a crucial step toward sustainable transportation. Advances in battery technology, charging infrastructure, and regulations supporting low-emission cars contribute to this change [2]. However, despite the increased supply of EVs and improving infrastructure related to EV charging, vehicle registration statistics indicate that customers still need to embrace EVs [3].

The EV is characterized by its low noise emissions, user-friendly operation, and absence of fuel-related expenses commonly associated with traditional vehicles. Despite these advancements, their limited operating range remains a significant barrier to the broad acceptance of battery electric vehicles (BEVs). The reluctance to fully embrace BEVs is partly due to current battery technology, which is rapidly evolving yet still does not

match the energy density and recharging speed of traditional gasoline engines [4]. Thus, the precise assessment of the SOC of batteries has become indispensable for drivers and the energy regulation systems of these vehicles. The SOC is a vital indicator of a battery's residual charge, which not only informs the potential driving range before the necessity of recharging arises but also aids in mitigating risks associated with overcharging and undercharging [5]. It is similar to the gasoline gauge in traditional automobiles but is more complex to estimate accurately due to the nature of battery chemistries [6]. Over the last decades, extensive scientific research has focused on determining the SOC of batteries. Owing to the complex and non-linear electrochemical properties of batteries, their performance hinges on a variety of factors, both intrinsic and environmental. Methods for estimating the SOC leverage available data through various techniques, including open circuit voltage (OCV), the Kalman filter (KF), and artificial intelligence (AI) approaches. It is also critical to evaluate not just individual battery performance but the collective performance of the entire battery assembly. Moreover, challenges like battery degradation, lifespan, temperature variations, among others, present considerable difficulties in precisely gauging the SOC [7].

In the past few years, machine learning (ML) has become an essential tool in many areas, including the automotive industry [8,9] and energy management systems (EMS) [10,11]. Using machine learning to estimate the SOC of BEVs is an important step forward in solving the problem of range anxiety. ML algorithms can learn from driving data, patterns of battery use, and environmental factors. They provide a more accurate and reliable SOC estimation than traditional methods. This enhancement is critical in improving the user experience by providing more precise range predictions and a more comfortable ride, thus fostering greater confidence and adoption of EVs among potential users. ML used in SOC estimations opens up smart energy management possibilities for BEVs. By accurately determining the SOC, the EMS can use the battery best, improve the vehicle's performance, and even make the battery last longer through efficient charging and discharging strategies [12]. Therefore, precise SOC estimation not only contributes to the user's convenience but also the vehicle's overall safety. By incorporating ML-based SOC estimation into intelligent transportation systems (ITS), BEVs can establish communication with charging stations, traffic management systems, and other vehicles in order to optimize traffic flow, charging schedules, and driving routes. The research work presented in the current paper focuses on the estimation of the battery electric vehicle SOC using data-driven approaches like regression models and artificial neural networks, as well as trained real-world driving data derived from two different BEVs.

The rest of this paper is organized as follows: Section 2 provides a comprehensive review of current methods for SOC estimation. Section 3 details the methodology we propose. Section 4 discusses the experimental results. The paper concludes with Section 5, which offers conclusions and future outlooks of this study.

2. Related Work

In general, battery SOC is characterized as the fraction of the remaining energy relative to the total energy capacity it can store (1). With Q_c representing the residual energy and Q_0 denoting the initial energy available under specific conditions, Q_T is the amount of energy that the battery has previously discharged. The SOC can, therefore, be mathematically expressed using these variables.

$$SOC = \frac{Q_c}{Q_0} = 1 - \frac{Q_T}{Q_0} \tag{1}$$

However, this mathematical representation of the battery's available capacity does not fully capture its practical variability, as SOC is frequently affected by a variety of conditions [13]. The battery SOC can be calculated using several methods, the most common of which involve the integration of current over time and voltage-based models. The methods for estimating SOC can be categorized into three groups:

a. Conventional Methods

Approaches that heavily depend on battery parameters, such as the straightforward look-up table methods [14], are primarily represented by the open-circuit voltage (OCV), which assumes that there is a linear relation between the voltage, the SOC and the ampere-hour (Ah) integration method. Yet, these methods exhibit a substantial estimation error and may not be well-suited for real-time estimation across different battery types. The association between the SOC and the open-circuit voltage (OCV) depends on the battery's type and capacity. A lead-acid battery, for instance, shows a direct linear relationship between SOC and OCV, unlike lithium-ion batteries with more complex and non-linear connections [15].

b. Model Based Methods

Model-based approaches were presented to overcome the limitations of traditional methods that rely solely on battery parameters. The main goal of model-based SOC estimation lies in utilizing the battery model to establish a connection between battery signals. Model-based techniques such as the Electrochemical Impedance Model (EIM) and electrochemical impedance spectroscopy (EIS) [16,17] are used to discern the relationship between a battery's impedance and its state of charge (SOC). Alternatively, linking the model to cellular measurements through the use of a filtering algorithm derived from current control theory can establish a closed-loop system for estimation. This method capitalizes on the variances between the modeled and actual battery outputs to refine the state predictions, thereby reducing the discrepancy between the modeled battery and the measured signals. The Kalman filter (KF) is a popular filter-based model; however, it is applicable only to linear systems. The extended Kalman filters (EKF) and Sigma-Point Kalman filters (SPKF) extend this capability to nonlinear systems [18–20]. The KF and its variants operate on the premise that all noises are characterized by zero mean and follow a Gaussian distribution. In contrast, the adaptive Kalman filter (AKF) estimates both the model and the noises online, while the adaptive extended Kalman filter (AEKF) is utilized to enhance state estimation accuracy [21].

c. Data Driven Methods

The inaccuracy of SOC estimation using model-based or traditional methods is still high when comparing the estimated output and the actual SOC recorded in the datasets. Furthermore, they are methods that are time-consuming and have low efficiency. Data-driven methods are established to accurately estimate the SOC based on recorded data; we are talking about artificial intelligence (AI) algorithms that take information from large datasets for model training [22]. Machine learning (ML) models forecast battery SOC based on time-series datasets, including features that could impact the SOC over time. SOC estimation often employs regression algorithms like linear regression (LR), support vector machine (SVM), support vector regressor (SVR), Gaussian process regression (GPR), random forest regression (RF), and decision tree regression (DT), as documented in various references [23–26]. The works of [27,28] introduced a hybrid SOC estimation approach combining model-based and data-driven methods for better accuracy. Deep learning (DL) or neural networks (NNs) in research have been shown to give better results related to SOC estimation [29]. An artificial neural network (ANN) model is created to estimate the SOC accurately. The estimation relies on load classification, which includes post-processing and boosting techniques to prevent overfitting, as discussed in [30]. Additionally, a comparative study of DL algorithms for lithium-ion SOC estimation was carried out in the works of [31,32]. The papers [33–35] also introduce the use of a long short-term memory recurrent neural network (LSTM-RNN) for SOC estimation. In [36], a deep multilayer perceptron network was developed and assessed using different activation functions, including the Sigmoid, tanh, ReLU, and Mish activation functions. Additional studies have focused on analyzing the effects of HVAC on electric vehicles, distributed power generation, and electrical storage using AI for accurate SOC estimation and to

improve range and battery lifetime [37]. Statistical analysis using ARIMA models and LTSM was dedicated to SOC forecasting in micro-grid systems [38].

While offering enhanced accuracy and adaptability in SOC estimation, data-driven methods also present several challenges. One significant drawback is their dependency on vast and high-quality datasets, which may not always be available or may require extensive preprocessing to be valid. These methods can also be computationally intensive, leading to longer processing times and requiring more powerful hardware, which can increase the cost and complexity of implementation. These methods may need to be more balanced, especially in scenarios where the variability in the training data do not accurately reflect real-world conditions, which can lead the models to overfit.

In this research, we evaluate the batteries' state of charge (SOC) estimation using real-world driving data, a notable departure from the controlled laboratory conditions in the existing literature. To comprehensively understand EV energy consumption, we underscore the importance of incorporating diverse driving scenarios and auxiliary components, such as HVAC systems, weather conditions, and specific route information. By leveraging datasets from European (Germany) and African (Morocco) contexts, we address the complexities of real-world EV operation, particularly highlighting the underexplored challenges of African road conditions. Furthermore, we compare four machine learning models and introduce a second testing phase to combat overfitting, ensuring model robustness and reliability across various driving conditions. This dual-phase testing enhances the practical applicability of our models, marking a significant contribution to the field.

3. Methodology

In a previous study [39], the paper analyzed how the EV's energy consumption could be affected by factors like road topography and ambient temperature when switching to comfort ride mode on Moroccan roads. However, the evaluation was insufficient for an accurate BEV range estimation in a real-life scenario, whether for a regular or a comfort ride option. Building upon the EV model in that study, the research was extended by evaluating the SOC estimation, including two crucial energy consumers as presented in Figure 1.

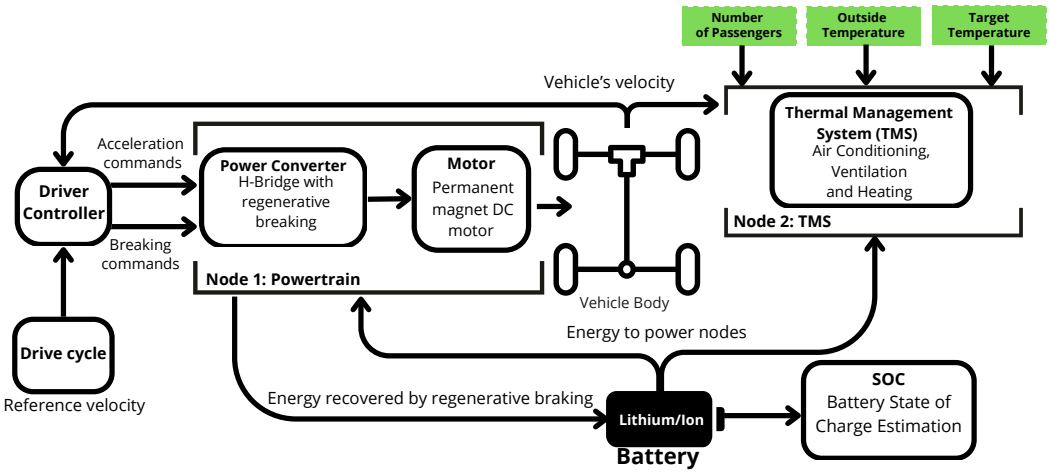


Figure 1. BEV model with powertrain and thermal management system.

The block diagram illustrates an electric vehicle's system, where the driver controller interprets the driver's input via the drive cycle to manage the vehicle's speed. The powertrain includes a power converter responsible for modulating the power from the battery to drive the motor, which in turn propels the vehicle. Additionally, the thermal management system (TMS) from [40–42], regulates the temperature for optimal performance and comfort, encompassing the HVAC system that handles air conditioning, ventilation, and heating within the vehicle's cabin and its components. The TMS takes as input parameters the target

cabin temperature, the outside temperature, the number of passengers, and the vehicle's speed. Finally, the lithium-ion type battery is central to the system, providing energy to the powertrain and TMS and recovering energy through regenerative braking. The battery block uses KF and EKF for SOC estimation. The estimation results using the model-based method were not as satisfying as expected, with a significant difference when comparing the simulation output and the actual recorded SOC in the datasets. This difference might be explained by several limits related to the MATLAB/SIMULINK existing model adaptation, as listed below:

(i) **Model Accuracy:** It can sometimes encourage idealized modeling that does not take into account all the non-linearities and uncertainties present in real systems.
(ii) **Documentation:** It may sometimes be insufficient, particularly for advanced functionalities or custom blocks.
(iii) **Limited Customization:** In some cases, it is impossible to introduce input parameters.
(iv) **Simulation parameters:** Managing simulation parameters to guarantee test reproducibility is a challenge when using real datasets.
(v) **Performance:** The simulated BEV is a model with a large number of Simulink and Simscape blocks, making simulation time-consuming and consuming a lot of system resources.

As an alternative approach, in this section, we present the chosen data-driven methods used to estimate the battery SOC, tested on actual data from a Moroccan and European context. The models were trained and tested with Python 3.10.12 version using ML libraries like scikit-learn (sklearn), Tensorflow, and Keras [43–45]. The flow chart in Figure 2 illustrates the process of SOC forecasting using ML algorithms.

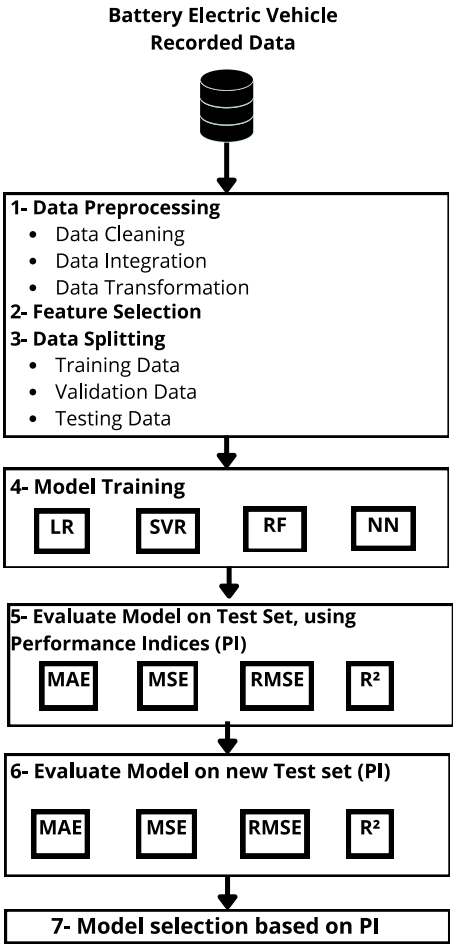


Figure 2. Flow chart of the proposed data-driven methods to estimate battery SOC.

Initially, data preprocessing is imperative to refine the collected data, involving cleaning to remove noise or irrelevant information, transformation to ensure consistency, and feature selection to identify the most relevant parameters for SOC estimation. Selecting appropriate features is crucial as it directly influences the model's accuracy and computational efficiency. The choice of algorithms—linear regression (LR), support vector regression (SVR), random forest (RF), and neural networks (NNs)—strikes a balance between simplicity and precision. LR, with its straightforward computation, offers a baseline for performance. SVR was selected because of its capacity to handle non-linear relationships. RF is included for its robustness to overfitting and its ability to rank the importance of features. At the same time, NNs were chosen because of their high potential for accuracy in complex relationships. Performance indices like mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R^2 are utilized to evaluate the models rigorously on an initial and second test sets. These metrics give us a comprehensive view of the model's accuracy and the variance explained by the model [46]. The second test set is additional recorded trips with the same number of features. Including a secondary test set, sourced from different operational contexts, is a strategic measure to assess the generalizability of the models and safeguard against overfitting. This dual-phase testing ensures that the algorithms perform well across various scenarios, which is essential for reliable SOC estimation in the dynamic environment of electric vehicles. The algorithms enlisted for SOC estimation in battery electric vehicles are detailed as follows:

1. **Linear Regression:** A statistical model that describes a linear relationship between dependent and independent variables. LR is a supervised ML algorithm applied to models with continuous linear variable output. It is often used for its simplicity and performance when tested on small- or medium-sized datasets. The empirical equation of the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (2)$$

where y is the predicted variable, $x_1, x_2, \ldots, x_n$ are the input feature values, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients, and ϵ is the error term as presented in Algorithm 1.

Algorithm 1: Linear Regression

Input: Dataset with predictor variable x and target variable y
Output: Model coefficients β_0 and β_1
Define the model: $y = \beta_0 + \beta_1 x + \epsilon$;
Estimate coefficients using least squares;
Evaluate the model using R-squared, MSE, and MAE;
Make predictions: Use the model to predict y for new x values;

2. **Support Vector Regression:** SVR, or support vector regressor, is a supervised learning algorithm employed for addressing various nonlinear classification and regression challenges. It operates on a principle similar to the support vector machine (SVM) but focuses on predicting the value of the output rather than classifying it into categories. As in Algorithm 2, SVR tackles regression problems by identifying a hyperplane within a high-dimensional feature space that optimally captures the connection between input variables (features) and the target variable (output). It proves to be a robust tool for regression tasks, particularly when dealing with datasets characterized by intricate relationships between variables.

Algorithm 2: Support Vector Regression (SVR)

Input: Training data, chosen kernel function, hyperparameters C and ϵ
Output: SVR model for prediction
Select kernel function (e.g., linear, polynomial, RBF);
Define the SVR model with the chosen kernel and hyperparameters C and ϵ;
Formulate the optimization problem to find the hyperplane that minimizes error within the ϵ margin;
Solve the optimization problem;
Identify support vectors;
Compute predictions;
Apply inverse transformations to predictions if needed;
Evaluate model performance using regression metrics;
Fine-tune hyperparameters (e.g., C and ϵ) using cross-validation;

3. **Random Forest regression:** An ensemble learning strategy. Algorithm 3 functions by generating several decision trees during the training process and deriving a collective prediction from the outputs of these individual trees. This technique focuses on the aggregated knowledge obtained from various trees instead of depending on a solitary model, thereby achieving a more robust and more dependable prediction through the mean of the outcomes from every tree that forms part of the ensemble.

Algorithm 3: Random Forest Regression

Input: Training data
Output: Ensemble model for prediction
Bootstrap Sampling: Create multiple random subsets of the data;
Build Decision Trees: Grow decision trees using different subsets of features;
Aggregate Predictions: Average the predictions from all decision trees;
Predict new data points using the ensemble model;

4. **Neural Networks (NNs):** These models are used to describe a non-linear or complex interaction using a set of input and output data. They consist of layers of interconnected nodes where each node performs a simple computation. The connections between neurons have associated weights tuned during the learning process as described in Algorithm 4. Different NN architectures exist in the literature, such as:
   (a) Feedforward Neural Networks (FNNs): The simplest architecture of NNs is a straightforward connection between the NN nodes. It includes the basic perceptron and multi-layer perceptrons (MLPs).
   (b) Convolutional Neural Networks (CNNs): Based on convolution operations for data processing as a matrix-like topology, they are primarily used in image processing and computer vision.
   (c) Recurrent Neural Networks (RNNs): Unlike FNNs, RNNs have connections that form cycles between nodes. These cyclical connections are designed to retain a form of memory, making RNNs particularly well-suited for applications involving sequential data. It includes, but is not limited to, time series analysis and natural language processing tasks.
   (d) Long Short-Term Memory (LSTM) Networks: A specialized variant of RNNs designed to learn and retain long-term dependencies in data sequences.

Algorithm 4: Artificial Neural Network (ANN)

Input: Training data with input and output variables
Output: Trained neural network model for predictions
Design Network Structure: Define layers and neurons;
Initialize weights and biases;
foreach *epoch* **do**
    Forward Propagation: Pass data through the network;
    Calculate Loss: Compute the difference between output and actual values;
    Backpropagation: Adjust weights and biases based on the loss;
Evaluate model performance on validation data;
Make predictions on new data;

The chosen PIs for the algorithm evaluation are listed in Table 1.

Table 1. Overview of Error Metrics.

Error Metric	Description	Equation
MAE (Mean Absolute Error)	A metric that measures the difference between two variables that are continuous. It is determined by computing the mean of the absolute variations between the forecasted values and the actual observed data. Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.	$\text{MAE} = \frac{1}{n}\sum\lvert y_i - \hat{y}_i\rvert$
MSE (Mean Squared Error)	Computes the average of the squared differences, representing the variance between the predicted and actual values. Due to the squaring of errors, MSE is more sensitive to outliers compared to the MAE. Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.	$\text{MSE} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$
RMSE (Root Mean Squared Error)	Quantifies the average magnitude of residuals, emphasizing larger errors more significantly. RMSE is particularly useful in scenarios where large errors are especially undesirable. Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n signifies the total number of observations	$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$
R^2 **(Coefficient of Determination)**	This statistical metric reflects the proportion of variance in the dependent variable that can be attributed to the independent variables in a regression analysis. It evaluates the adequacy of the model, demonstrating the extent to which the predicted values correspond with the actual data. A perfect R^2 of 1 means the predictions match the data exactly. Where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of actual values, and n is the number of observations.	$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$

For a comprehensive evaluation of the model's performance, the inclusion of MAE, MSE, and RMSE as complementary metrics is justified by:

- MAE provides an average measure of absolute error.
- MSE emphasizes more significant errors by penalizing them quadratically.
- RMSE offers an interpretable error measure in the same units as the target variable, allowing us to analyze the models from different perspectives.

4. Experimental Results

In this section, we estimate the EV battery SOC based on data-driven approaches by considering impact factors that can significantly influence energy consumption. We use two different datasets [47,48] under two different contexts:

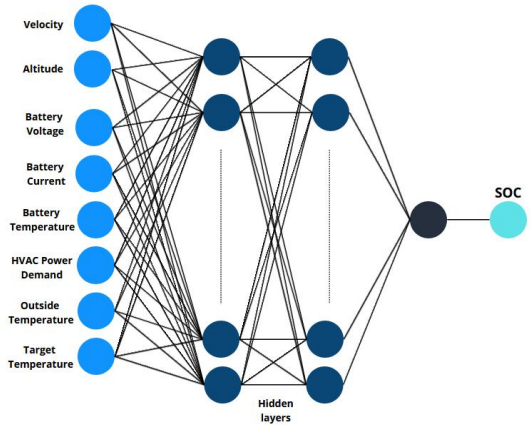
- Dataset 1: Referred to as Munich20, it consists of 72 actual driving recorded trips in Munich, Germany, of a BMW I3 (60 Ah), highlighting the HVAC system's significant power demand. This dataset validates a comprehensive vehicle model, including the powertrain and the HVAC. Each trip contains environmental, vehicle, battery, and heating circuit data. It includes Category A (recorded trips in summer) and Category B (winter).

- Dataset 2: known as HELECAR-D, is a dataset collected in Morocco from three routes within the cities of Rabat and Sala Al Jadida. The data were gathered using a Renault TWIZY vehicle and includes various features such as date, time, battery SOC, vehicle speed, driving mode, GPS data, weather conditions, traffic information, and speed limits. It is worth noting that two recorded trajectories took place on urban roads, and the dataset includes 3 categories: T1 (light traffic), T2 (dense traffic), and T3 (moderate traffic). The works of [27] have used this dataset for SOC forecasting using a hybrid approach as stated in the related work section.

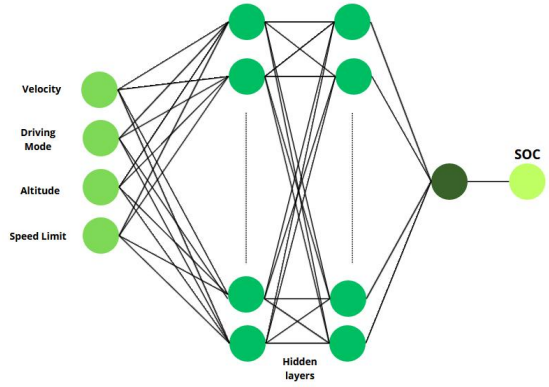
We reconstituted the databases for the second test evaluation as follows:

1. Munich20 Database:
   (a) DB_test1_Munich20: A mix of trips A (70%) and B (30%). It is divided into 2/3 for training and 1/3 for testing.
   (b) DB_test2_Munich20: A mix of trips A (30%) and B (70%), used entirely for testing (100%).
2. HELECAR-D Database:
   (a) DB_test1_HELECAR-D: This dataset includes a mix of trips with light traffic (T1, 50%) and dense traffic (T2, 50%). It is divided into 2/3 for training and 1/3 for testing.
   (b) DB_test2_HELECAR-D: This dataset comprises trips with moderate traffic (T3) and is used entirely for testing (100%).

Data preprocessing consists of handling missing values and outliers of each trip and then integrating values of chosen trips to have a large dataset for better model training; the input features are then selected as presented in Figure 3.



(**a**)



(**b**)

Figure 3. NN architecture with selected features for each driving scenario ((**a**) Munich20, (**b**) HELECAR-D 2).

Finally, all obtained data are normalized to mitigate fluctuations associated with the varying scales of each feature. Both datasets are split into training and testing as highlighted previously. All the chosen algorithms were tested twice on the corresponding test sets, and each algorithm's performance should be assessed using error metrics as indicated in the flow chart of the proposed evaluation methodology.

In this study, we developed two separate neural network (NN) models for SOC estimation to account for the differences in features between the datasets. The MUNICH20 dataset includes HVAC energy consumption as a feature but does not provide information on driving modes, whereas the HELECAR-D dataset lacks HVAC energy consumption but includes driving modes. These features are critical as they are strongly correlated with SOC changes. Training a single NN model across both datasets would introduce significant bias. For instance, including the driving mode feature in the MUNICH20 model would always result in the feature's value being zero, which could skew the model's learning process. Similarly, adding HVAC energy demand to the HELECAR-D model would misrepresent the data since the vehicle used in this dataset does not have an HVAC system. To ensure accurate learning from the available features in each dataset, two independent NN models were trained, each optimized for its respective dataset.

First, we trained regression models (LR, SVR, and RF) on both datasets.

Figures 4 and 5 present the estimated SOC output.

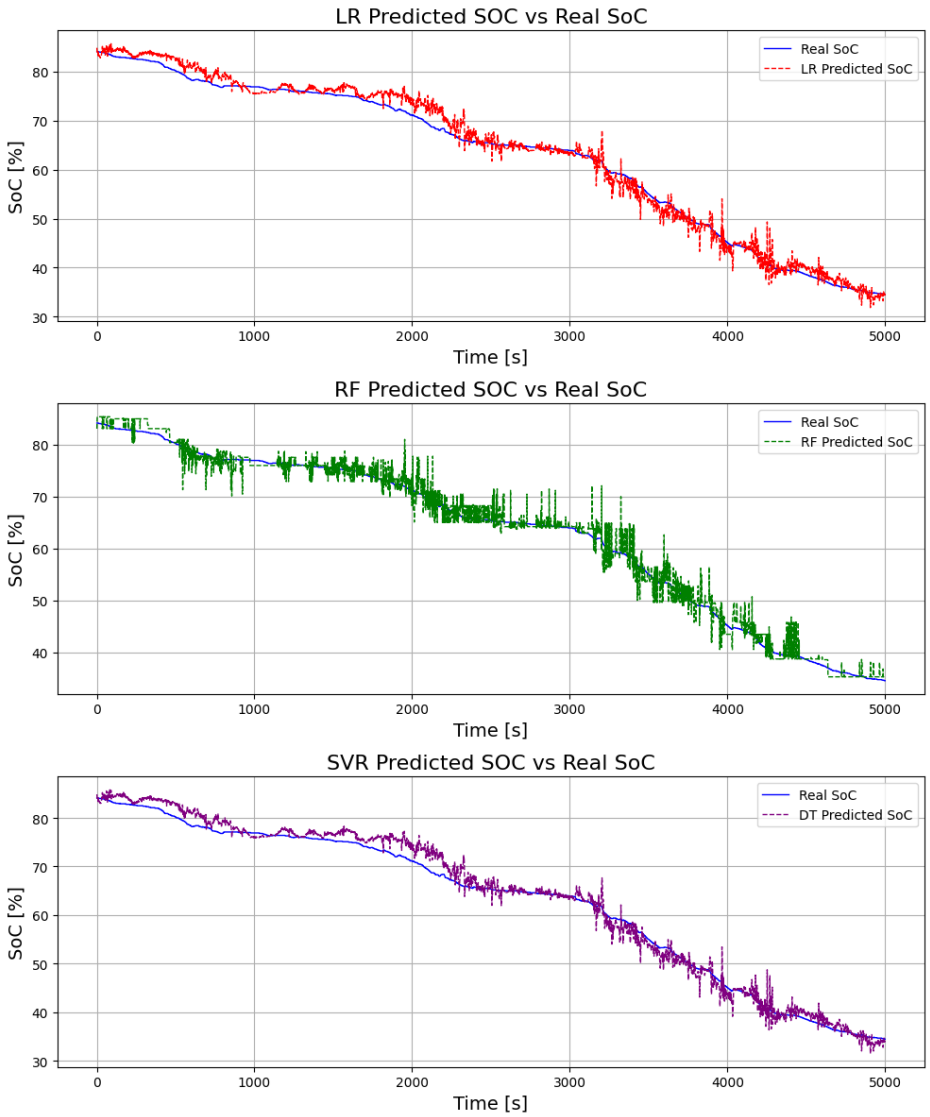


Figure 4. Recorded SOC vs. Estimated SOCusing Munich20.

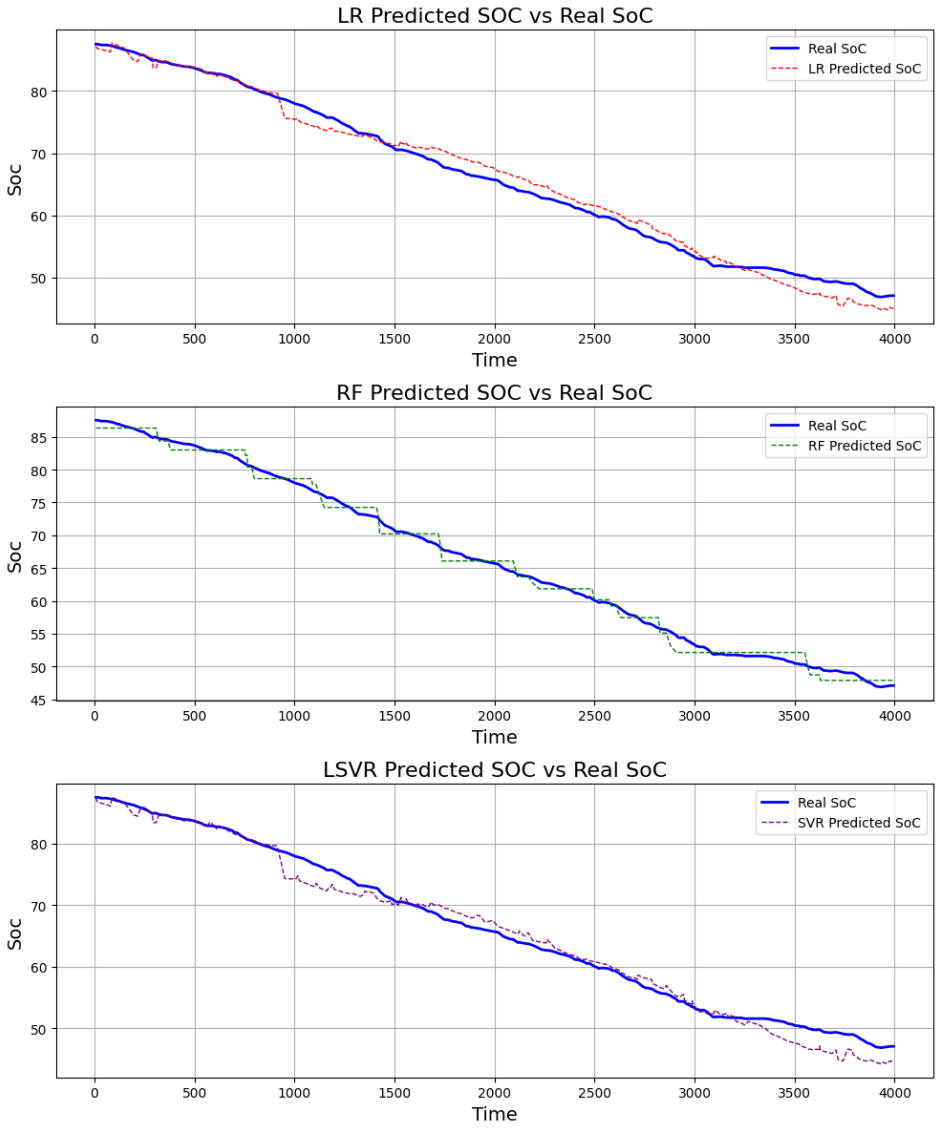


Figure 5. Recorded SOC vs. Estimated SOC using HELECAR-D.

(a) **Regression performance**

In the first scenario as well as the second, all the regression algorithms have performed commendably in estimating the SOC values over time on the first test sets. The RF algorithm showed a relatively lower performance compared to LR and SVR for Munich20, with an MSE of 3.44 and R^2 of 0.972, but was still fairly accurate. For HELECAR-D, RF excelled in the first test with the lowest MSE of 1.20 and the highest R^2 of 0.997, indicating very accurate estimations. However, to validate the regression models for SOC forecasting, we must consider the different input features selected in both datasets. It is prudent to test on a new set for both scenarios. Table 2 presents the performance indices (PI) of the regression models on both tests for each dataset. Notably, the second test represents actual trips considered only for a second-test model evaluation. Performance of all algorithms decreased in Test 2 compared to Test 1, with SVR maintaining the best balance of error metrics (MSE of 2.32 and MAE of 1.27) and R^2 (0.73), indicating relatively better predictive capability for Munich20. However, the performance on the second test set for both datasets decreases, evidenced by LR's R^2 dropping to 0.56 in Munich20, Test 2, even after extensive hyperparameter tuning and model regularization. Most regression models show promise when splitting the datasets into training, validation, and test sets due

to the model's tendency to overfit with large datasets and numerous features, even after thorough data preprocessing. Consequently, it is essential to conduct additional testing on different scenarios before validating the models.

(b) **Neural Network performance**

The NN training takes longer time than the previous algorithms; however, it has also performed well in the first test sets for both scenarios, as presented in Figures 6 and 7 and also in the second test sets, as in Table 2, compared to the ML performance on the second test set. These findings demonstrate that the modeled NN for both datasets can handle overfitting much better than the ML algorithms.

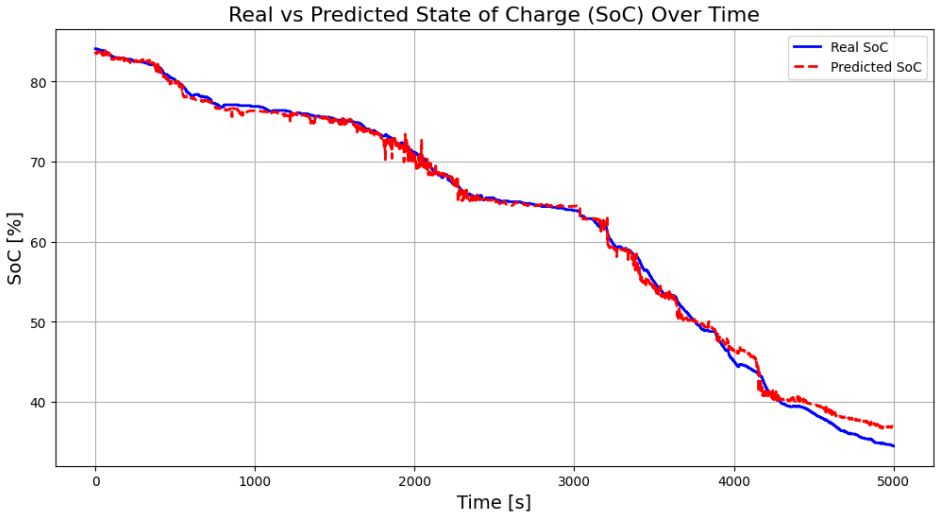


Figure 6. Recorded SOC vs. NN-Estimated SOC using Munich20.

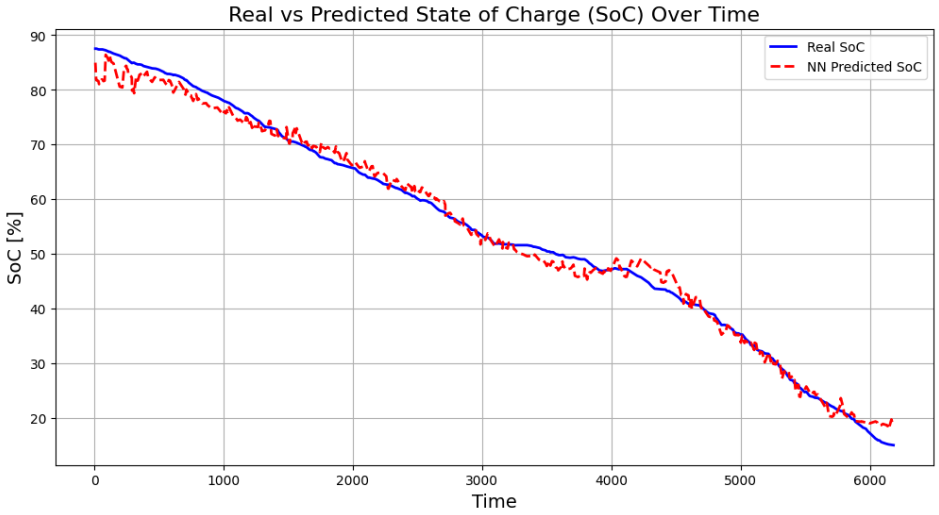


Figure 7. Recorded SOC vs. NN-Estimated SOC using HELECAR-D.

Across most tests, neural networks (NN) consistently outperformed other algorithms, especially in Munich20, where they exhibited remarkable adaptability to various data conditions, reflected in their low error rates (MSE of 0.62) and high R^2 value (0.992) in the first test. Conversely, random forest's (RF) performance varied: it was outstanding in the first test of HELECAR-D (with an MSE of 1.20 and R^2 of 0.997) but less impressive in other scenarios, particularly in the second test of Munich20, where it had an MSE of 7.29 and an R^2 of 0.75. While reliable, linear regression (LR) and support vector regression (SVR) did not reach the performance levels of NN and were occasionally surpassed by RF, as seen with LR's R^2 of 0.56 in Munich20, Test 2. This fluctuation in effectiveness across different tests and datasets underscores the impact of specific data characteristics on the

algorithms' performance. NN's strong performance is likely due to their superior capability in modeling complex and non-linear relationships, a feature less effectively captured by the other models. Although each algorithm has its advantages, NN emerged as the most flexible and effective across the conducted tests and datasets, as shown by their consistent R^2 values above 0.9 in all but one scenario. Specifically, in the second test of Munich20, NN achieved an MSE of 0.62 and an R^2 of 0.92, and in HELECAR-D, Test 2, they maintained a high R^2 of 0.96 despite the challenges. These findings highlight NN's robustness and adaptability in various testing conditions.

Table 2. Comparison of algorithm performance using Munich20 and HELECAR-D.

Dataset	Test	Algorithm	MSE	MAE	RMSE	R^2
Munich20	1	LR	2.82	1.29	1.68	0.97
Munich20	1	SVR	2.92	1.27	1.71	0.976
Munich20	1	RF	3.44	1.31	1.85	0.972
Munich20	1	NN	0.62	0.49	0.79	0.995
Munich20	2	LR	3.83	1.71	1.96	0.56
Munich20	2	SVR	2.32	1.27	1.52	0.73
Munich20	2	RF	7.78	2.21	2.79	0.11
Munich20	2	NN	0.62	0.65	0.79	0.92
HELECAR-D	1	LR	2.29	1.26	1.51	0.994
HELECAR-D	1	SVR	2.63	1.24	1.62	0.993
HELECAR-D	1	RF	1.20	0.93	1.10	0.997
HELECAR-D	2	NN	3.60	1.29	1.90	0.96
HELECAR-D	2	LR	37.84	5.64	6.15	0.65
HELECAR-D	2	SVR	17.25	3.55	4.15	0.84
HELECAR-D	2	RF	70.39	6.80	8.39	0.36
HELECAR-D	1	NN	3.63	1.51	1.91	0.991

In electric mobility research, accurately estimating the SOC is critical for operational reliability and efficiency. This work contributes to the field by using real-world driving data, contrasting with most existing literature, which uses battery data from controlled laboratory conditions as presented in the works of [9,24,32] or overly simplified driving scenarios. Our findings indicate that real-world data, including a more extensive range of driving scenarios and auxiliary components like HVAC systems, weather conditions, and specific route information, results in a more significant difference between Estimated SOC and Recorded SOC values from the datasets. This deviation emphasizes the complexities of real-world EV operation, as opposed to the idealized settings commonly depicted in simulation-based research. In our analysis, neural networks (NN) demonstrated superior performance in handling these complexities, especially when compared to more traditional models like linear regression (LR) and support vector regression (SVR). The NN's ability to minimize MSE and maximize R^2 is particularly notable in HELECAR-D. The findings align with results from [27] that employed a hybrid model using LR and LSTM networks, which also leveraged real-world data. However, our research adds more to the conversation by integrating an additional test set and additional datasets that highlight the energy consumption of HVAC systems, a critical factor that should not be overlooked in SOC estimation for BEVs. Moreover, our research incorporated an essential second testing phase, a step often neglected in related studies. This second phase is crucial for reducing the risk of overfitting, which happens when models are too fine-tuned to a particular dataset, making them less effective in real-world situations. This part of our study helps ensure

that the utilized NN models are robust and can be trusted to work well across different driving conditions.

5. Conclusions

Battery state of charge (SOC) estimation in electric vehicles is crucial for intelligent energy management; therefore, in this paper, we evaluated four different algorithms: linear regression (LR), support vector regression (SVR), random forest regression (RF), and neural networks (NN) on two datasets from different contexts (European and African). With each specific number of features for SOC estimation, two tests were applied on both algorithms for each dataset. It can be concluded based on the findings that NNs consistently displayed high predictive accuracy across different scenarios in Munich20, surpassing models with the lowest MSE and highest R^2 values. While LR and SVR demonstrated commendable model fits with high R^2 values, NN generally outperformed them. RF exhibited a more variable performance, with relatively lower accuracy than LR and SVR in Munich20, but excelled in HELECAR-D, Test 1, with the best predictive results among the algorithms. However, RF's performance notably declined in other tests. SVR emerged as a robust algorithm in HELECAR-D, Test 2, showing a favorable balance between error metrics and R^2, suggesting a reliable performance even when other algorithms faltered. Overall, while each algorithm had strengths in specific tests, NN proved to be the most consistent regarding predictive accuracy. RF showed the ability to achieve high accuracy under certain conditions.

In upcoming projects, we plan to use the neural network (NN) models to estimate the state of charge (SOC) for a vehicle's journey. This estimation will then make intelligent decisions about distributing the vehicle's energy. By knowing the SOC in advance, we can better plan how to use the energy throughout the trip. This could help in using less power when it is not needed, which saves energy, or ensuring there is enough power for the parts of the trip that require more, like uphill stretches. The goal is to manage the vehicle's energy most effectively, making the ride smoother and the vehicle's energy use more efficient.

Author Contributions: S.A.: Conceptualization, Methodology, Data curation and pre-processing, Algorithm testing and validation, Writing—original draft. Z.B., A.E.O. and A.E.: Data curation, Formal analysis, Visualization, Writing—review & editing. A.T. and S.E.: Supervision, Project administration, Reviewing and Editing, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from: The Ministry of Europe and Foreign Affairs (MEAE), the Ministry of Higher Education and Research (MESR), and the Ministry of Higher Education, Scientific Research, and Innovation (MESRSI). The funding falls under the Franco-Moroccan Hubert Curien Partnership (PHC-TOUBKAL) program 2022–2023, with Grant number: 48530UC.

Data Availability Statement: We have provided the link to the datasets used in this paper [47,48].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. López-Claros, A.; Dahl, A.L.; Groff, M. *Global Governance and the Emergence of Global Institutions for the 21st Century*; Cambridge University Press: Cambridge, UK, 2020. [CrossRef]
2. Hossain; Kumar, L.; Assad, M.E.H.; Alayi, R. Advancements and Future Prospects of Electric Vehicle Technologies: A Comprehensive review. *Complexity* **2022**, *2022*, 3304796. [CrossRef]
3. International Energy Agency (IEA). Global EV Outlook 2021. 2021. Available online: https://www.iea.org/reports/global-ev-outlook-2021 (accessed on 15 January 2024).
4. She, Z.Y.; Sun, Q.; Ma, J.J.; Xie, B.C. What are the barriers to widespread adoption of battery electric vehicles? A survey of public perception in Tianjin, China. *Transp. Policy* **2017**, *56*, 29–40. [CrossRef]
5. Lu, L.; Han, X.; Li, J.; Hua, J.; Ouyang, M. A review on the key issues for lithium-ion battery management in electric vehicles. *J. Power Sources* **2013**, *226*, 272–288. [CrossRef]
6. Zhang, M.; Fan, X. Review on the state of charge estimation methods for electric vehicle battery. *World Electr. Veh. J.* **2020**, *11*, 23. [CrossRef]

7. Bhatt, D.K.; Darieby, M.E. An Assessment of Batteries form Battery Electric Vehicle Perspectives. In Proceedings of the 2018 IEEE International Conference on Smart Energy Grid Engineering (SEGE), Oshawa, ON, Canada, 12–15 August 2018; pp. 255–259. [CrossRef]
8. Norouzi, A.; Heidarifar, H.; Borhan, H.; Shahbakhti, M.; Koch, C.R. Integrating Machine Learning and Model Predictive Control for automotive applications: A review and future directions. *Eng. Appl. Artif. Intell.* **2023**, *120*, 105878. [CrossRef]
9. Zhao, J.; Ling, H.; Liu, J.; Wang, J.; Burke, A.; Lian, Y. Machine learning for predicting battery capacity for electric vehicles. *eTransportation* **2023**, *15*, 100214. [CrossRef]
10. Zekić–Sušac, M.; Mitrović, S.; Has, A. Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. *Int. J. Inf. Manag.* **2021**, *58*, 102074. [CrossRef]
11. Musbah, H.; Aly, H.H.H.; Little, T.A. Energy management of hybrid energy system sources based on machine learning classification algorithms. *Electr. Power Syst. Res.* **2021**, *199*, 107436. [CrossRef]
12. Singirikonda, S.; Obulesu, Y.P. Battery modelling and state of charge estimation methods for Energy Management in Electric Vehicle—A review. *IOP Conf. Ser.* **2020**, *937*, 012046. [CrossRef]
13. Naguib, M.; Kollmeyer, P.J.; Emadi, A. Lithium-Ion battery pack robust state of charge estimation, cell inconsistency, and balancing: Review. *IEEE Access* **2021**, *9*, 50570–50582. [CrossRef]
14. Zhou, W.; Zheng, Y.; Pan, Z.; Lu, Q. Review on the battery model and SOC estimation method. *Processes* **2021**, *9*, 1685. [CrossRef]
15. Hannan, M.A.; Lipu, M.S.H.; Hussain, A.; Mohamed, A. A review of lithium-ion battery state of charge estimation and management system in electric vehicle applications: Challenges and recommendations. *Renew. Sustain. Energy Rev.* **2017**, *78*, 834–854. [CrossRef]
16. Wang, Y.; Tian, J.; Sun, Z.; Wang, L.; Xu, R.; Li, M.; Chen, Z. A comprehensive review of battery modeling and state estimation approaches for advanced battery management systems. *Renew. Sustain. Energy Rev.* **2020**, *131*, 110015. [CrossRef]
17. Wu, Y.; Sundaresan, S.; Balasingam, B. Battery Parameter Analysis through Electrochemical Impedance Spectroscopy at Different State of Charge Levels. *J. Low Power Electron. Appl.* **2023**, *13*, 29. [CrossRef]
18. Cheng, K.W.E.; Divakar, B.; Wu, H.; Ding, K.; Ho, H.F. Battery-Management System (BMS) and SOC development for electrical vehicles. *IEEE Trans. Veh. Technol.* **2011**, *60*, 76–88. [CrossRef]
19. Plett, G.L. Sigma-point Kalman filtering for battery management systems of LiPB-based HEV battery packs. *J. Power Sources* **2006**, *161*, 1356–1368. [CrossRef]
20. Zhao, Y.; Xie, W.; Wu, J. A novel SOC estimation method for supercapacitor cell module based on EKF-MF hybrid filtering algorithm. *J. Electr. Eng. Technol. Electr. Eng. Technol.* **2024**, *19*, 4927–4940. [CrossRef]
21. Sepasi, S.; Ghorbani, R.; Liaw, B.Y. SOC estimation for aged lithium-ion batteries using model adaptive extended Kalman filter. In Proceedings of the 2013 IEEE Transportation Electrification Conference and Expo (ITEC), Detroit, MI, USA, 16–19 June 2013; pp. 1–6. [CrossRef]
22. How, D.N.; Hannan, M.; Lipu, M.H.; Ker, P.J. State of charge estimation for lithium-ion batteries using model-based and data-driven methods: A review. *IEEE Access* **2019**, *7*, 136116–136136. [CrossRef]
23. Castanho, D.S.; Guerreiro, M.; Silva, L.C.A.; Eckert, J.J.; Alves, T.A.; De Souza Tadano, Y.; Stevan, S.L.; Siqueira, H.V.; Corrêa, F.C. Method for SOC estimation in Lithium-Ion batteries based on multiple linear regression and particle swarm optimization. *Energies* **2022**, *15*, 6881. [CrossRef]
24. Jumah, S.; Elezab, A.; Zayed, O.; Ahmed, R.; Narimani, M.; Emadi, A. State of Charge Estimation for EV Batteries Using Support Vector Regression. In Proceedings of the 2022 IEEE Transportation Electrification Conference & Expo (ITEC), Anaheim, CA, USA, 15–17 June 2022; pp. 964–969. [CrossRef]
25. Venkatesan, C.; Patil, C.K.; Karthick, A.; Dharmaraj, G.; Rahim, R.; Ghosh, A. State of charge estimation of Lithium-Ion battery for electric vehicles using machine learning algorithms. *World Electr. Veh. J.* **2021**, *12*, 38. [CrossRef]
26. Li, C.; Chen, Z.; Cui, J.; Wang, Y.; Zou, F. The lithium-ion battery state-of-charge estimation using random forest regression. In Proceedings of the 2014 Prognostics and System Health Management Conference (PHM-2014 Hunan), Zhangjiajie, China, 24–27 August 2014; pp. 336–339. [CrossRef]
27. NaitMalek, Y.; Najib, M.; Lahlou, A.; Bakhouya, M.; Gaber, J.; Essaaidi, M. A hybrid approach for State-of-Charge forecasting in Battery-Powered electric vehicles. *Sustainability* **2022**, *14*, 9993. [CrossRef]
28. Wang, Q.; Sun, C.; Gu, Y. Research on SOC estimation method of hybrid electric vehicles battery based on the grey wolf optimized particle filter. *Comput. Electr. Eng.* **2023**, *110*, 108907. [CrossRef]
29. Das, K.; Kumar, R. A comprehensive review of categorization and perspectives on State-of-Charge estimation using deep learning methods for electric transportation. *Wirel. Pers. Commun.* **2024**, *133*, 1599–1618. [CrossRef]
30. Lee, D.T.; Shiah, S.J.; Lee, C.M.; Wang, Y.C. State-of-charge estimation for electric scooters by using learning mechanisms. *IEEE Trans. Veh. Technol.* **2007**, *56*, 544–556. [CrossRef]
31. Guo, S.; Ma, L. A comparative study of different deep learning algorithms for lithium-ion batteries on state-of-charge estimation. *Energy* **2023**, *263*, 125872. [CrossRef]
32. El Fallah, S.; Kharbach, J.; Hammouch, Z.; Rezzouk, A.; Jamil, M.O. State of charge estimation of an electric vehicle's battery using Deep Neural Networks: Simulation and experimental results. *J. Energy Storage* **2023**, *62*, 106904. [CrossRef]
33. Chemali, E.; Kollmeyer, P.J.; Preindl, M.; Emadi, A. State-of-charge estimation of Li-ion batteries using deep neural networks: A machine learning approach. *J. Power Sources* **2018**, *400*, 242–255. [CrossRef]

34. Azkue, M.; Miguel, E.; Martinez-Laserna, E.; Oca, L.; Iraola, U. Creating a Robust SoC Estimation Algorithm Based on LSTM Units and Trained with Synthetic Data. *World Electr. Veh. J.* **2023**, *14*, 197. [CrossRef]
35. Bockrath, S.; Rosskopf, A.; Koffel, S.; Waldhör, S.; Srivastava, K.; Lorentz, V.R. State of charge estimation using recurrent neural networks with long short-term memory for lithium-ion batteries. In Proceedings of the IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal, 14–17 October 2019; Volume 1, pp. 2507–2511. [CrossRef]
36. Li, X.; Jiang, H.; Guo, S.; Xu, J.; Li, M.; Liu, X.; Zhang, X. SOC estimation of Lithium-Ion battery for electric vehicle based on deep multilayer perceptron. *Comput. Intell. Neurosci.* **2022**, *2022*, 3920317. [CrossRef]
37. Kim, S.; Lim, H. Reinforcement learning based energy management algorithm for smart energy buildings. *Energies* **2018**, *11*, 2010. [CrossRef]
38. NaitMalek, Y.; Najib, M.; Bakhouya, M.; Essaaidi, M. Embedded real-time battery State-of-Charge forecasting in Micro-Grid systems. *Ecol. Complex.* **2021**, *45*, 100903. [CrossRef]
39. Ariche, S.; Boulghasoul, Z.; El Ouardi, A.; Elbacha, A.; Tajer, A.; Espie, S. Energy Consumption of a Battery Electric Vehicle for a Comfort Ride on Moroccan Roads. In Proceedings of the 2023 9th International Conference on Control, Decision and Information Technologies (CoDIT), Rome, Italy, 3–6 July 2023; pp. 1–5. [CrossRef]
40. Electric Vehicle Thermal Management. Mathworks. Available online: https://www.mathworks.com/help/hydro/ug/sscfluids_ev_thermal_management.html (accessed on 25 November 2024).
41. Vehicle Electrical and Climate Control Systems. Mathworks. Available online: https://www.mathworks.com/help/simulink/slref/vehicle-electrical-and-climate-control-systems.html (accessed on 25 November 2024).
42. EV Battery Cooling System. Mathworks. Available online: https://www.mathworks.com/help/hydro/ug/ev-battery-cooling.html (accessed on 25 November 2024).
43. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
44. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**, arXiv:1603.04467.
45. Chollet, F.; et al. Keras. 2015. Available online: https://keras.io (accessed on 15 April 2024).
46. Naser, M.; Alavi, A.H. Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences. *Archit. Struct. Constr.* **2021**, *3*, 499–517. [CrossRef]
47. Steinstraeter, M. Battery and Heating Data in Real Driving Cycles. IEEE Dataport. 2020. Available online: https://ieee-dataport.org/open-access/battery-and-heating-data-real-driving-cycles (accessed on 24 November 2024).
48. NaitMalek, Y.; Najib, M.; Bakhouya, M.; Gaber, J. HELECAR-D: A dataset for urban electro mobility in Moroccan context. *Data Brief* **2023**, *48*, 109080. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.