State-Of-Health Estimation for Lithium-Ion Batteries Using Supervised Machine Learning: XGBoost

Muhamad Maulanal Haq[1,2], Irsyad Nashirul Haq[1] & Justin Pradipta[1]

[1] Department of Engineering Physics, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40132, Indonesia
[2] PT PLN (Persero) Head Office, Jl. Trunojoyo Blok M-I No.135, Jakarta 12160, Indonesia

Email: muhamad.maulanal@pln.co.id

Abstract. Batteries are energy storage systems used in almost every aspect. As the battery will degrade over time as it is used, the Battery Management System (BMS) needs to be able to monitor its health so that the right battery replacement time can be predicted. Several State-of-Health (SoH) estimation methods have been studied, one of which is the data-driven method. This paper proposed SoH estimation for universal lithium-ion batteries using supervised machine learning: XGBoost that trained with only one battery material, which is Lihtium-Nickel-Cobalt-Alumunium (NCA). From Model evaluation results show that the model is able to predict well on other data with the same material as the training data with RMSE 1.4320% MAE 0.9174% and MAPE 0.0100%. However, to make predictions on other types of material data, the model has difficulty because XGBoost is not able to make good predictions outside of the training data.

Keywords: *battery; battery management system; lithium-ion; state-of-health; supervised machine learning, xgboost.*

1 Introduction

Batteries are energy storage media commonly used in daily life applications such as electronic devices, backup energy systems on the power grid, and electric vehicles. In addition to the support of electrical energy providers, the use of batteries in daily applications also requires a Battery Management System (BMS) as a management system to monitor and maximize the battery's performance. BMS will check battery parameters such as voltage, current, and internal temperature during the charging and draining process and estimate battery conditions such as actual capacity and battery health [1]. State of Health (SoH) indicates the level of battery degradation or available capacity compared to the battery capacity in new condition. The types of battery degradation namely Loss of Lithium Inventory (LLI) and Loss of Active Material (LAM) cause a decrease in capacity [2]

Received _____, Revised _____, Accepted for publication _____

Estimation of SoH parameter values generally consists of three methods, direct calculation methods, modeling-based methods, and data-driven methods [3]. The direct calculation method will measure the value of the open circuit voltage which will then be compared with the voltage and capacity curve. While the model-based method compares the voltage, current, and temperature values with the modeling results that have been made. The data-driven method is analogous to a 'black box' that performs SoH estimation [3].

The data-driven method uses Machine Learning to estimate SoH, so it requires a large database on the charge and discharge cycle of a battery. The large database is also because in actual conditions there are differences in operating conditions from each battery use, so the resulting degradation patterns vary [4]. With existing historical data, it is expected to be able to provide an accurate estimate of the degradation pattern of the State of Health of the battery.

Based on the results of the literature study, it is found that there have been previous studies that use various machine learning models such as extreme learning, self-supervised, and deep neural networks. Each of them uses different features as parameters in the estimator. In SoH estimation research by Pan, Rui in 2023 using up to 15 Health Features (HFs), with the resulting RMSE value in the range of 1,168 to 2,290. Then research from Chen, Si-Zhe in 2023 [5] only used 4 Health Features with RMSE results in the range of 0.56 to 0.96. This shows that it is necessary to choose features that have a good correlation with KK and able to produce accurate values.

In this study, the estimation of battery SoH was carried out using the Supervised Machine Learning method, where the Supervised method provides the speed of the learning process on a large database and already has a label on the work parameters that have been measured so that it can minimize the resources used.

2 Material and Data Understanding

This study used data derived from widely published battery research results. There are three types of battery materials from three different data sources.

2.1 Lithium Nickel-Cobalt-Aluminum Oxide (NCA)

This type of NCA battery was obtained from Jöst, Dominik et al in 2021 [6], manufactured by Samsung (INR18650-35E) with a nominal voltage of 3.6V and a nominal capacity of 3.4Ah. In this database contains 28 pcs high energy NCA/C+Si cycle aging test, and every battery performed the same profile test which is a series of cycling tests and checkup tests. In the cycling test the battery was charged with CCCV at 1.02 A current until 4.05V and 0.068 A cutoff.

For this study, the battery data used for the data training is a battery with ID 020, and ID 027 for the evaluation process. The discharge current varies in every cycle, as seen in Figure 1, maximum discharge current at 10A, and charging current around 2A.

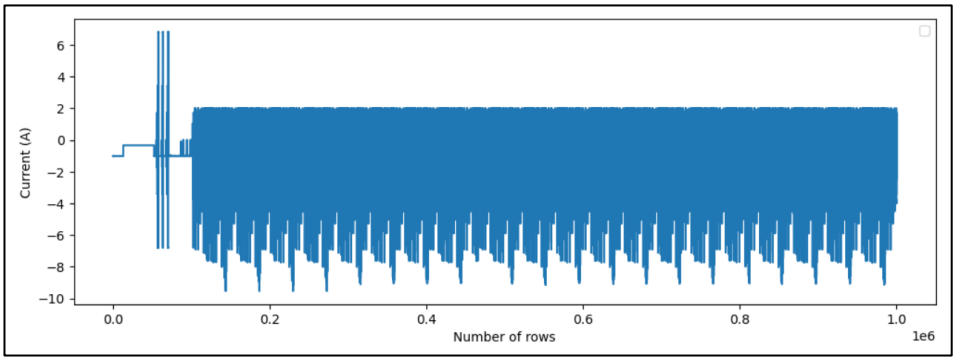

Figure 1 Battery current discharge from partial data of NCA ID 020

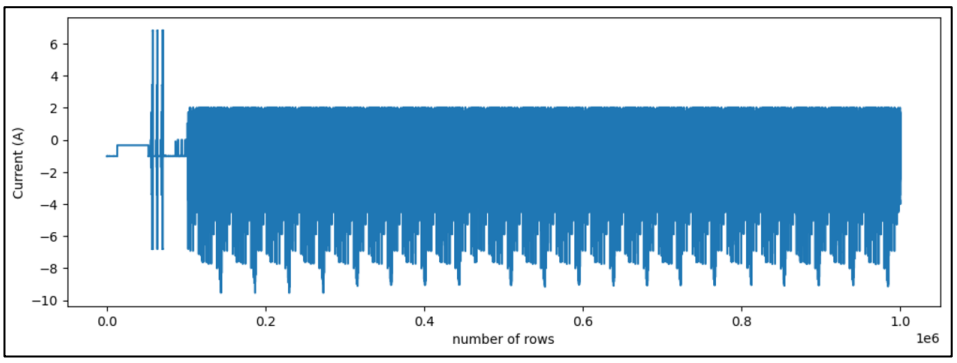

Figure 2 Battery current discharge from partial data of NCA ID 027

2.2 Lithium Cobalt Oxide (LCO)

This LCO battery data from a fast charging test research by Gun, Define et al in 2015 [7] from Berkeley University, with four cell ID were performed several CCCV cycles that gradually increased in C-rates to examine its behavior. Then, MCC, CP-CV, and Boostcharge cycles at various C-rates with an additional 1C CCCV baseline capacity test. The batteries were manufactured by Sanyo (18650) with a nominal capacity of 2.6 Ah and a nominal voltage of 3.7 V.

In this paper, the LCO ID 4 is use as evaluation data. The battery discharge profile used a constant current 2.5A in every cycle as seen in Figure 2.

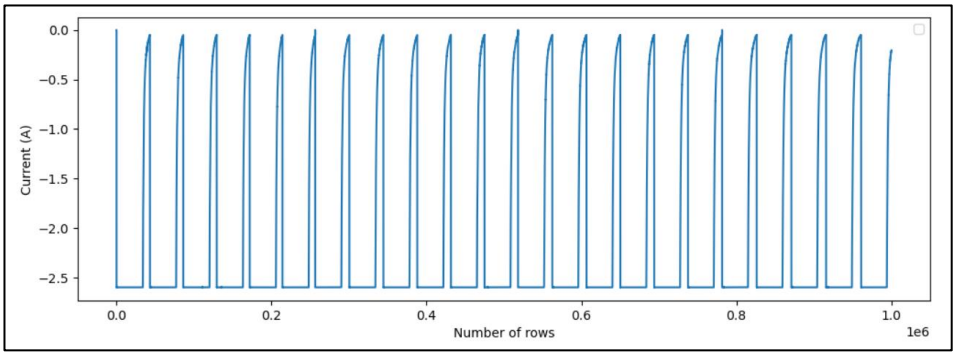

Figure 3 Battery current discharge data of LCO ID 4

2.3 Lithium Iron Phosphate (LFP).

This LFP battery that used in this paper was provided by Toyota Research Institute from Severson et al in 2019 [8], and consists of 124 commercial lithium-ion batteries manufactured by A123 Systems (APR18650M1A) were cycled to failure under different fast-charging conditions. The cell have a nominal capacity of 1.1 Ah and a nominal voltage of 3.3 V.

The LFP ID 47 use as evaluation data has discharged current profile as shown in Figure 3. The discharge current gradually increased with peak at 4A, and slowly decreased before it drained.

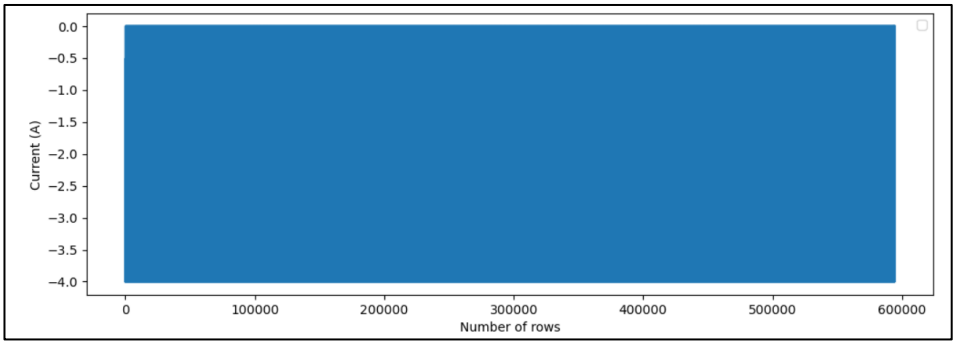

Figure 4 Battery current discharge data of LFP 47

3 Methodology

The study runs in few step including data preparation, model training and evaluation process.

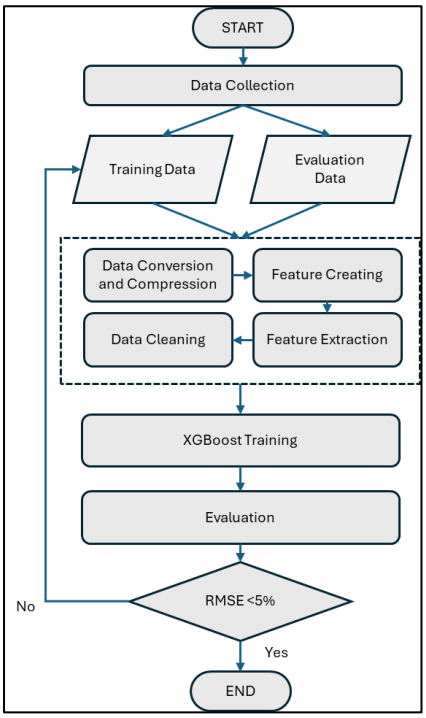

Figure 5 Reasearch Methodology

3.1 Data Preparation

Data preparation in this study consists of five steps. The first step is data conversion and compression. The collected raw data needs to be converted from its original file to be processed in the pandas Python library. The parquet extension was chosen because it can store data with a smaller memory, making it easier for the computing device to process data at low cost.

The next step is feature creation. Raw data already has several features from data acquisition, including Voltage, Current, Time, and Temperature, which are then used to create new features.

In this study, feature extraction based on Incremental Capacity Analysis (ICA) techniques was carried out to investigate the battery degradation mechanism [11]. This ICA technique is used by comparing changes in charge/discharge capacity with changes in voltage [12].

$$ICA = \frac{dQ}{dV} = \frac{\Delta Q}{\Delta V} = \frac{(Q_i - Q_{i-1})}{(V_i - V_{i-1})} \quad (1)$$

The ICA can be obtained from the constant current (CC) charging curve of the battery [13]. The results of the calculation of the ICA will form a curve in relation to the voltage during constant current charging. From the ICA curve can be obtained peak shape values and variations in the position of the peak which is closely related to the decrease in battery capacity [12]. In a study conducted by Wang, Guangfeng [14], it was shown that there is a high correlation between health features such as the position and size of the IC/DV peak and SOH [14]. From the Incremental Capacity Analysis curve generation process as shown in Figure 6, several features can be obtained, the first is 'max_ica' as the ICA value of the peak of the curve, the secondly 'voltage_atmaxica' as the voltage value when the peak value is formed, and the last is 'peak_area' as the peak area under the curve with a voltage range of ±0.025V from 'voltage_atmaxica'.

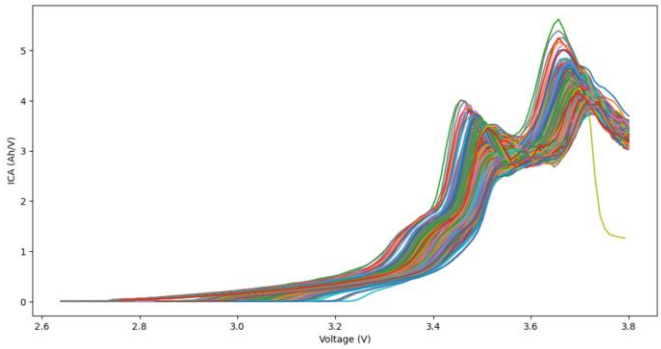

Figure 6 ICA Curve of NCA-020 Charging Data

The SoH feature is calculated using the actual value of the battery capacity at a given cycle (C_{act}) compared to the nominal capacity when the battery is in new condition (C_{new}).

$$SoH = \frac{C_{act}}{C_{new}} \tag{2}$$

The third step is Feature Extraction, where existing feature data is selected to be used in the training data. In this research there are several features used based on ICA including ‘voltage_atmaxica’, ‘voltage_atmaxica’, and ‘peak_area’. And two features that has low correlation to SoH are dQ_mean as the mean of capacity differences and temp_mean as the mean of temperature during the charging or discharging.

The fourth step is Data Cleaning, which is intended to remove noise in the data that can be caused by measurement errors or conditions that are not ideal during data acquisition.

Using correlation function from sklearn, the correlation of health features can be obtained as shown in the table below.

Table 1 Features Correlation with SOH

Feature	Correlation
max_ica	0.98
voltage_atmaxica	-0.88
peak_area	0.98
dQ_mean	0.69
temp_mean	-0.64

From the 5 health features that have been determined, five feature combinations are made to be used as experimental variations in this study. Here are the feature combinations shown in the table below.

Table 2 Feature Combinations

Feature Name	Combination	Health Features
	KF1	max_ica; voltage_atmaxica; peak_area
	KF2	max_ica; voltage_atmaxica; peak_area; dQ_mean
	KF3	max_ica; voltage_atmaxica; peak_area; dQ_mean; temp_mean

3.2 XGBoost Model Training

To build State-of-Health estimator, this paper uses Supervised Machine Learning with XGBoost Regressor as its algorithm. XGBoost is claimed to run faster up to ten times than other popular algorithms [9]. This is considered because the amount of battery data obtained as training data is quite large, so it is expected to require low computational costs. In another study by Chen, Si-Zhe. 2023 [5], XGBoost is the algorithm that has the smallest error as an estimator compared to RF, SVR, and KRR.

XGBoost uses hyperparameter to control the training process and the complexity of the resulting model. In this study use hyperparameter: objective, alpha, lambda, learning_rate, max_depth, and n_estimators.

3.3 Evaluation Process

In this Evaluation Step, the trained model estimate state-of-health with the Evaluation Data. The evaluastion metrics used in this research is RMSE, MAE and MAPE.

RMSE: Root Mean Squared Error

$$RMSE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(y_i - \hat{y}_i)^2} \quad (3)$$

MAE: Mean Absolute Error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(y_i - \hat{y}_i)| \quad (4)$$

MAPE: Mean Absolute Percentage Error

$$MAPE = 100 \text{ x } \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(y_i - \hat{y}_i)}{y_i}\right| \quad (5)$$

There are three different data for the evaluation from each material, the NCA battery with ID 027, LFP ID 47, and LCO ID 4 which represent three different battery materials.

4 Results and Discussion

Each model of SoH estimator trained with discharging data is through the evaluation process with battery data that has different material types.

In the initial stage, after the data is processed from data preparation stage, a hyperparameter search for XGBoost is carried out using the GridSearch function, where the computer automatically iterate using the XGBoost algorithm to determine the best hyperparameter that gives the minimum error. The computing device used in this research has a specification of 8 CPU Cores, and 16 Threads, with a base clock up to 4Ghz with 16GB RAM. The XGBoost computation used is CPU-based.

These hyperparameters used in all models made, the results of the selected hyperparameters include: 'alpha': 0.01, 'lambda': 0.1, 'learning_rate': 0.3, 'max_depth': 5, 'n_estimators': 200, 'objective': 'reg:squarederror'. After the hyperparameters have been determined, the estimator trained using training data on three different models with different feature combinations. The training data used is NCA ID 020 data, with 70% training data and 30% used as test data.

From the experimental results, the error value is obtained as in the Table 3.

Table 3 Performance Using Three Different Feature Combinations

Feature Combinations	KF1	KF2	KF3
RMSE	0.6782	0.6007	0.5568
MAE	0.3748	0.2999	0.2659
MAPE	0.0041	0.0032	0.0029

Based on the estimator prediction results, the acceptable RMSE value is less than 5% RMSE [10].

From the first evaluation results of three different feature combinations, the use of the KF3 combination gives the lowest RMSE value among the other combinations. Furthermore, an evaluation process was carried out on the three types of battery materials from the collected batteries database, NCA ID 027, LFP ID 47, and LCO ID 4. The results of model evaluation using validation data are shown in the table below.

Table 4 Model Performance in Validation Test

Validation Data	NCA-027 Charging	LFP-47 Charging	LFP-47 Discharging	LCO-4 Charging	LCO-4 Disharging
RMSE	1.4320	7.8864	6.2003	3.5607	4.7642
MAE	0.9174	6.9229	5.8380	2.5214	4.3211
MAPE	0.0100	0.0761	0.0632	0.0266	0.0455

After the validation process performed with other battery data, it was found that the model still has acceptable prediction results on the NCA-027 Charging data. This NCA-027 validation data has the same charging and discharging method as NCA-020, so the pattern of the ICA curve also has the same pattern. Therefore, the validation results show that the model predicts well on the same material.
In the validation process on the LCO-4 data, the RMSE value was obtained below the 5% target. Although it has a good RMSE, the SOH plot does not show a good prediction, there are several battery cycles that have a large enough error between the actual SOH value and the prediction as shown in Figure 10, especially in the LCO-4 Discharging data. In these data, the prediction results tend to be the same until the end of the cycle, so it can be said that the model is not able to see a decrease in SOH over the cycle.
In validation with LFP-47 data which has RMSE results above 5%, 7.88% on charging data, and 6.20% on discharging data, the estimator model is unable to predict the charging data from LFP-47. This can be caused by the charging method performed on the LFP-47 is a constant current consisting of 3 large currents, which causes the ICA curve formed to be a mixture of several constant currents so that the Incremental Capacity Analysis method is unable to capture the curve pattern over the cycle. In the discharge data from LFP-47, the SOH prediction shows a decrease at a cycle value of around 1700 and is able to approach the actual SOH value. The discharging data of LFP-47 has the same constant current method during the experiment cycle.

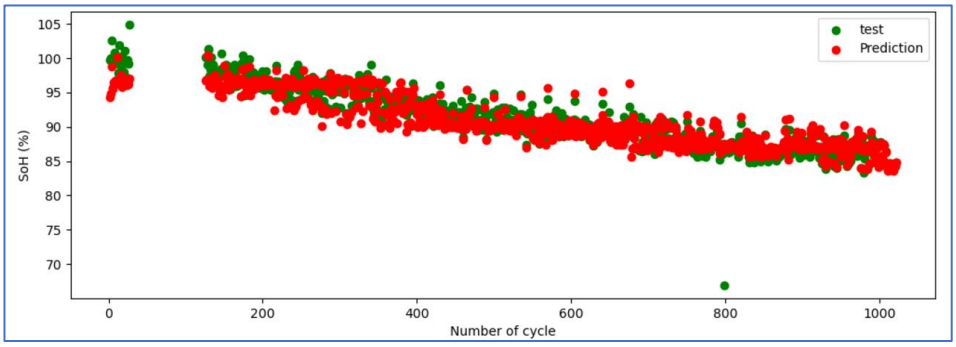

Figure 7 SoH prediction compare to SoH actual of NCA ID 027 Charging

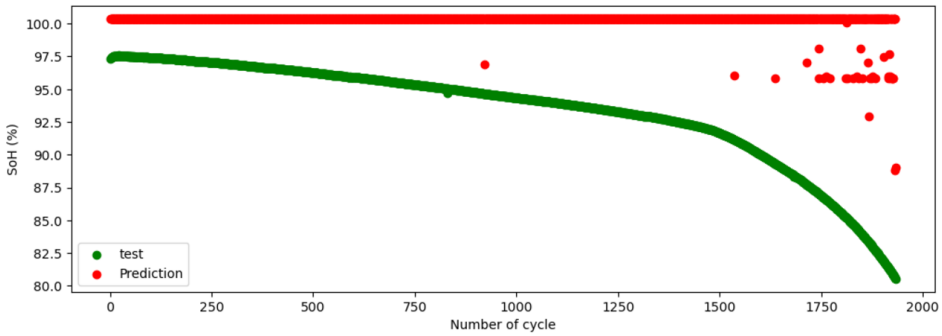

Figure 8 SoH prediction compare to SoH actual of LFP ID 47 Charging

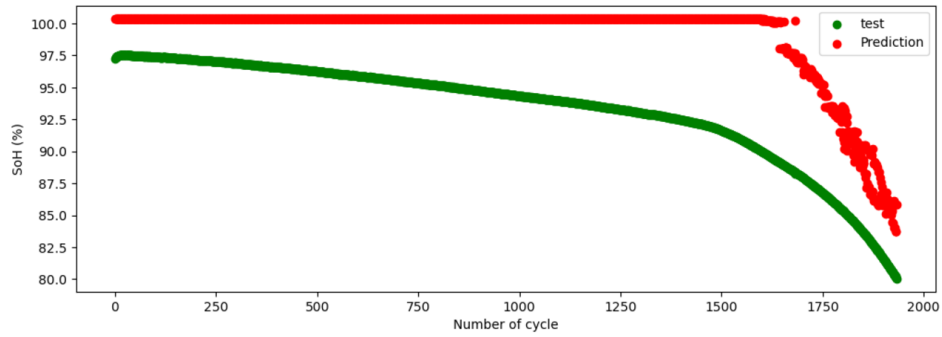

Figure 9 SoH prediction compare to SoH actual of LFP ID 47 Discharging

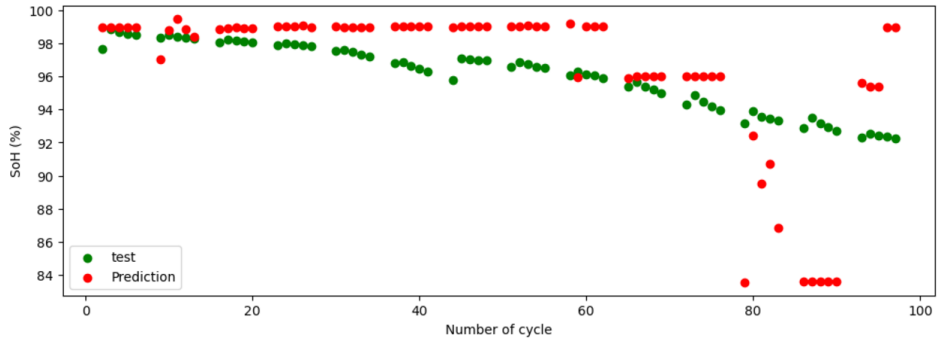

Figure 10 SoH prediction compare to SoH actual of LCO ID 4 Charging

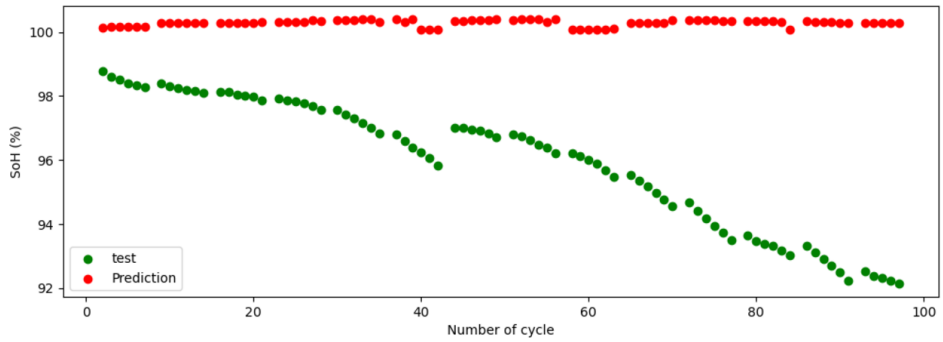

Figure 11 SoH prediction compare to SoH actual of LCO ID 4 Discharging

5 Conclussions

From the experiments conducted, the use of features obtained from Incremental Capacity Analysis can provide a small error in the SOH estimation application in the same material with RMSE 1.4320% MAE 0.9174% and MAPE 0.0100%, in this case, the NCA-027 battery material. However, in SOH prediction applications in other types of material data, the model has difficulty estimating SOH with error values that tend to be high, and prediction plot results that are not able to approach the actual SOH value. This can be caused by the charging and discharging methods applied or the machine learning method used. In further research, researchers have conducted several experiments using different learning methods. It was found that the XGBoost model was unable to make predictions on data outside its training data range, which is NCA battery data. So a more optimal supervised machine learning model is obtained using the Stacking method which combines the capabilities of XGBoost with SVR.

6 References

[1] Haq, Irsyad N. *Development of Battery Management System for Cell Monitoring and Protection*. 2014 IEEE International Conference on Electrical Engineering and Computer Science. 2014.

[2] Birkl, Christoph R. *Degradation diagnostics for lithium ion cells*. Journal of Power Sources. 2017

[3] Chen, L., Lü, Z., Lin, W., Li, J., Pan, H. *A new state-of-health estimation method for lithium-ion batteries through the intrinsic relationship between ohmic internal resistance and capacity*. Measurement 116, 586–595. http: //dx.doi.org/10.1016/J.MEASUREMENT.2017.11.016. 2018

[4] Deng, Zhongwei. Dkk. Battery health estimation with degradation pattern recognition and transfer learning, *Journal of Power Sources*, 2023

[5] Chen, Si-Zhe. Et al. *Li-ion battery state-of-health estimation based on the combination of statistical and geometric features of the constant-voltage charging stage.* Journal of Energy Storage. 2023

[6] Jöst, Dominik et al. *Timeseries data of a drive cycle aging test of 28 high energy NCA/C+Si round cells of type 18650*. DOI: 10.18154/RWTH-2021-02814. 2021.

[7] Gun, Defne et al. *Fast Charging Tests [Dataset]. Dryad. https://doi.org/10.6078/ D1MS3X*. 2015

[8] Severson *et al. Data-driven prediction of battery cycle life before capacity degradation.* Nature Energy volume 4, pages 383–391. 2019

[9] Chen,T. Guestrin,C. XGBoost: A Scalable Tree Boosting System. 2016.

[10] Venugopal, P.; T., V. *State-of-Health Estimation of Li-ion Batteries in Electric Vehicle Using IndRNN under Variable Load Condition*. Energies 2019, 12, 4338. https://doi.org/10.3390/en12224338

[11] Zheng, L., Zhu, J., Lu, D. D. C., Wang, G., and He, T. (2018): Incremental capacity analysis and differential voltage analysis based state of charge and capacity estimation for lithium-ion batteries, *Energy*, 150, 759–769. https://doi.org/10.1016/j.energy.2018.03.023

[12] Li, Y., Abdel-Monem, M., Gopalakrishnan, R., Berecibar, M., Nanini-Maury, E., Omar, N., van den Bossche, P., and Van Mierlo, J. (2018): A quick on-line state of health estimation method for Li-ion battery with incremental capacity curves processed by Gaussian filter, *Journal of Power Sources*, 373, 40–53. https://doi.org/10.1016/j.jpowsour.2017.10.092

[13] Han, X., Ouyang, M., Lu, L., and Li, J. (2014): A comparative study of commercial lithium ion battery cycle life in electric vehicle: Capacity loss estimation, *Journal of Power Sources*, 268, 658–669. https://doi.org/10.1016/j.jpowsour.2014.06.111

[14] Wang, G., Cui, N., Li, C., Cui, Z., and Yuan, H. (2023): A state-of-health estimation method based on incremental capacity analysis for Li-ion

battery considering charging/discharging rate, *Journal of Energy Storage*, 73. https://doi.org/10.1016/j.est.2023.109010

Acknowledgement

This research was supported by the PT PLN (Persero).