RSC Sustainability

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *RSC Sustainability*, 2025, **3**, 3133

# Machine learning prediction of the reversible capacities of a biomass-derived hard carbon anode for sodium-ion batteries†

Stephen Yaw Owusu *

Received 21st May 2025
Accepted 2nd June 2025

DOI: 10.1039/d5su00360a

rsc.li/rscsus

This project is among the pioneering works that incorporate machine learning (ML) modeling into the development of biomass-derived sodium-ion battery anodes for sustainable energy storage technologies. It was conceptualized and executed to satisfy a desire to use computational techniques to fill the research gap in a paper authored by Meenatchi *et al.* in 2021. The authors asserted that an activated orange peel-derived hard carbon (AOPDHC) can be used as an anode for sodium-ion batteries, yet the evidence for this claim was lacking. This work therefore sought to utilize ML to verify the claim by investigating the reversible capacities of AOPDHC at different initial coulombic efficiencies (ICE) and current densities. Data used to train the algorithms were mined from literature and applied in a 4 : 1 training-to-testing data split. Models that gave good correlations between experimental and predicted capacities for some assumed unknowns were used to predict the reversible capacities of AOPDHC. The maximum capacity obtained for AOPDHC was 341.1 mA h g$^{-1}$ at a current density of 100 mA g$^{-1}$ and an ICE of 48% and the minimum capacity was 170.3 mA h g$^{-1}$ at a current density of 100 mA g$^{-1}$ and an ICE of 43%. Lastly, the modeling found ICE to be a very important factor that influences the reversible capacities of hard carbon anodes for sodium-ion batteries, which matches literature findings, and possibly validates the modeling procedure. This study is of utmost importance since biomass-derived hard carbons are versatile, cost-effective, environmentally friendly and sustainable.

### Sustainability spotlight

This study helps to advance knowledge on how to computationally verify the potential use of biomass-derived hard carbon anodes for battery applications-exemplified with an activated orange peel-derived hard carbon (AOPDHC). The modeling procedures described in this work not only accelerate the development of novel biomass-derived materials but also provide new insights for the development of sustainable energy storage technologies. This effort is in tune with the UN Sustainable Development Goal 7: to provide affordable, reliable, sustainable, and modern energy for all. The paper mainly revolves around machine learning modeling and a few kinetic studies which can eventually create an energy-dominant circular bioeconomy and reduce waste and exposure to hazardous substances generated from conducting numerous trial-and-error experiments.

## 1. Introduction

Modern civilization has increased the demand for energy storage devices with quick response time, rapid construction and cycling flexibility.[1,2] Additionally, factors such as global warming, increasing fossil fuel consumption, and environmental pollution significantly hinder the progress of renewable energy source usage.[3,4] A challenge, which has made it inevitable to develop energy storage systems that can sustain the power output demands. Batteries can conveniently foot the bill due to their high energy capacity, high power and rapid response.[5,6] Lithium-ion batteries (LIBs) have distinguished themselves as better alternatives in the energy storage sector due to their high energy density, elevated open-circuit voltage, and cycling stability.[7,8] However, they struggle to meet the increasing demand for large scale energy storage systems because of their low natural abundance and limited global reserves. This problem has encouraged research advancements to find alternatives to LIBs. Sodium-ion batteries (SIBs) emerged as an option that can curb the shortcomings of LIBs since sodium is highly abundant and cost effective. However, it is not free from limitations. SIBs have lower energy density, and a shorter life cycle compared to LIBs. Putting all these together, it was recommended that future research on energy storage focus on fabricating anode and cathode materials for SIBs, which have higher specific capacities, higher voltages and can produce energies closer to 200 Wh kg$^{-1}$, which is similar to that

*Department of Chemistry, Missouri S&T, Rolla, MO 65409, USA. E-mail: sadnd@mst.edu*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5su00360a

of LIBs.[9] An appreciable number of cathodes for SIBs have been fabricated, however, a suitable anode is needed to facilitate their commercialization.[10] Also, a search for anodes is needed since the anode is an important factor that influences the energy density, cycling stability and efficiency of the battery.[11]

Hard carbon stands out among the various anode materials for SIBs due to their structure, which facilitates the adsorption and desorption of sodium ions. They are prepared by heating thermosetting carbon-containing precursors under an inert atmosphere (pyrolysis).[12] Hard carbons are non-graphitizable, and disordered materials. Due to their excellent sodium storage performance, alongside their low cost, low voltage and high capacity, hard carbon anodes appear to be the most likely anode to be commercialized.[13] The cost of the anode can be further reduced by using biomass as the hard carbon precursor. Biomass is obtained from organic substances such as plants, animals, and their waste products. Compared to other hard carbon precursors, biomass-derived hard carbons are highly abundant, cheap, renewable, and mostly rich in heteroatoms, which is essential for battery applications.[14] The preparation process of hard carbons is often optimized by tuning the synthesis parameters or changing the biomass precursor used. Biomass derived hard carbons prepared at pyrolysis temperatures between 1200 °C and 1400 °C have been identified to exhibit improved structural properties beneficial to sodium-ion storage performance.[12,15]

Traditional experimental methods of analyzing the electrochemical properties of hard carbons include measuring the surface area, ratio of defective to graphitic carbon, ICE, and other material properties, and correlating them with their reversible capacities at different charge densities. However, these strategies are laborious, time-consuming and costly to perform. It is therefore advantageous to use machine learning (ML), deep learning and data mining techniques to study the electrochemical properties of novel hard carbons. ML is a subdivision of artificial intelligence used to analyze large datasets. It has been extensively used in scientific research to study and establish structure–property–performance relationships, which can eventually cause significant advancements in the fabrication of novel materials.[16,17]

The ML approach has been demonstrated in this work using AOPDHC as an anode for sodium ion batteries. AOPDHC was chosen primarily to fill the research gap in a paper authored by Meenatchi *et al.*[18] Additionally, orange peels are cheap, abundant, environmentally safe, and sustainable. Compared to other biochar-based materials such as those derived from rape straw pyrolysis,[19] AOPDHC presents a significantly lower surface area ($60.16\ m^2\ g^{-1}$) as against HC from rape straw ($2046.92\ m^2\ g^{-1}$) at the same pyrolysis temperature (700 °C). This high surface area of rape straw-derived HC could hinder their applicability for sodium-ion batteries as a low surface area is required to reduce the solid electrolyte interphase (SEI), and improve ICE.[20]

Seven ML algorithms have been used in this study to predict the reversible capacities of AOPDHC as an electrode material. Material characterization data was used as input features and reversible capacity values as response or target. Features input into the models as its training conditions were ICE, pyrolysis temperature, current density, surface area, pore volume, interlayer spacing (d002), crystallite sizes (La and Lc), annealing time, heating rate and the ratio of defective to graphitic carbon. The study combined statistical and mathematical analysis to evaluate which algorithm best fits our dataset in terms of prediction accuracy. The best models were then used to predict the reversible capacities of the unknown sample (AOPDHC) at different current densities and ICEs.

Presently, a minimal number of studies have been reported, which utilize ML algorithms for investigating the performance of sodium-ion batteries. Amongst the few, this work presents a significant difference and advancement. Tianshuang *et al.* recently used ML to predict the discharge performance of hard carbon materials for sodium-ion batteries.[21] Though it was an extensive study, they didn't incorporate pyrolysis temperature and annealing time of the hard carbons as input features. Also, current density and ICE were missing from the features for their modeling. This presents a significant limitation since these factors are known to strongly influence battery capacity performance.[20] Here, the limitation has been catered for by incorporating all these features into the modeling of reversible capacities for AOPDHC. A different work conducted by Yang *et al.* used ML models to predict the specific capacitance of biomass-derived carbon materials and compared the predicted results to their corresponding experimental values in literature.[22] Their work is similar to this one as they didn't perform actual laboratory experiments to augment the modeling results. However, it is still a little beneath this work due to some significant differences in the validation approaches. In this study, the validation procedure is significantly improved by employing both random sampling and cross-validation techniques. Additionally, some known experimental values were assumed to be unknown to the algorithms and predicted by the models for validation purposes. Furthermore, just like the study conducted by Tianshuang *et al.*, Yang *et al.*'s work omitted some key features such as interlayer spacing (d002), crystallite sizes (La, and Lc), ICE, current density, and $I_D/I_G$ from the modeling, which limits its reliability and applicability to some extent. These limitations have been catered for in this study. Another notable difference in this work compared to the others is that models that gave good predictions to reversible capacities unknown to the algorithms were used to predict the actual unknowns (AOPDHC) in addition to the test and score results obtained from the modeling. Hence, this work provides new insights and approaches to test the usage of biomass-derived anodes for sustainable battery technologies.

Finally, feature datasets were investigated for their contribution and impact on the reversible capacities of AOPDHC. This was done through model accuracy analysis and was illustrated by further visualizations using shapley additive explanations analysis (SHAP). The SHAP findings were further authenticated using feature ablation. The findings from this project will hopefully provide a better understanding of the relationship between the various factors that influence the reversible capacities of biomass-derived hard carbons and will rapidly and accurately guide future experimental or computational studies through quick optimizations in the hard carbon's preparation and application process.

## 2. Methodology

### 2.1 Dataset

In mining the data, precision and quality were prioritized. Furthermore, the number of observations analyzed was more than ten times the number of features. These were carefully done to ensure the accuracy of the modeling and predictions.[23,24]

Data for 170 samples are used for all analysis in this work. 141 samples were used to train the models, and the remaining were used to validate the electrochemical data obtained for AOPDHC. Data used for this study was mined through a combination of small experimental datasets from several publications as shown in Table 1.

### 2.2 Feature selection

Factors that are generally known to significantly influence the reversible capacity of hard carbons for sodium-ion battery applications include ICE, pyrolysis temperature, current density, surface area, pore volume, interlayer spacing (d002), crystallite sizes (La and Lc), annealing time, heating rate and the ratio of defective carbon to graphitic carbon obtained from Raman analysis of the hard carbon. Given this, only publications that have data for almost all these factors were selected for the modeling of this study. The distribution in relation to each of these key influencing factors used to determine the electrochemical performance of the hard carbon anode is presented in Table 2.

Though a few of the input features (*e.g.* $I_D/I_G$) have a high standard deviation relative to their means, attempting to manipulate the data largely will affect the accuracy of study. In this study, the accuracy of the data and its' findings were prioritized compared to a good-looking statistic. Given this, all data were retained rather than replacing them with a new data that is different from the actual experimental values.

Table 1 Sources used for dataset collection

| No. | Hard carbon precursor | Abbreviation | Reference | Samples in dataset |
|---|---|---|---|---|
| 1 | Epoxy phenol novalac | EPNHC | 25 | 4 |
| 2 | Navel orange | NOHC | 26 | 3 |
| 3 | Natural cotton | HCT | 27 | 3 |
| 4 | Corn cobs | — | 28 | 3 |
| 5 | Banana peel | BPPG | 29 | 6 |
| 6 | Cedarwood bark | CBC | 30 | 3 |
| 7 | Orange peel | AOPDHC | 18 | 6 |
| 8 | Pomegranate peel | PGPC | 31 | 3 |
| 9 | Cellulose | CHC | 32 | 4 |
| 10 | Corn cobs | CC | 33 | 4 |
| 11 | Mangosteen shell | MHC | 34 | 19 |
| 12 | Rice husk | RHHC | 35 | 3 |
| 13 | Agar/urea/graphene oxide | CA | 36 | 4 |
| 14 | Cellulose nanocrystals | — | 37 | 4 |
| 15 | Polyurea-Si | PUA@Si | 20 | 4 |
| 16 | Sugarcane-bagasse | SCA | 38 | 1 |
| 17 | Phenolic resin | — | 39 | 2 |
| 18 | Resorcinol-formaldehyde | RFHC | 40 | 4 |
| 19 | Phenolic resin | — | 41 | 1 |
| 21 | Biomass starch | — | 42 | 6 |
| 22 | Biomass | — | 43 | 6 |
| 23 | Eucalyptus wood | EHC | 44 | 8 |
| 24 | Date palm |  | 45 | 10 |
| 25 | Camellia seed shell | TS | 46 | 6 |
| 26 | P-doped carbon nanofibers | CFs | 47 | 2 |
| 27 | Sycamore fruit seed | SFS | 48 | 4 |
| 28 | Maple tree | MAHC | 49 | 3 |
| 29 | F-doped hard carbon | F-HC | 50 | 3 |
| 30 | Corn straw piths | — | 51 | 4 |
| 31 | N-doped hard carbon | — | 52 | 3 |
| 32 | Walnut shell | WAHC | 53 | 6 |
| 33 | N-doped hard carbon | — | 24 | 3 |
| 33 | P-doped olive kernel | OHC | 54 | 3 |
| 34 | P-doped sisal fiber | PSHC | 55 | 1 |
| 35 | Asphalt/pecan shells | — | 56 | 2 |
| 36 | Bamboo | HCB | 57 | 4 |
| 37 | Corn starch | SCHC | 58 | 2 |
| 38 | Sawdust | HC | 59 | 2 |
| 49 | Petroleum asphalt | PHC | 60 | 3 |
| 50 | Shaddock peel | — | 61 | 4 |
| 51 | Phenolic resin | — | 62 | 3 |
| 52 | Hydroxymethylfurfural | — | 63 | 1 |

Table 2 Input features used in this study and their corresponding statistical values

| Input features | Median | Mean | Standard deviation |
|---|---|---|---|
| Reversible capacity (mA h g$^{-1}$) | 262.80 | 252.52 | 93.47 |
| $I_D/I_G$ | 1.02 | 1.84 | 7.68 |
| Surface area | 41.91 | 126.21 | 192.62 |
| Current density (mA g$^{-1}$) | 30.00 | 72.10 | 149.25 |
| Pyrolysis temperature | 1100.00 | 1138.50 | 318.15 |
| ICE | 70.00 | 66.22 | 18.18 |
| Interlayer spacing (A$^\circ$) | 3.84 | 3.85 | 0.17 |
| Crystallite size (Lc) | 1.58 | 2.75 | 3.04 |
| Crystallite size (La) | 4.29 | 5.66 | 3.37 |
| Annealing time (h) | 2.00 | 2.11 | 0.79 |
| Heating rate (°C min$^{-1}$) | 5.00 | 4.22 | 1.54 |
| Pore volume (cm$^3$ g$^{-1}$) | 0.04 | 0.12 | 0.24 |

Additionally, in dealing with outliers in a dataset for ML modeling, several factors must be considered. This includes their potential impact on the analysis. Since a factor such as $I_D/I_G$ critically affects the performance of hard carbons for battery applications, it is better to maintain them. Furthermore, attempting to remove all outliers can cause a significant reduction in the amount of data used for the study. This can affect the overall reliability of the modeling, since it is generally not acceptable to train ML algorithms with a very small dataset.

Aside from all these factors, outliers contribute to the weakness of some models in giving accurate predictions. There are other models that can effectively handle outliers and prevent them from significantly affecting the accuracy of the modeling. For this reason, some models performed better than others in the ML predictions. In the end, the best modeling results were used to draw conclusions for this work.

### 2.3 Data preprocessing

Preprocessing of data is commonly done in ML and data analysis to enhance the usefulness of the data and improve the accuracy of the modeling results.[64] To obtain only valid data for all analysis of this study, literature that had data for almost all the input features was selected. For the few missing data, the average of all data for that parameter was entered to fill the gaps. This is an acceptable procedure for numerical variables in ML modeling and was done to prevent the presence of missing data from potentially affecting the performance of the machine learning algorithms.[65]

### 2.4 Software

The Orange software was used for all computations. This software was developed by the bioinformatics lab at the University of Ljubljana, Slovenia, in collaboration with the open-source community. Orange is an open-source data mining and machine learning toolkit that can be used to visualize and analyze data.

### 2.5 Machine learning modeling

Seven machine learning models: linear regression, support vector machines (SVM), k-nearest neighbors (KNN), random forest, gradient boosting, adaboost, and trees were used to train the reversible capacities data to facilitate accurate predictions of the unknowns. These are existing programs in the orange software used for this study.

Linear regression ML model is an algorithm that models the relation between a dependent (target) and one or more independent variables (predictor). The model assumes that this relationship is linear. Due to the linearity it assumes, it is particularly useful and easy to interpret when analyzing a small dataset.[66]

Unlike the linear regression algorithm, decision trees (DT) or trees can fit complex datasets. It learns from layers of problems to arrive at relevant conclusions. Its working principle is to utilize an entire dataset and divide it by a true or false designation of a certain test condition. To get the best training process, the dataset is further divided by a true/false designation continuously for as many times as possible.[67] Decision tree is good for predictions and classification since it is not affected by data scaling.[68,69] For example, Cosgun *et al.*, used DT to optimize biomass growth and lipid yield conditions for producing renewable biofuel from microalgae, which guided their new experimental work.[70]

Random Forest (RF) also utilizes a decision tree, just like trees (it optimizes the decisions and improves on the results of simple trees by using multiple trees for its predictions). During the modeling with RF, each tree is established with independent and randomly selected samples. This randomness reduces the possibility of overfitting, which leads to more accurate predictions.[71] Feature selection is not required for this model, and this makes it particularly useful for processing high-dimensional data. Besides, unbiased estimation is used in training this model. This unbiased estimation enables it to have a strong model generalization ability.[72]

For gradient boosting algorithm, simple decision trees are combined with larger ones to make predictions. Since this model makes minimal assumptions about the data, it works better compared to random forest when dealing with large and complex data. The minimal assumptions made also cause it to have a reduced mean square error and improved $R^2$ value compared to random forest.[73] AdaBoost uses adaptive learning technique where weak learners are manipulated to favor previously misclassified data. This method makes adaboost less susceptible to overfitting compared to other algorithms. Though the individual learners may be weak, they can join to form a stronger learner if the performance of each data is better than random guessing. As a result, adaboost works best when handling a noisy data full of outliers.[74]

Support vector machine (SVM) model makes decisions on training data in a manner that results in maximizing the decision border margins in the featured space. By doing so, classification errors are minimized, and a better generalization ability is obtained making this model useful for both small and complex dataset.[75]

K-nearest neighbor algorithm classifies data points by finding the most common point to them. It basically makes predictions of data points based on the values of its neighbors and is useful for small datasets. However, it tends to be overly sensitive to unnecessary data.[76]

### 2.6 Predicting unknown reversible capacities

For the analysis conducted with ML here, the reversible capacities of various biomass-based hard carbons are predicted by first using a training set to train the models and allowing the models to use the patterns learnt to predict the unknowns. Training data is prepared in an excel spreadsheet and uploaded as a file into the orange software. The input parameters are characterized as features or categorical. In this work all input parameters are features. The file is then transposed as a data table and mapped to the various models. Model parameters are adjusted till a best fit is obtained, which can be seen in the test and score tab. After dragging from test and score to predictions, the reversible capacity for each set is predicted based on the $R^2$ values obtained for each model under test and score. To predict the reversible capacity for unknown variables, all unknowns are entered with question marks in a spreadsheet and uploaded as File 1 and Data Table 1. After mapping File 1 and the best models to Predictions 1 as illustrated in Fig. 1, the algorithms give predictions of the data for each unknown variable.

This procedure is followed to obtain data for all samples with unknown reversible capacities. For all analysis, data output from the algorithms are based on the average of a random sampling procedure, which was set to twenty repeated runs.

**2.6.1 Balancing dataset and hyperparameter tuning in running the algorithms.** An imbalanced data, where the effect of some dataset classes is severely suppressed, can lead to a bias in ML modeling and predictions. To alleviate this challenge, a balanced data split of 80% for training and 20% for validation was employed in all analysis. Also, the models were not run on only their default settings since this approach significantly hindered the performance of the models. Instead hyperparameter tuning strategies were employed. For number of trees, gradient boosting is set to 250 and random forest is set to 300. The learning rate for both models is 0.1. The number of estimators in the adaboost model is 600 and the learning rate is set to 1. The number of attributes considered at each split for random forest was 5. KNN model was given ten neighbors. For tree algorithm, subsets smaller than 5 were not split and their classification were set to stop when the majority reaches 95%. Detailed parameters used for each model prediction are given. (see ESI† (Fig. S-1)).

### 2.7 Validation of modeling results

No actual experimental data from AOPDHC is accrued in this study to verify the modeling data, however, experimental results

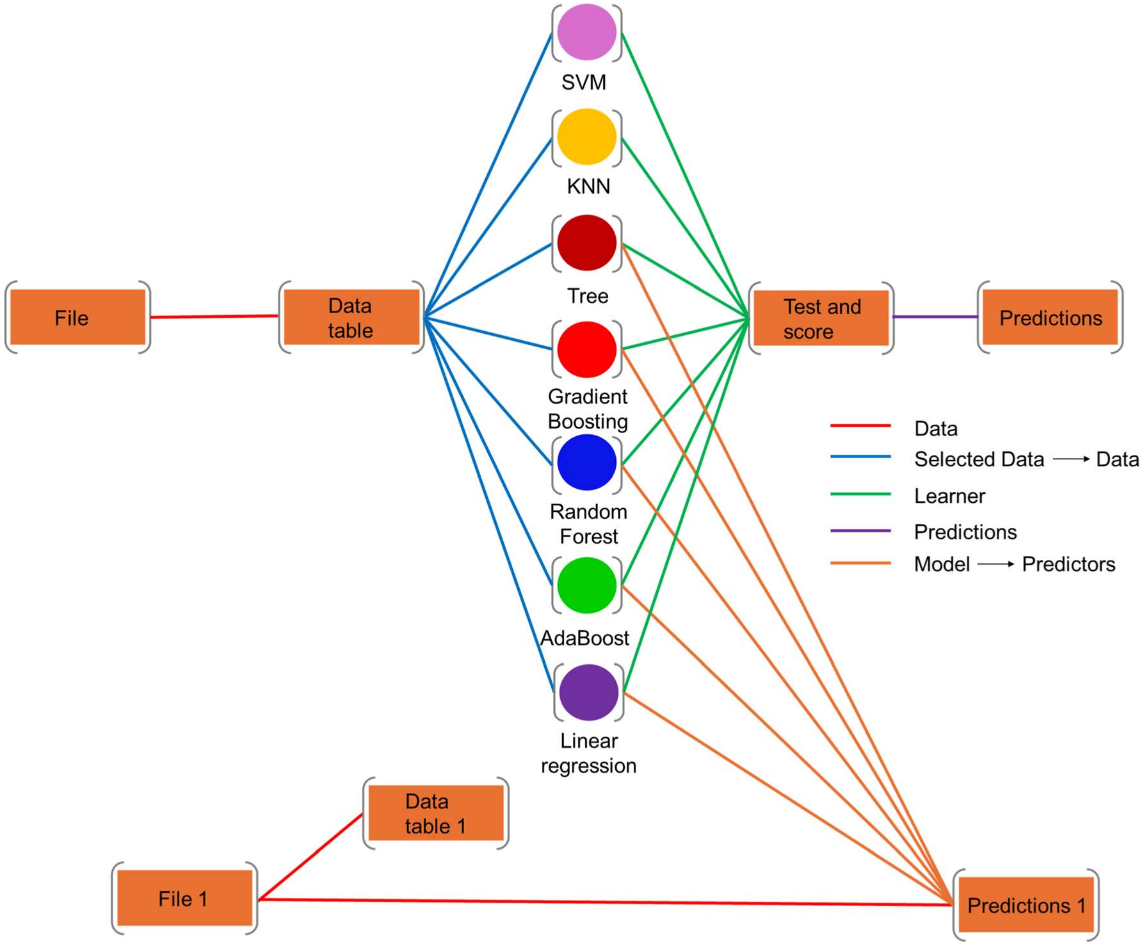

Fig. 1 Flow of machine learning modeling procedure.

from other biomass-derived hard carbons were modeled simultaneously with AOPDHC for authentication purposes (see Table S-1†). The known reversible capacities from these samples were assumed to be unknown and left blank when imputing data into the algorithms. This was done to investigate if there will be a correlation between the experimental and predicted values and thereby enhance the reliability of the modeling results. The approach described here was employed in a previous study for the first time[77] and is an improvement to other works where no known data were assumed to be unknown, but rather only a comparison between the imputed data and predicted data was used to conclude.[78]

This technique is useful since obtaining experimental data can be hindered by a lack of resources, time, effort, funding, safety and many other factors. Computational works in data science come in handy in these circumstances as it fill the gaps where it is practically impossible to obtain actual experimental data. To further enhance the validation approach of this work, both random sampling and cross-validation techniques (K-fold = 10) were employed.

### 2.8 Evaluating the impact of training features on predictions

To evaluate the impact of training features on the predictions, the relative importance of ICE, pyrolysis temperature, current density, surface area, pore volume, interlayer spacing (d002), crystallite sizes (La and Lc), annealing time, heating rate and $I_D/I_G$ on the reversible capacity of AOPDHC was evaluated and ranked. (see Fig. 5). Ranking was performed *via* two techniques, the shapley additive explanations analysis (SHAP) and the feature importance analysis *via* the ablation technique. These methods were programmed into the orange software using the explain and feature importance widgets, respectively. In the SHAP ranking, the model is trained with all features in the first interaction and values are computed for each feature. Then the widget ranks the most important features for modeling the target variable based on game theory. On the other hand, feature importance conduct ranking in a single step during the modeling.[79]

## 3. Results and discussion

### 3.1 Training of the algorithms

Pyrolysis temperature for all hard carbons used in this study ranged from 700 to 1600 °C, but was mainly concentrated around 1200 °C. This is about the average temperature mostly used for preparing hard carbons for sodium-ion batteries.[12,80] The statistical results indicate that the distribution of features is not within a very close range (based on the standard deviations from Table 2). The observation here could be because the hard carbons used to train the models are obtained from different precursors, which was done to assist in limiting the potential for a biased training of the algorithms and add to the veracity of the predictions.

The correlation between the predictions and actual values indicates that KNN and SVM are poor algorithms for this study as their $R^2$ values were below 0.5. This may result from the unbiased nature of KNN modeling, which makes it very sensitive to unnecessary dataset features.[81] The non-suitability of SVM on the other hand may be because this model assumes a balanced class,[82] which is not the case in this study. Comparatively, Fig. 3 shows a better correlation between the experimental and predicted values for random forest, gradient boosting, adaboost, and linear regression models used for the training dataset, with their $R^2$ values being greater than 0.5. These analyses were based on a random sampling technique using a balanced 80% training and 20% validation data split.

According to Table 3, which indicates the coefficient of determination values ($R^2$), and the root mean squared error (RMSE) for each model, gradient boosting, adaboost and random forest are best in predicting the reversible capacities of the training data set.

A cross-validation approach was also conducted on the predictions to ensure the accuracy of random sampling results, and its findings are given in Table 4.

Comparing both validation techniques, it is observed that cross validation (K-fold = 10) gave similar outcomes as the random sampling technique.

### 3.2 Predicting reversible capacities of assumed unknowns

All the algorithms were used to predict the reversible capacity data for select samples whose experimental reversible capacities were assumed to be unknown (see Table S-1†), and the results are presented in Fig. 2.

To get a better view of how each model performs in terms of predicting reversible capacity values that are approximately the

Table 3 The mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination ($R^2$) of the various ML models used in this study – random sampling technique employed

| Model | MSE | RMSE | MAE | MAPE | $R^2$ |
|---|---|---|---|---|---|
| Adaboost | 5252.123 | 72.472 | 50.370 | 0.263 | 0.402 |
| Random forest | 5338.342 | 73.064 | 53.134 | 0.289 | 0.392 |
| Gradient Boosting | 4680.299 | 68.413 | 47.120 | 0.255 | 0.467 |
| Tree | 7077.600 | 84.128 | 63.688 | 0.319 | 0.194 |
| SVM | 6092.398 | 78.054 | 57.319 | 0.315 | 0.307 |
| KNN | 6687.577 | 81.778 | 59.193 | 0.324 | 0.239 |
| Linear Regression | 7650.113 | 87.465 | 67.640 | 0.351 | 0.129 |

Table 4 The mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination ($R^2$) of the various ML models used in this study-cross validation technique employed

| Model | MSE | RMSE | MAE | MAPE | $R^2$ |
|---|---|---|---|---|---|
| Adaboost | 5554.037 | 74.525 | 49.970 | 0.241 | 0.361 |
| Random forest | 5530.483 | 74.367 | 52.507 | 0.260 | 0.364 |
| Gradient Boosting | 4788.885 | 69.202 | 45.886 | 0.228 | 0.449 |
| Tree | 6833.056 | 82.662 | 59.094 | 0.272 | 0.214 |
| SVM | 6327.576 | 79.546 | 57.126 | 0.283 | 0.272 |
| KNN | 6456.778 | 80.354 | 56.420 | 0.283 | 0.257 |
| Linear Regression | 7349.207 | 85.728 | 65.972 | 0.320 | 0.155 |

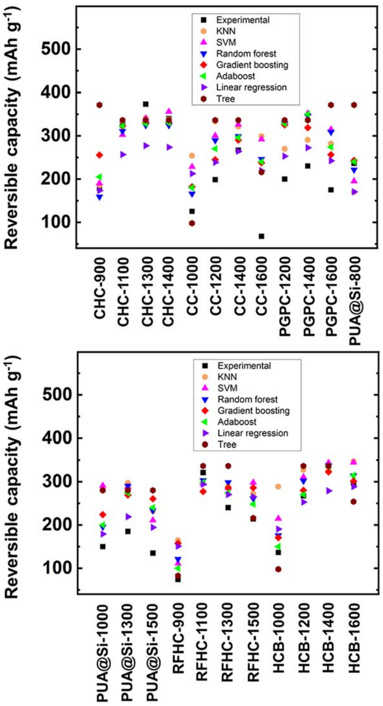

Fig. 2 Prediction of the reversible capacities of some samples whose capacities were assumed to be unknown and modeled for authentication purposes. Predicted values for reversible capacities are compared with their corresponding experimental values.

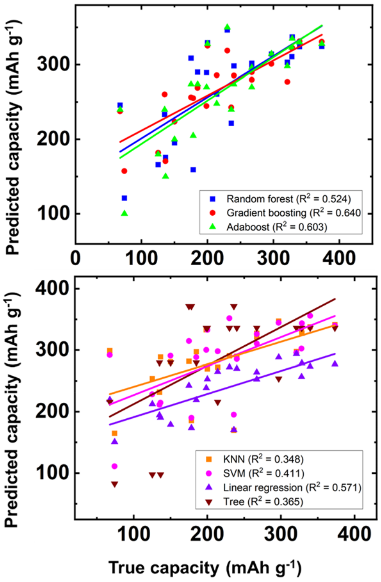

Fig. 3 Prediction of the reversible capacities of some samples whose capacities were assumed to be unknown and modeled for authentication purposes. $R^2$ values are determined to evaluate each model's performance in their potential to give accurate predictions for the assumed unknowns.

same as those obtained experimentally, data for each model was plotted independently and their correlation coefficients were determined as shown in Fig. 3.

Based on the $R^2$ values obtained, Fig. 3 shows a better predictability potential for the unknown using linear regression compared to the remaining models. However, since it failed to give good predictions during the model training, it couldn't be relied upon. As a result, the findings from Fig. 3 lead to the conclusion that gradient boosting is the best algorithm for this study.

### 3.3 Predicting reversible capacities of actual unknowns (AOPDHC)

Inspired by the good predictions obtained for the assumed unknowns, all known samples were used as training set for the algorithms to determine the reversible capacities for AOPDHC at different current densities and ICEs, and the findings are presented in Fig. 4. This procedure was followed to determine how different parameter switches affect the reversible capacity of the material and to observe any potential trends in the reversible capacities observed due to changing current densities and ICEs.

As seen from Fig. 4, the maximum capacity obtained for AOPDHC was 341.1 mA h $g^{-1}$ at a current density of 100 mA $g^{-1}$ and an ICE of 48% and the minimum capacity was 170.3 mA h $g^{-1}$ at a current density of 100 mA $g^{-1}$ and an ICE of 43%. Even at low current density (30 mA $g^{-1}$), gradient boosting could still predict a capacity greater than 250 mA h $g^{-1}$, which is like the experimental results obtained from most of the literature used here. Though the model's performance in predicting the data in the training set was average, they performed better when predicting the reversible capacities of the actual samples, which were assumed to be unknown. The data for AOPDHC were modeled simultaneously with the assumed unknowns, and since good correlations were obtained between the experimental and predicted data for the assumed unknowns, it suggests that the data obtained for the actual unknown

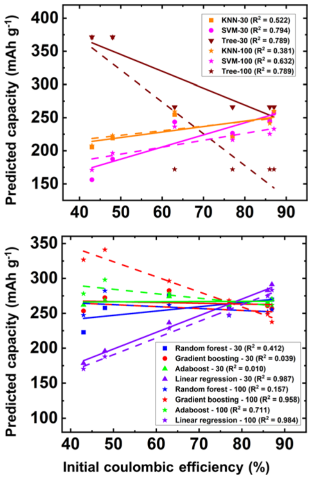

Fig. 4 Trend in reversible capacity changes due to varying ICE and current densities for all ML algorithms except SVM and KNN. For the current densities, the dashed line = 100 mA $g^{-1}$, and solid line = 30 mA $g^{-1}$.

(AOPDHC) is also reliable, regardless of the average performance of the model with the training dataset.

Computational modeling from this study *via* machine learning could therefore address the research gap left by Meenatchi, *et al.* in their manuscript[18] and proves that AOPDHC can be used as an anode for sodium-ion batteries. Though there were a few outliers, it could be seen from Fig. 4 that for each current density, improvement in ICE resulted in an increase in the reversible capacities. This is expected since for hard carbons, a higher ICE means less energy is lost during the initial charge cycle due to irreversible reactions, which translates to the storage and retrieval of more energy from the battery. This trend is like the observations made with other hard carbon systems for sodium-ion batteries.[20,83] The linear relationship seen with the models further confirms the reliability of these predictions since it is known from literature that increasing ICE leads to an increase in reversible capacity values.[36,84]

### 3.4 Ranking the impact of training features on predictions

The relative importance of ICE, pyrolysis temperature, current density, surface area, pore volume, interlayer spacing (d002), crystallite sizes (La and Lc), annealing time, heating rate and $I_D/I_G$ on the reversible capacity of the AOPDHC was modeled and ranked. Ranking was performed *via* two techniques as shown in Fig. 5. The shapley additive explanations analysis (SHAP) (using the explain model widget) and the feature importance analysis *via* the ablation technique (using the feature importance widget). In the SHAP results, the closer the color of the points to red, the higher its contribution to the model prediction and the closer the color of the points to blue, the lower its contribution to the model prediction. A SHAP value greater than zero indicates that it promotes model prediction and *vice versa*. For feature importance *via* the ablation technique, a larger decrease in $R^2$ value when a particular feature is removed during the analysis, the greater the effect of that feature on the reversible capacity predictions and *vice versa*.

Fig. 5 ranks the importance of each feature in predicting the reversible capacities of AOPDHC *via* four modeling techniques (gradient boosting, random forest, adaboost and linear regression) using the training dataset whose data summary is given in Table 2. These models were chosen since they gave good predictions of the assumed unknowns. Since some models performed better than others, it is useful to investigate how each model arrives at their predictions. The reason for the differences seen in the ranking by each model may be because each model has a different way of analyzing the data to arrive at their predictions. Hence, the rankings can provide a better understanding of the features that each model prioritizes in arriving at their conclusions. To further validate the SHAP results, feature importance *via* ablation analysis was conducted as a follow-up. In view of this, only rankings that are similar between the two techniques are used in drawing conclusions and analyzing how each model interprets feature relevance.

It is seen from Fig. 5 that, except for adaboost, all models ranked ICE as the most important factor that influences the reversible capacity predictions. This is supported by literature as a low ICE, often caused by the poor reversibility of sodiation/desodiation reaction and the decomposition of electrolytes to form SEI in the first cycle, has a direct negative effect on the battery's capacity and *vice versa*.[83]

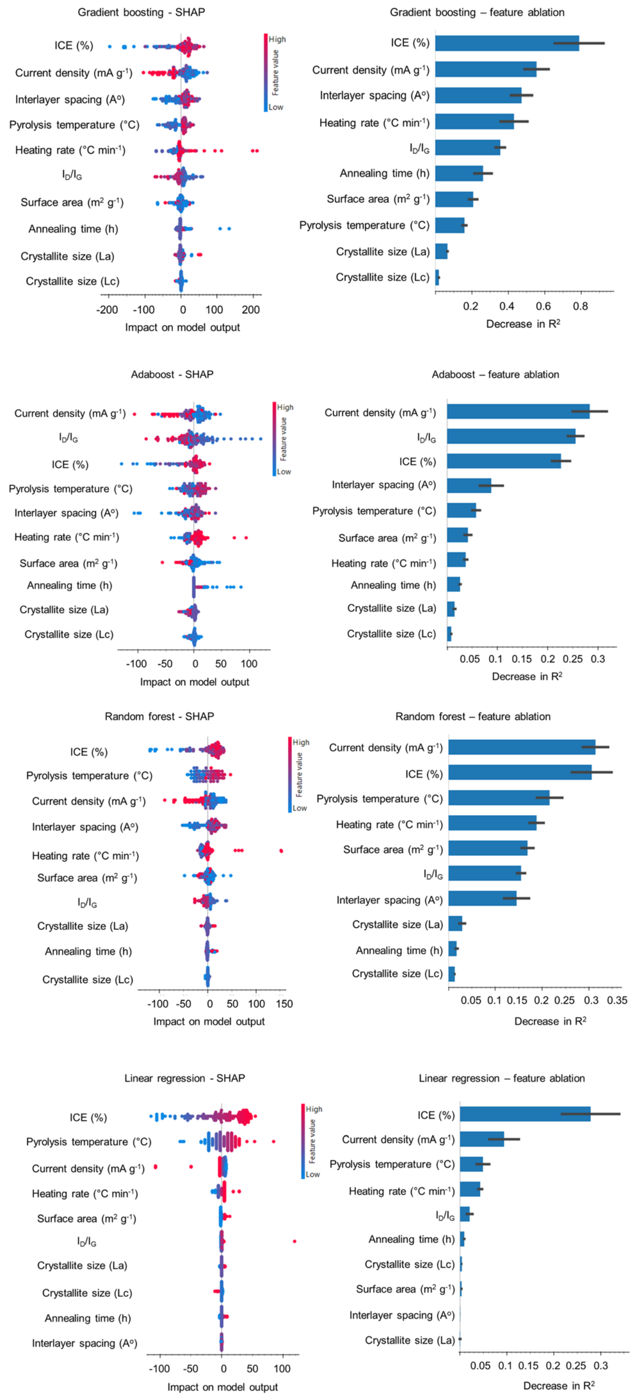

Fig. 5 Ranking the influence of features on the reversible capacity predictions as interpreted by four models with good predictabilities in this study. Ranking was performed with SHAP and feature importance *via* ablation widgets. The models were gradient boosting, adaboost, random forest and linear regression.

The second most important factor per the rankings of all models, except adaboost, was current density. This feature was ranked first by adaboost, and these observations signify the importance of current density to the reversible capacity predictions. The findings here are supported by literature, which has experimentally proven that high current densities lead to low reversible capacities.[85]

The boosting algorithms (gradient boosting and adaboost) considered interlayer spacing as an important factor, whilst linear regression ranked it among the least influential factors. Also, linear regression ranked crystallite size (Lc) over interlayer spacing, whereas for all remaining models, it was ranked the worst factor that influenced the capacity predictions. The rankings observed here with linear regression could be because the interlayer spacing values for all hard carbons were similar with a lower standard deviation relative to their mean. This may hinder the ability of the model to find a linear relationship between the input (interlayer spacing) and the output (reversible capacity). This observation differs from that seen with crystallite size (Lc) since that parameter has values with a higher standard deviation relative to their mean. The boosting algorithms on the other hand improves the performance of weak learners to build a better predictive model and therefore are powerful even in modeling complex, non-linear relationships.[86]

Adaboost ranked $I_D/I_G$ as the second most influential factor in predicting the reversible capacities, whereas it was ranked as an average influencing factor for gradient boosting. This could be because in improving weak learners to create better predictive models, adaboost starts by building short trees as compared to gradient boosting that dives deeper by starting with building leaves.[77] As a result, the high standard deviation of the $I_D/I_G$ data may have little influence on the adaboost predictions.

The two ranking techniques have different principles and hence give slightly different results. This difference could be because SHAP results not only rank features based on their importance but also based on how much each feature contributes to the models' predictions. Regardless, the overall findings are similar, and the effect of key features on each factor is identified.

Overall, it is reported that surface area has a direct relationship with ICE value, which was ranked as the most influential factor by most of the models, while pyrolysis temperature directly affects the graphitization degree of the carbon materials as reflected in their $I_D/I_G$.[25] Due to the interrelationships between all the features considered in reversible capacity predictions, it is recommended to take all these into consideration when preparing hard carbons for sodium-ion batteries regardless of their rank in this study.

# 4. Conclusion

Machine learning algorithms successfully predicted the reversible capacity of a biomass-derived hard carbon (AOPDHC) using ICE, pyrolysis temperature, current density, surface area, pore volume, interlayer spacing (d002), crystallite sizes (La and Lc), annealing time, heating rate and the ratio of defective carbon to graphitic carbon obtained from Raman analysis of the hard carbons as input features. The maximum capacity obtained for AOPDHC was 341.1 mA h g$^{-1}$ at a current density of 100 mA g$^{-1}$ and an ICE of 48% and the minimum capacity was 170.3 mA h g$^{-1}$ at a current density of 100 mA g$^{-1}$ and an ICE of 43%. The results were validated by simultaneously predicting the capacities of other samples with known experimental values alongside the unknown (AOPDHC). Upon ranking the input features based on their contribution to the predictions (using SHAP and feature importance), it was found that ICE is the most important factor that needs to be optimized to realize a high reversible capacity. The rankings therefore confirmed what is experimentally proven and thereby validated the rationale for the modeling. The correlation between the experimental and predicted values (Fig. 2 and 3), the trend of ICEs *versus* the predicted values (Fig. 4), and the match between the feature importance rankings and experimentally proven facts (Fig. 5) enhances the reliability of the modeling results. This work assists in laying a foundation for further understanding of the relationships between the factors that influence the performance of hard carbon anodes and advances knowledge on how to computationally explore the potential use of hard carbon anodes for sodium-ion batteries.

# Data availability

Data has been made available in the manuscript.

# Conflicts of interest

There are no conflicts to declare.

# References

1 B. Dunn, H. Kamath and J. M. Tarascon, *Science*, 2011, **334**, 928–935.

2 U. Akram, M. Nadarajah, R. Shah and F. Milano, *Renewable Sustainable Energy Rev.*, 2020, **120**, 109626.

3 Q. Wang, X. Zhu, Y. Liu, Y. Fang, X. Zhou and J. Bao, *Carbon*, 2018, **127**, 658–666.

4 O. Ellabban, H. Abu-Rub and F. Blaabjerg, *Renewable Sustainable Energy Rev.*, 2014, **39**, 748–764.

5 A. Aghmadi and O. A. Mohammed, *Batteries*, 2024, **10**, 141.

6 X. Fan, B. Liu, J. Liu, J. Ding, X. Han, Y. Deng, X. Ly, Y. Xie, B. Chen, W. Hu and C. Zhong, *Trans. Tianjin Univ.*, 2020, **26**, 92–103.

7 B. Diouf and R. Pode, *Renew. Energy*, 2015, **76**, 375–380.

8 G. Huang, H. Zhang, F. Gao, D. Zhang, Z. Zhang, Y. Liu, Z. Shang, C. Gao, L. Luo, M. Terrones and Y. Wang, *Carbon*, 2024, **228**, 119354.

9 K. M. Abraham, *ACS Energy Lett.*, 2020, **5**, 3544–3547.

10 E. Goikolea, V. Palomares, S. Wang, I. R. Larramendi, X. Guo, G. Wang and T. Rojo, *Adv. Energy Mater.*, 2020, **10**, 2002055.

11 G. Huang, H. Zhang, F. Gao, D. Zhang, Z. Zhang, Y. Liu, Z. Shang, C. Gao, L. Luo, M. Terrones and Y. Wang, *Carbon*, 2024, **228**, 119354.

12 M. Thompson, Q. Xia, Z. Hu and X. S. Zhao, *Mater. Adv.*, 2021, **2**, 5881–5905.

13 J. Liu, L. Wang, Z. Huang, F. Fan, L. Jiao and F. Li, *Chem. Pap.*, 2022, **76**, 7465–7473.

14 B. Zhong, C. Liu, D. Xiong, J. Cai, J. Li, D. Li, Z. Cao, B. Song, W. Deng, H. Peng, H. Hou, G. Zou and X. Ji, *ACS Nano*, 2024, **18**, 16468–16488.

15 Y. Jin, Z. Shi, T. Han, H. Yang, H. D. Asfaw, R. Gond, R. Younesi, P. G. Jönsson and W. Yang, *Processes*, 2023, **11**, 764.

16 I. H. Sarker, *SN Comput. Sci.*, 2021, **2**, 160.

17 B. Oral, B. Tekin, D. Eroglu and R. Yildirim, *J. Power Sources*, 2022, **549**, 232126.

18 T. Meenatchi, V. Priyanka, R. Subadevi, W.-R. Liu, C.-H. Huang and M. Sivakumar, *Carbon Lett.*, 2021, **31**, 1033–1039.

19 S. Shao, L. Ma, X. Li, H. Zhang and R. Xiao, *Ind. Crops Prod.*, 2023, **192**, 115912.

20 S. Sundaramoorthy, R. U. Soni, S. Y. Owusu, S. Bhattacharya, A. B. M. S. Doulah, V. A. Edlabadkar, C. Sotiriou-Leventis and A. Choudhury, *ACS Appl. Energy Mater.*, 2024, **7**, 9289–9299.

21 T. Qi, X. Zhang, K. Xiong, H. Yang, S. Zhang and H. Chen, *J. Mater. Chem. A*, 2025, DOI: 10.1039/d5ta01528f.

22 X. Yang, C. Yuan, S. He, D. Jiang, B. Cao and S. Wang, *Fuel*, 2023, **331**, 125718.

23 Z. Li, J. Yoon, R. Zhang, F. Rajabipour, W. V. Srubar III, I. Dabo and A. Radlińska, *NPJ Comput. Mater.*, 2022, **8**, 127.

24 N. Sun, R. Zhao, M. Xu, S. Zhang, R. A. Soomro and B. Xu, *J. Power Sources*, 2023, **564**, 232879.

25 C. Fan, R. Zhang, X. Luo, Z. Hu, W. Zhou, W. Zhang, J. Liu and J. Liu, *Carbon*, 2023, **205**, 353–364.

26 Y. Gao, S. Piao, C. Jiang and Z. Zou, *Diamond Relat. Mater.*, 2022, **129**, 109329.

27 Y. Li, Y.-S. Hu, M.-M. Titirici, L. Chen and X. Huang, *Adv. Energy Mater.*, 2016, **6**, 1600659.

28 P. Liu, Y. Li, Y.-S. Hu, H. Li, L. Chen and X. Huang, *J. Mater. Chem. A*, 2016, **4**, 13046–13052.

29 E. M. Lotfabad, J. Ding, K. Cui, A. Kohandehghan, W. P. Kalisvaart, M. Hazelton and D. Mitlin, *ACS Nano*, 2014, **8**, 7115–7129.

30 M. Lu, Y. Huang and C. Chen, *Energy Fuels*, 2020, **34**, 11489–11497.

31 M. G. Karthick Babu, R. Sampath, D. Kumar and K. Ramesha, *Ionics*, 2025, DOI: 10.1007/s11581-024-06048-5.

32 L. Qin, S. Xu, Z. Lu, L. Wang, L. Chen, D. Zhang, J. Tian, T. Wei, J. Chen and C. Guo, *Diamond Relat. Mater.*, 2023, 110065.

33 N.-J. Song, N. Guo, C. Ma, Y. Zhao, W. Li and B. Li, *Molecules*, 2023, **28**, 3595.

34 K. Wang, Y. Jin, S. Sun, Y. Huang, J. Peng, J. Luo, Q. Zhang, Y. Qiu, C. Fang and J. Han, *ACS Omega*, 2017, **2**, 1687–1695.

35 T. Wang, D. Su, D. Shanmukaraj, T. Rojo, M. Armand and G. Wang, *Electrochem. Rev.*, 2018, **1**, 200–237.

36 H. Wang, S. Liu, C. Lei, H. Qiu, W. Jiang, X. Sun, Y. Zhang and W. He, *Electrochim. Acta*, 2024, 143812.

37 Y. E. Kim, S. J. Yeom, J.-E. Lee, S. Kang, H. Kang, G.-H. Lee, M. J. Kim, S. G. Lee, H.-W. Lee and H. G. Chae, *J. Power Sources*, 2020, **468**, 228371, DOI: 10.1016/j.jpowsour.2020.228371.

38 Y. Yang, Z. Wu, J. Yao, T. Guo, F. Yang, Z. Zhang, J. Ren, L. Jiang and B. Li, *Energy Rev.*, 2024, 100068.

39 N. Sun, R. Zhao, M. Xu, S. Zhang, R. A. Soomro and B. Xu, *J. Sources*, 2023, **564**, 232879.

40 Q. Zhang, X. Deng, M. Ji, Y. Li and Z. Shi, *Ionics*, 2020, **26**, 4523–4532.

41 A. Beda, P.-L. Taberna, P. Simon and C. Matei Ghimbeu, *Carbon*, 2018, **139**, 248–257.

42 Y. Chen, F. Li, Z. Guo, Z. Song, Y. Lin, W. Lin, L. Zheng, Z. Huang, Z. Hong and M.-M. Titirici, *J. Power Sources*, 2023, **557**, 232534.

43 T. K. Kumaresan, S. A. Masilamani, K. Raman, S. Z. Karazhanov and R. Subashchandrabose, *Electrochim. Acta*, 2021, **368**, 137574.

44 M. Liu, Y. Wang, F. Wu, Y. Bai, Y. Li, Y. Gong, X. Feng, Y. Li, X. Wang and C. Wu, *Adv. Funct. Mater.*, 2022, 2203117.

45 I. Izanzar, M. Dahbi, M. Kiso, S. Doubaji, S. Komaba and I. Saadoune, *Carbon*, 2018, **137**, 165–173.

46 Y. Jia, X. Chen, H. Lu, F. Zhong, X. Feng, W. Chen, X. Ai, H. Yang and Y. Cao, *J. Phys. D:Appl. Phys.*, 2020, **53**, 414002.

47 J. Feng, F. Cai, Y. Zhao, X. Zhang, X. Zhan and S. Wang, *Energy*, 2024, 131474.

48 G. Zhang, Y. Zhao, L. Yan, L. Zhang and Z. Shi, *J. Mater. Sci.*, 2021, **32**, 5645–5654.

49 Y. Wang, Z. Feng, W. Zhu, V. Gariépy, C. Gagnon, M. Provencher, D. Laul, R. Veillette, M. Trudeau, A. Guerfi and K. Zaghib, *Materials*, 2018, **11**, 1294.

50 L. Kong, Y. Li and W. Feng, *Trans. Tianjin Univ.*, 2022, **28**, 123–131.

51 Y.-E. Zhu, H. Gu, Y.-N. Chen, D. Yang, J. Wei and Z. Zhou, *Ionics*, 2017, **24**, 1075–1081.

52 G. Huang, q. kong, W. Yao and Q. Wang, *ChemSusChem*, 2023, **16**, e202202070.

53 S. Zhang, Y. Li and M. Li, *JOM*, 2018, **70**, 1387–1391.

54 P. Liang, Z. Huo, Y. Liu, Z. Bo, Y. Wu, X. Hu and Z. Wen, *Batteries Supercaps*, 2025, e202400694.

55 Y. Wang, Y. Luo, X. Li, S. Lin, Y. Qin, K. Guo, L. Liao, W. Wang, K. Zhang and A. Qin, *J. Mater. Sci.*, 2025, **36**, 581.

56 Y. Yin, Y. Tan, Y. Lu, Y. Wang, J. Yang, Y. Li and B. Huang, *J. Energy Storage*, 2025, **113**, 115649.

57 T. Gao, Y. Zhou, Y. Jiang, Z. Xue and Y. Ding, *Diamond Relat. Mater.*, 2024, **150**, 111737.

58 Q. Xin, Y. Feng, S. Gan, X. Deng, Z. Feng, D. Xiong and M. He, *Ionics*, 2025, **31**, 4309–4320.

59 Z.-T. Liu, T.-H. Hsieh, C.-W. Huang, M.-L. Lee, W.-R. Liu and J. Inst, *Chem. Eng.*, 2023, 104889.

60 H. Wang, S. Liu, C. Lei, H. Qiu, W. Jiang, X. Sun, Y. Zhang and W. He, *Electrochim. Acta*, 2024, 143812.

61 N. Sun, H. Liu and B. Xu, *J. Mater. Chem. A*, 2015, **3**, 20560–20566.

62 A. Kamiyama, K. Kubota, T. Nakano, S. Fujimura, S. Shiraishi, H. Tsukada and S. Komaba, *ACS Appl. Energy Mater.*, 2019, **3**, 135–140.

63 E. O. Eren, E. Senokos, Z. Song, B. Mondal, A. Perju, T. Horner, E. B. Yılmaz, E. Scoppola, P.-L. Taberna, P. Simon, M. Antonietti and P. Giusto, *Mater. Horiz.*, 2025, **12**, 886–898.

64 T. Qi, X. Zhang, K. Xiong, H. Yang, S. Zhang and H. Chen, *J. Mater. Chem. A*, 2025, DOI: 10.1039/d5ta01528f.

65 T. Makaba and E. Dogo, in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, IEEE, 2019.

66 T. M. Hope, Linear regression, in *Machine Learning*, Academic Press., London, 2020, pp. 67–81.

67 Q. Tao, Z. Li, J. Xu, N. Xie, S. Wang and J. A. K. Suykens, *Expert Syst*, With Appl., 2020, p. 114214.

68 X. Yang, S. Sun and X. Zhao, *Mater. Lett.*, 2024, **377**, 137414.

69 A. Coşgun, M. E. Günay and R. Yıldırım, *Renewable Energy*, 2021, **163**, 1299–1317.

70 X. Yang, C. Yuan, S. He, D. Jiang, B. Cao and S. Wang, *Fuel*, 2023, **331**, 125718.

71 Z. Ullah, M. khan, S. Raza Naqvi, W. Farooq, H. Yang, S. Wang and D.-V. N. Vo, *Bioresour. Technol.*, 2021, **335**, 125292.

72 J. Wang, Z. Xu, J. Eloi, M. Titirici and S. J. Eichhorn, *Adv. Funct. Mater.*, 2022, **32**, 2110862.

73 T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani and J. Friedman, Boosting and additive trees, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2009, pp. 337–387.

74 Y. Cao, Q.-G. Miao, J.-C. Liu and L. Gao, *Acta Mech. Sin.*, 2014, **39**, 745–758.

75 J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua and A. Lopez, *Neurocomputing*, 2020, **408**, 189–215.

76 N. S. Altman, *Am. Stat.*, 1992, **46**, 175–185.

77 S. Y. Owusu, M. Amo-Boateng and R. U. Soni, *J. Mater. Chem. B*, 2025, **13**, 6233–6245.

78 X. Yang, C. Yuan, S. He, D. Jiang, B. Cao and S. Wang, *Fuel*, 2023, **331**, 125718.

79 H. Wang, Q. Liang, J. T. Hancock and T. M. Khoshgoftaar, *J. Big Data*, 2024, **11**, 8.

80 Y. Cheng, J. Zhao, L. Zhang, J. Wan, J. Yang and H. Wang, *Ion*, 2023, **402**, 116374.

81 S. Uddin, I. Haque, H. Lu, M. A. Moni and E. Gide, *Sci. Rep.*, 2022, **12**, 6256.

82 R. Guido, S. Ferrisi, D. Lofaro and D. Conforti, *Information*, 2024, **15**, 235.

83 Y. Wan, Y. Liu, D. Chao, W. Li and D. Zhao, *Nano Mater. Sci.*, 2023, **5**, 189–201.

84 S. Guo, Y. Chen, L. Tong, Y. Cao, H. Jiao, Z. long and X. Qiu, *Electrochim. Acta*, 2022, **410**, 140017.

85 C. Chen, R. Agrawal and C. Wang, *Nanomaterials*, 2015, **5**, 1469–1480.

86 L. Yang, Y. Liang, Q. Zhu and X. Chu, *Ann. GIS*, 2021, **27**, 273–284.