BIBECHANA
Vol. 21, No. 2, August 2024, 159-170

ISSN 2091-0762 (Print), 2382-5340 (Online)
Journal homepage: http://nepjol.info/index.php/BIBECHANA
Publisher: Dept. of Phys., Mahendra Morang A. M. Campus (Tribhuvan University) Biratnagar

Comparative study of machine learning based prediction of supercapacitance performance of activated carbon prepared from bio-based materials

Kirti Bir Rajguru[1], Sujan Bhandari[2], Ganesh Kumar Shrestha[1], Chhabi Lal Gnawali[1,*], Bhadra Prasad Pokharel[1]

[1]*Department of Applied Sciences and Chemical Engineering, Pulchowk Campus, Institute of Engineering (IOE), Tribhuvan University, Lalitpur, Nepal*
[2]*Central Department of Physics, Tribhuvan University, Kirtipur, Kathmandu, Nepal*

[*]Corresponding author: Email: chhabig123@ioe.edu.np

Abstract

The performance of electrochemical double-layer capacitors (EDLCs) is evaluated by the capacitance of activated carbon (AC) electrodes. The capacitance of AC electrodes is influenced by many factors such as precursor type, activation method, pore structure, surface chemistry and electrolytic properties. In this paper, we present a comparative study of machine learning based prediction of surface area, mesopore volume and total pore volume of activated carbon for energy storage applications. The ML models were trained on a dataset of synthetic data that were generated from the limited number of experimental data and which included the activation temperature, methylene blue number and iodine number of the activated carbon (AC). The best performing ML model was Random Forest and XGboost model. The results of this study can be used to optimize the production of activated carbon and improve its performance in energy storage applications.

Keywords
Machine Learning; Activated Carbon; Energy storage; Capacitance.

Article information
Manuscript received: January 31, 2024; Revised: April 18, 2024; Accepted: April 23, 2024
DOI https://doi.org/10.3126/bibechana.v21i2.62465

1 Introduction

A data based method called the machine learning algorithm has been applied as a substitute tool to deal with a number of real-world problems. Creating models and techniques that enable computers to learn from data and make predictions or judgments without being explicitly programmed is a key component of the artificial intelligence field of machine learning [1, 2]. It is utilized in a variety of programs, including recommendation engines, natural language processing, and identifying pictures. There are various types of machine learning such as unsupervised machine learning and supervised machine learning. In supervised machine learning, the algorithm develops predictions or classifications based on input data after learning from labeled training data. Unsupervised machine learning uses an unlabeled data collection as its foundation [3].

Supercapacitors, also known as electrochemical capacitors, are an intermediate form of energy storage between batteries and ordinary capacitors. Supercapacitors can store more energy than regular capacitors due to their substantially higher energy density. Supercapacitors have received a lot of interest recently, due to its potential applications in a wide range of fields, such as consumer electronics and renewable energy systems. Whereas batteries, store energy chemically, supercapacitors do so electrostatically [4, 5]. Supercapacitors rely on the rapid movement of ions to store and release energy quickly. Larger surface area and appropriate pore volumes enhance the device's energy and power density. Optimizing these parameters help in designing energy storage materials that can store more energy, charge and discharge faster, and has improved overall performance and durability in various applications [6–9].

Fossil fuels can lead to energy crisis due to their finite supply and environmental impact. Transitioning to renewable energy source is essential to mitigate these issues. Activated carbon is highly porous form of carbon with a large surface area. Activated carbon (AC) is prepared from various carbonaceous material such as peat, wood, etc. In this study, the methylene blue number (MBN), iodine number and temperature were utilized to predict the surface area, mesopore volume, and total pore volume of produced activated carbon by using machine learning algorithm [10, 11]. A larger surface enhances the size of room for interaction between the electrolyte and electrode material. More active places for electrochemical processes to take place can be made possible by this, increasing the amount of energy stored and accelerating the rate of charge/discharge. Larger pores called mesopores allow ions to pass through them and enter and exit the material [12]. A higher mesopore volume ensures better ion accessibility and diffusion within the material, which improves the overall charge/discharge efficiency and, the device's power performance. Total Pore Volume includes all sizes of pores, from microspores to mesopores. It is a measure of the overall storage capacity for ions [13, 14]. This research employs statistical and experimental data, to assess the most suitable model for the prediction of more accurate dataset. The aspiration is to employ machine learning techniques which enhance comprehension of the interplay between different aspects of activated carbon and surface area, mesopore volume and total pore volume. This, in turn, will streamline and enhance the precision of experimental guidance for future experiments [15, 16].

1.1 Machine Learning Models

There are different types of machine learning models that can be used to predict the electrochemical properties of activated carbon and its features, as they can capture the non-linearity in the data. Among all these models, we selected those that were most suitable for our data type so that we can predict the electrochemical features of the AC with high degree of accuracy. Various studies like that of Ziang, Su, Wang, Donthula has been done in the use of machine learning for prediction of different types of properties of activated carbon [17].

2 Materials and Methods

2.1 Experimental

Terminalia chebula seed stones were used for the preparation of activated carbon(AC). Terminalia seeds were collected from the market and cleaned to remove any impurities. Phosphoric acid was used as an activating agent. Terminalia seeds powder was mixed in a phosphoric acid solution. The impregnation process ensured that the seeds powder absorbed the acid solution uniformly. The impregnated powder was subjected to tubular furnace, typically in the range of 400°C- 700°C. It helped in the formation of development of pores and the removal of the volatile components. The phosphoric acid acted as a dehydrating agent and helped the carbon substance (activated carbon) to develop holes. To get rid of any remaining acids or other contaminants, the prepared activated carbon was thoroughly rinsed with distilled water. The synthesized activated carbon (AC) was analyzed by determining its iodine number, methylene blue number, surface area, mesopore volume and total pore volume [18–20]. We took eight data for each feature of methylene blue number, iodine number, surface area, mesopore volume, and total pore volume.

2.2 Computational

2.2.1 Data Collection

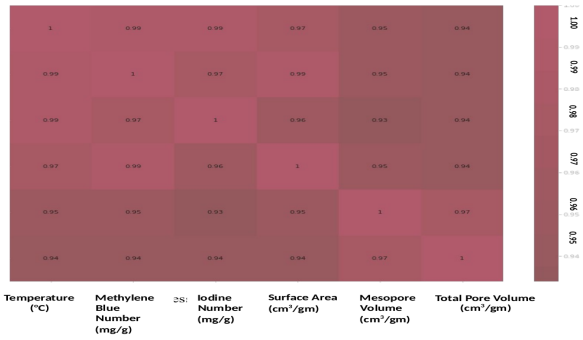

Figure 1: Heatmap of regression values of different input and output features of AC.

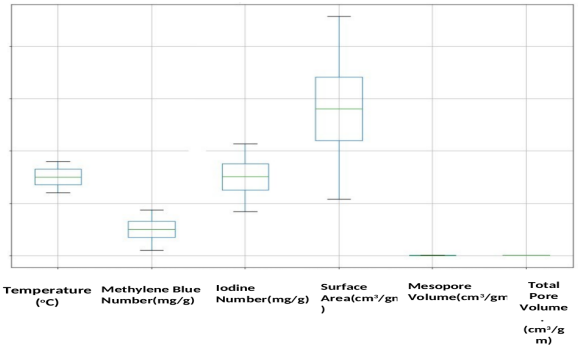

Figure 2: Box plot featuring range of input and output values of features of the AC.

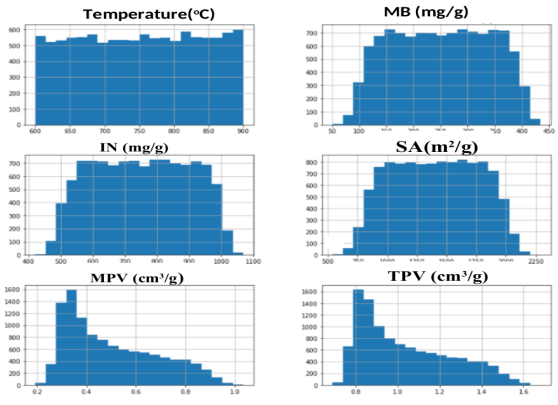

Figure 3: Histogram of all the inputs and output features.

The data from the experimental method was used to define the range and variance of each of the resultant values obtained using formulas for specific capacitance, surface area, mesopore volume, and total pore volume.

The synthetic data was generated using the algorithm of producing random number but in range and variance obtained from the experimental data. The large number of synthetic data generation is crucial for machine learning as it helps to expand the dataset, making the model more robust and capable of better generalization. This broader dataset can lead to improved training and more accurate predictions. Machine learning algorithms often require a significant amount of diverse data to effectively learn patterns and relationships. The details of the generated data can be studied using the heat map, the box plot and histogram shown in figure 1, 2 and 3 respectively.

By creating synthetic data that closely resembles the original experimental data, we provide the algorithm with more examples which will potentially enhance its performance. The large number of data were generated. To generate a normal random distribution value firstly, we limited the range of input and out values by using an algorithm function. Finally, the generated value is scaled to fit within the range as mentioned above. These techniques ensured that the synthetic data closely resembles our experimental data [21, 22].

2.2.2 Machine Learning Process

Firstly, the data were checked for missing values and missing values were computed using mean. In the following process, firstly the importance of input features and their order of importance were noted for each output feature as shown in the figure 3. Methylene blue number was the most important parameter for the prediction of mesopore volume followed by temperature and iodine number respectively. Finally, iodine number, methylene blue number and temperature in the same order are important for prediction of total pore volume. To ensure fair treatment of features, we applied scaling to bring all the features to a similar scale.

Random Forest model was used to calculate feature importance. The normalized reduction in the criterion brought by each feature was guided our selection process.

A variety of machine learning algorithms were chosen based on their relevance to the task. The dataset was partitioned into training and validation sets in the ratio of 80:20. Subsequently, each algorithm underwent training using the training data, enabling them to grasp underlying patterns [23].

Analysis of the results provided insights into algorithm performance, guiding the refinement of the approach for improved outcomes in subsequent iterations [24].

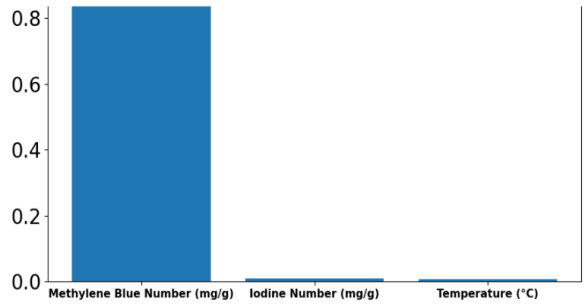

Figure 4: Feature importance of input variables for output of ‘’Surface Area".

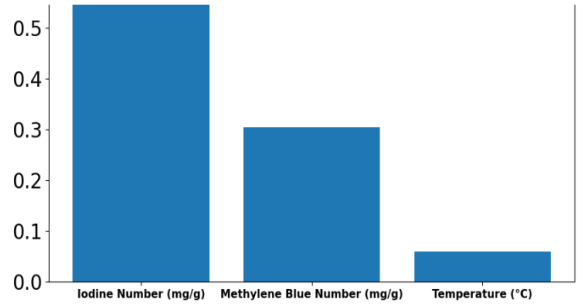

Figure 5: Feature importance of input variables for output of Total Pore Volume.

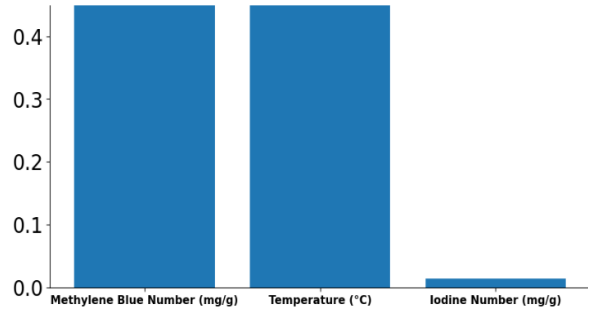

Figure 6: Feature importance of input variables for output of ‘ ’Mesopore Volume’ ’.

2.2.3 Proposed Model Selection

There are several machine learning algorithms that are suitable for regression tasks. In this paper following machine learning algorithms were described.

Linear Regression

A fundamental approach model connects between input variable and a continuous output using a linear equation. Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The rationale behind linear regression is being simple and interpretable and can be used for robust regression and classification tasks. It assumes a linear relationship between the variables, attempting to find the best-fitting straight line (or hyperplane in higher dimensions) through the data points [25]. If the true relationship between the input feature and surface area is not linear, linear regression may not capture it. The goal of linear regression is to find the coefficients that minimize the sum of squared differences between the predicted values and the actual values in the training data. This is often done by using techniques like Ordinary Least Squares (OLS) or gradient descent [26].

Ridge Regression

Ridge regression, a regularization method in linear regression, introduces a penalty term based on the L2 norm of coefficients to mitigate overfitting and multi-collinearity [27].

Lasso Regression

Similar to Ridge, Lasso adds a penalty, but it tends to create simpler models by encouraging some feature contributions to become zero. Both Ridge and Lasso are used when dealing with multicollinear data. Both Lasso and Ridge assume linearity and independence. Regularization helps to mitigate overfitting, but their performance depends on the true underlying relationship [28].

Decision Tree Regression

This method dissects data into hierarchical decision structures to predict the target output. A decision tree is a widely used machine learning algorithm for both classification and regression tasks [29]. It is a tree-like model where an input is passed through a series of binary decisions based on the values of its features. Each decision node in the tree represents a feature, and the branches from that node represents the possible feature values. The leaf nodes of the tree contain the predicted output or class. When a new input is given, it traverses the tree based on the feature values, following the decisions at each node, and finally arrives at a leaf node that provides the predicted class (classification) or value (Regression). Decision Trees tend to over-fit noisy data and feature importance may not accurately reflect the true

relevant features of data.

Random Forest Regression

An ensemble technique constructs multiple decision trees and averages their predictions, by enhancing accuracy and avoiding overfitting. Random Forest is a machine learning algorithm used for both classification and regression tasks. It is an ensemble method that builds multiple decision trees during training and combines their outputs to make predictions. By introducing randomness in tree construction and feature selection, Random Forest reduces overfitting and improves generalization. It is a powerful and versatile algorithm often used for complex datasets and can provide valuable insights into feature importance. Random Forests addresses decision tree limitations by averaging predictions across multiple trees. They handle non-linearites better but may still struggle with subtle relationships [30].

Support Vector Regression (SVR)

Support Vector Regression is to used determine a regression line that best suits the data, allowing some flexibility around the line. Support Vector Regression (SVR) is a machine learning algorithm that is used for regression tasks. It is an extension of Support Vector Machines (SVM) and is particularly useful while dealing with continuous or numeric target variables. SVR works by finding a hyperplane that best fits the data points which minimizes the error within a certain margin.SVR uses different kernel functions (such as linear, polynomial, radial basis function, etc.) to map the input data into a higher dimensional space, which can help to capture complex relationship between variables. SVR can model non-linear relationships effectively. The choice of kernel function impacts the performance [31].

Gradient Boosting Regression

An ensemble strategy that sequentially builds weak learners, with each one rectifying the errors of its predecessor. Gradient boosting adapts well to complex patterns but might over fit if not tuned properly.

Cat Boost Regression

A gradient boosting algorithm proficient in handling categorical features; can be configured to run without displaying training progress. Cat Boost is a gradient boosting machine learning algorithm which particularly well suits for categorical features. It stands for "Categorical Boosting." similar to other gradient boosting methods, Cat Boost also builds an ensemble of decision trees to make predictions. Cat Boost also includes features like automatic handling of missing values, built-in cross validation, and the ability to monitor training progress. It has gained popularity for its ease of use, efficiency, and competitive performance on various machine learning tasks [32].

XG Boost Regression

Another robust gradient boosting algorithm designed to overcome limitations of traditional gradient boosting. It is important to assess dataset size, feature complexity, interpretability, and computational resources while choosing the most suitable algorithm for research paper.

2.2.4 Model Evaluation

In this research, a comprehensive evaluation of the model's performance is carried out to assess its generalizability and stability. The technique of cross-validation is employed, which involves partitioning the dataset into subsets for both training and validation purposes. This allows us to test the model's ability to perform consistently across different subsets of data, ensuring that it can handle a variety of scenarios effectively. To gauge the accuracy and reliability of the model, appropriate evaluation methods are applied. These methods could encompass metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and possibly more domain specific metrics tailored to the nature of the dataset. These metrics collectively offer insights into the model's predictive performance and its closeness to the actual experimental data [33] A crucial aspect of this analysis involves a comparison between the actual experimental data and the model's predicted values for specific characteristics like capacitance, surface area, mesopore volume, micro pore volume, and total pore volume. By contrasting these values, researchers can quantify the model's accuracy in predicting intricate material properties. The degree of agreement between the predicted and actual values will offer insights into the model's capability to capture the underlying relationships within the data. In the context of Terminalia chebula seeds, the effectiveness of the model in predicting the material's characteristics is a focal point of analysis. The findings obtained through rigorous evaluation methods shed light on how well the model can cap-

ture the nuanced features of the activated carbon derived from these seeds. Researchers will interpret the results to discern the strengths and limitations of the model in reproducing the complex properties of the material. This interpretation forms a critical part of the paper, as it guides the reader in understanding the implications of the study's outcomes and its broader implications for the field. CatBoost and XGBoost combine ensemble power with categorical feature handling. They perform well but require tuning [34,35].

2.2.5 Model Evaluation Parameters

In the field of predictive modeling and statistical analysis, the selection of appropriate performance metrics is of paramount importance. This choice determines how effectively a model's predictions are evaluated and compared against actual values. Four commonly employed metrics for this purpose are Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared [36].

Mean Squared Error

Mean Squared Error is a widely utilized metric that quantifies the average squared differences between the model's predicted values and the actual observed values. By squaring the errors, MSE emphasizes larger deviations and penalizes them more than smaller errors. This characteristic makes MSE sensitive to outliers and may lead to models that are overly focused on minimizing large errors [37]. The following formula is used to calculate the mean squared error:

$$MSE = \frac{1}{n}\sum_{i}^{n}(Y_i - \bar{Y}^2) \quad (1)$$

were, MSE= Mean squared error
n= number of data points
Y_i= Observed values
$\hat{Y}_i$= Predicted values

Root Mean Squared Error

Root Mean Square Error is derived from MSE, RMSE is the square root of the average of squared errors. It has an advantage of presenting errors in the same unit as the target variable, making it more interpretable. RMSE offers insights into the spread or dispersion of prediction errors. While it retains the sensitivity to outliers, the square root transformation mitigates the influence of extreme values [38, 39]. The formula to calculate root mean square error is given as:

$$RMSE = \sqrt{MSE} \quad (2)$$

Mean Absolute Error

Mean Absolute Error calculates the average of the absolute differences between predicted and actual values. Unlike MSE, it treats all errors equally, making it less sensitive to outliers. This characteristic makes MAE a favorable choice where outliers should not disproportionately affect the evaluation. However, it may lack the emphasis on larger errors that MSE provides. The following formula used to calculate the mean squared error:

$$MAE = \frac{\sum_{i}^{n}(Y_i - \bar{Y}_i)}{n} \quad (3)$$

where,
MAE=Mean absolute error
y_i=Prediction
x_i= True value
n= Total number of data points

R-squared

R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model [40]. It offers insights into the goodness of fit and the model's explanatory power. Higher R-squared values indicate a larger portion of the variability in the data is accounted by the model. Nevertheless, it is important to note that R-squared can be misleading when applied to complex models, as it tends to increase with the addition of more variables, even if they are not meaningful [40]. The following formula used to calculate the R-squared:

$$R^2 = 1 - \frac{SSR}{SST} \quad (4)$$

While selecting the appropriate metric for evaluating a model's performance, researcher's must consider the nature of the data, the goals of the analysis, and the desired balance between sensitivity to errors and robustness to outliers [41]. In many cases, a combination of these metrics can provide a more comprehensive assessment of a model's strengths and weaknesses, contributing to a more thorough and nuanced interpretation of results.

3 Results and Discussion

3.1 Statistical analysis of datasets

Table.1: depicts the entirety of the generated dataset. Within this dataset, the variables exhibited mesopore volume size ranging from 0.3 to 0.899 cm3/g and surface area from 800.09 to 1999.70 m2/g. To assess the degree of dispersion for each

variable, we employed mean values and their corresponding standard deviations(SD). Additionally, the range of each variable was captured through the utilization of minimum and maximum values. Complementing these measures, quartile ranges were furnished to provide supplementary insights into the distribution's central tendency. "SD" typically stands for "standard deviation," which measures the amount of variation or dispersion in a set of values. So, "SD mean" could refers to the standard deviation of the mean, indicating how much the average value varies from the overall average.

3.2 Machine Learning Prediction

3.2.1 Model Evaluation

Nine (9) machine learning models were employed based on the testing dataset. Table.2: displays the error values for various models, including support vector re gression, linear regression, ridge regression, decision tree regression, and random forest regression, together with surface area, mesopore volume and total pore volume. Table.2: Error values for the three outputs using different machine learning models.

3.2.2 Comparative Analysis with Chemical Methods

We compared ML predictions (e.g. using Random Forest, Gradient Boosting, etc.) with traditional methods. ML models outperform traditional methods due to their ability to capture intricate patterns. ML models benefit from larger datasets, whereas traditional methods can work with smaller samples. ML implementation involved computational costs, but traditional methods required expensive equipment. ML models handled noisy data better, but traditional methods were robust even with limited data.

3.2.3 Hyperparameter tuning

Hyperparameter tuning is the process of selecting the best combination of hyperparameters for a machine learning model which optimizes its performance in a given task.

Hyperparameters are parameters that are not learned by the model during training, but rather set before training begins, such as learning rate, regularization parameter, number of hidden layers, etc. The selection of appropriate hyperparameters can greatly impacts the performance of a machine learning model. For example, using a high learning rate may cause the model to converge quickly but also result in unstable results or overshooting the optimal solution, while a low learning rate may cause the model to converge slowly but also result in more stable and accurate results. Similarly, selecting an appropriate value for regularization can help to prevent overfitting, while choosing the optimal number of hidden layers or neurons can impact the model's ability to learn complex patterns in the data.

Hyperparameter tuning involves searching through a range of hyperparameters using various techniques such as grid search, random search and Bayesian optimization to find the combination that results in the performance on a validation set. By optimizing the hyperparameters, the model's accuracy and generalization ability can be improved [42].

Hyperparameter tuning in random forest is the process of finding the optimal values for the hyperparameters of the model which maximize its performance on a given dataset. Hyperparameters are values set before training the model that control its behavior, such as the number of trees in the forest, maximum depth of each tree, and minimum number of samples required to split a node.

Table 1: Statistical Analysis of **MBN, IN, SA, MPV** and **TPV**. Note: **MBN**: Methylene Blue Number, **IN:** Iodine Number, **SA**: Surface Area, **MPV:** Mesopore volume, **TPV**: Total Pore Volume, **SD:** Standard deviation, **Min:** Minimum value, **Max:** Maximum value

Item	MBN (mg/g)	IN (mg/g)	SA(m^2/g)	MPV (cm^3/g)	TPV (cm^3/g)
Count	11001	11001	11001	11001	11001
Mean	250.97	748.21	1399.32	0.600	1.150
SD	86.28	143.96	345.04	0.172	0.203
Min.	100.00	500.06	800.09	0.300	0.800
25%	176.40	625.57	1100.07	0.451	0.975
50%	252.19	746.86	1404.11	0.599	1.150
75%	325.67	872.67	1693.13	0.751	0.975
Max.	399.93	999.93	1999.70	0.899	1.499

Table 2: Error values for the three outputs using different machine learning models.

S.N.	Model	MAE	MSE	RMSE	R^2	Adj.R^2
0	LR	47.109881	3459.895945	58.820880	0.973111	0.973074
1	LR	0.046539	0.003337	0.057767	0.902342	0.902208
2	LR	0.055253	0.004676	0.068379	0.902178	0.902044
3	Ridge	47.112481	3460.045858	58.822154	0.973110	0.973073
4	Ridge	0.046538	0.003337	0.057766	0.902343	0.902210
5	Ridge	0.055254	0.004676	0.068378	0.902182	0.902048
6	Lasso	47.146393	3461.876753	58.837715	0.973096	0.973059
7	Lasso	0.157934	0.034177	0.184871	-0.000206	-0.001572
8	Lasso	0.186585	0.047804	0.218642	-0.000131	-.001498
9	DTR	68.094400	7183.560785	84.755889	0.944172	0.944096
10	DTR	0.039345	0.002447	0.049466	0.928393	0.928295
11	DTR	0.047635	0.003575	0.059798	0.925204	0.925102
12	RFR	51.035035	4071.936105	63.811724	0.968354	0.968311
13	RFR	0.029206	0.001357	0.036833	0.960296	0.960242
14	RFR	0.033933	0.001826	0.042730	0.961801	0.961749
15	SVR	47.745394	5438.636734	73.747113	0.957733	0.957675
16	SVR	0.028200	0.001268	0.035602	0.962906	0.962855
17	SVR	0.032529	0.001679	0.040970	0.964883	0.964835
18	GBR	47.685568	3549.730217	59.579612	0.972413	0.972375
19	GBR	0.027841	0.001228	0.035046	0.964057	0.964007
20	GBR	0.032529	0.001653	0.040657	0.965418	0.965371
21	MLPR	51.592602	4156.769301	64.473012	0.977695	0.967651
22	MLPR	0.027841	0.001240	0.035212	0.963715	0.963665
23	MLPR	0.033108	0.001719	0.041458	0.964040	0.963991
24	KNR	52.935184	4374.378187	66.139082	0.966004	0.965957
25	KNR	0.030165	0.001431	0.037834	0.958109	0.958052
26	KNR	0.034718	0.001930	0.043932	0.959621	0.959566
27	CBR	48.094239	3602.027347	60.016892	0.972006	0.971968
28	CBR	0.027934	0.001252	0.035384	0.963359	0.963309
29	CBR	0.032460	0.001658	0.040722	0.965307	0.965259
30	XGBR	49.447313	3847.236973	62.026099	0.970101	0.970060
31	XGBR	0.029149	0.001350	0.036742	0.960494	0.960438
32	XGBR	0.033727	0.001813	0.042582	0.962065	0.962014

Note: LR: Linear Regression, **DTR:** Decision Tree Regression, **RFR**: Random Forest Regression, **SVR:** Support Vector Regression **GBR:** Gradient Boosting Regressor**, MLPR**: MLP Regressor, **KNR**: KNeighbors Regressor, **CBR**: CatBoost Regressor, **XGBR**: XGB Regressor

Table 3: Metrics for evaluation of Random Forest model after hyperparameter tuning

S.N.	Model	MAE	MSE	RMSE	R^2	Adj. R^2
0	RFR	47.984737	3578.597581	59.821381	0.972188	0.972150
1	RFR	0.028127	0.001263	0.035535	0.963045	0.962994
2	RFR	0.032706	0.001698	0.041210	0.964470	0.964421

Table 4: Metrics for evaluation of XG boost regression model after hyperparameter tuning

S.N.	Model	MAE	MSE	RMSE	R^2	Adj-R^2
0	XGBR	47.889026	3580.548067	59.837681	0.972173	0.972135
1	XGBR	0.023751	0.001221	0.034947	0.964259	0.964210
2	XGBR	0.032215	0.001640	0.040494	0.965694	0.965647

Table 5: Best performance metrics and performance models

S.N.	Model	MAE	MSE	RMSE	R^2	Adj.R^2	Best parameters
0	LR	47.109881	3459.8945	58.82088	0.973111	0.973104	{}
1	RR	47. 109881	3459.895958	58.82088	0.973111	0.973104	{`alpha`:1}
2	Lasso	47.11211	3460.137574	58.822934	0.973109	0.973102	{`alpha`:1}
3	DTR	48.808878	3694.968815	60.786255	0.971284	0.971276	{`max-depth`:5,`min-sample-split`
4	RFR	47.923388	3567.013048	59.724476	0.972278	0.972271	{'max-depth':3,'n-estimators':50}
5	SVR	47.125706	3460.319712	58.024402	0.972413	0.9731	{'c':1, 'kernel': 'linear'}
6	GBR	47.773029	35499.733574	59.579641	0.972413	0.972405	{'max-depth':3,'n-estimators':50}
7	MLPR	47. 803276	3572.308247	59.76879	0.972237	0.97223	{'alpha':0.1,'hidden-layer-sizes': (100,)}
8	KNNR	52.502503	4354.524219	65.988819	0.966150	0.966149	{'n-neighbors':5}
9	CBR	47.905911	3585.895292	59.882345	0.972132	0.972124	{'depth':3,'iterations':200}
10	XGBR	48.372938	3656.927191	60.472533	0.97158	0.971572	{'max-depth':3,n'-estimators':100}
11	Elastic Net	47.1101	3459.913411	58.821029	0.973111	0.973103	{`alpha`:0.1
12	PLSR	47. 109881	3459.895945	58.82088	0.973111	0.973104	{`n-components`:3,`scale:`True}

Note: LR: Linear Regression, **DTR:** Decision Tree Regression, **RFR**: Random Forest Regression, **SVR:** Support Vector Regression **GBR:** Gradient Boosting Regressor, **MLPR**: MLP Regressor, **KNR**: KNeighbors Regressor, **CBR**: CatBoost Regressor, **XGBR**: XGB Regressor

Hyperparameter tuning in random forest is the process of finding the optimal values for the hyperparameters of the model that maximize its performance on a given dataset. Hyperparameters are values set before training the model that control its behavior, such as the number of trees in the forest, the maximum depth of each tree, and the minimum number of samples required to split a node.

For each of the model we have calculated the evaluation metrics like MAE, MSE etc. and also the best parameters so that we can get the best pathway for each model.

4 Conclusion

Fossil fuels contribute to air pollution, greenhouse gas emissions, and environmental degradation. Supercapacitors can help to reduce these effects by providing efficient energy storage and recovery, especially in applications like electric vehicles and renewable energy systems. They can capture and release energy quickly, minimizing the need for constant fossil fuel consumption. Using supercapacitors can lead to cleaner air and reduced carbon emissions, promoting a more sustainable energy

landscape. A model that predicts the surface area, mesopore volume and total pore volume of bio-based activated carbon was created using various machine learning algorithms. The machine uses the input features namely temperature, methylene blue number and iodine number, and the test circumstances. The prediction performance of the Random Forest Regressor and cat Boost Regressor models was superior than that of the ANN, Ridge, MLP, GBR, DTR and SVR models. The multi-faceted correlations between input parameters such iodine amount and Methylene blue number and output variables include surface area, mesopore volume, and total pore volume have been studied using machine learning (ML). In addition, ML might be used for generating models for prediction that includes surface area, mesopore volume, and total pore volume. By using surface area, mesopore volume, and total pore volume, machine learning algorithms can assist researchers in choosing the best model and in enhancing the efficacy and reliability of the energy storage process.

Machine Learning (ML) is used to analyze the effects of the input parameters on the basis of value of surface area, mesopore volume, and total pore volume, as well as to optimize the required output parameters. We located Random Forest Regressor and XG Boost Regressor models, among nine different ML methods, are the most suitable models. Since, produced activated carbon possesses substantial values for surface area, mesopore volume, and total pore volume, it can be employed as an electrode material for energy storage in future.

Acknowledgement

The authors would like to express their gratitude to the Department of Applied Sciences and Chemical Engineering (IOE), Pulchowk Campus.

References

[1] M. Rahimi, M.H. Abbaspour-Fard, and A. Rohani. Synergetic effect of n/o functional groups and microstructures of activated carbon on supercapacitor performance by machine learning. *Journal of Power Sources*, 521:230968, 2022.

[2] C. Wang, W. Jiang, G. Jiang, T. Zhang, K. He, L. Mu, J. Zhu, D. Huang, H. Qian, and X. Lu. Machine learning prediction of the yield and bet area of activated carbon quantitatively relating to biomass compositions and operating conditions. *Industrial & Engineering Chemistry Research*, 62:11016–11031, 2023.

[3] A.G. Saad, A. Emad-Eldeen, W.Z. Tawfik, and A.G. El-Deen. Data-driven machine learning approach for predicting the capacitance of graphene-based supercapacitor electrodes. *Journal of Energy Storage*, 55:105411, 2022.

[4] S. Jha, S. Bandyopadhyay, S. Mehta, M. Yen, T. Chagouri, E. Palmer, and H. Liang. Data-driven predictive electrochemical behavior of lignin-based supercapacitors via machine learning. *Energy & Fuels*, 36:1052–1062, 2021.

[5] H. Su, S. Lin, S. Deng, C. Lian, Y. Shang, and H. Liu. Predicting the capacitance of carbon-based electric double layer capacitors by machine learning. *Nanoscale Advances*, 1:2162–2166, 2019.

[6] J. Abdi, T. Pirhoushyaran, F. Hadavimoghaddam, S.A. Madani, A. Hemmati-Sarapardeh, and S.H. Esmaeili-Faraj. Modeling of capacitance for carbon-based supercapacitors using super learner algorithm. *Journal of Energy Storage*, 66:107376, 2023.

[7] C.L. Gnawali, S. Shahi, S. Manandhar, G.K. Shrestha, M.P. Adhikari, R. Rajbhandari, and B.P. Pokharel. Porous activated carbon materials from triphala seed stones for high-performance supercapacitor applications. *BIBECHANA*, 20(1):10–20, 2023.

[8] R. Dubey and V. Guruviah. Machine learning enabled performance prediction of biomass-derived electrodes for asymmetric supercapacitor. In *International Conference on Futuristic Communication and Network Technologies*, volume 995, pages 453–460, 2021.

[9] L. Leng, L. Yang, X. Lei, W. Zhang, Z. Ai, Z. Yang, H. Zhan, J. Yang, X. Yuan, and H. Peng. Machine learning predicting and engineering the yield, n content, and specific surface area of biochar derived from pyrolysis of biomass. *Biochar*, 4:63, 2022.

[10] C.L. Gnawali, L.K. Shrestha, J.P. Hill, R. Ma, K. Ariga, M.P. Adhikari, R. Rajbhandari, and B.P. Pokharel. Nanoporous activated carbon material from terminalia chebula seed for supercapacitor application. *Journal of Carbon Research*, 9:109, 2023.

[11] H. Thanh, S.E. Taremsari, B. Ranjbar, H. Mashhadi-Moslem, E. Rahimi, M. Rahimi, and A. Elkamel. Hydrogen storage on porous carbon adsorbents: Rediscovery by nature-derived algorithms in random forest machine learning model. *Energies*, 16:2348, 2023.

[12] S. Zhu, J. Li, L. Ma, C. He, E. Liu, F. He, C. Shi, and N. Zhao. Artificial neural network enabled capacitance prediction for carbon based supercapacitors. *Materials Letters*, 233:294–297, 2018.

[13] J. Wang, Z. Li, S. Yan, X. Yu, Y. Ma, and L. Ma. Modifying the microstructure of algae-based active carbon and modelling supercapacitors using artificial neural networks. *RSC Advances*, 9:14797–14808, 2019.

[14] X. Yang, C. Yuan, S. He, D. Jiang, B. Cao, and S. Wang. Machine learning prediction of specific capacitance in biomass derived carbon materials: Effects of activation and biochar characteristics. *Fuel*, 331:125718, 2023.

[15] Z. Yang, Y. Lin, X. Gu, and X. Liang. Prediction and optimization model of activated carbon double layer capacitors based on improved heuristic approach genetic algorithm neural network. *Engineering Computations*, 35:1625–1638, 2018.

[16] M. Rahimi, M.H. Abbaspour-Fard, and A. Rohani. Machine learning approaches to rediscovery and optimization of hydrogen storage on porous bio-derived carbon. *Journal of Cleaner Production*, 329:129714, 2021.

[17] K. Donthula, N. Thota, S.B. Anne, and M. Kakunuri. Prediction of capacitance using artificial neural networks for carbon nanofiber-based supercapacitors. In *Computer Aided Chemical Engineering*, volume 52, pages 1015–1020, 2023.

[18] A. Albalasmeh, M.A. Gharaibeh, O. Mohawesh, M. Alajlouni, M. Quzaih, M. Masad, and A. El Hanandeh. Characterization and artificial neural networks modelling of methylene blue adsorption of biochar derived from agricultural residues: Effect of biomass type, pyrolysis temperature, particle size. *Journal of Saudi Chemical Society*, 24:811–823, 2020.

[19] S. Joshi, R.G. Shrestha, R.R. Pradhananga, K. Ariga, and L.K. Shrestha. High surface area nanoporous activated carbons materials from areca catechu nut with excellent iodine and methylene blue adsorption. *C*, 8:2, 2021.

[20] S. Joshi and B.P. Pokharel. Preparation and characterization of activated carbon from lapsi (choerospondias axillaris) seed stone by chemical activation with potassium hydroxide. *Journal of the Institute of Engineering*, 9:79–88, 2013.

[21] C. L. Gnawali, S. Manandhar, S. Shahi, R. G. Shrestha, M. P. Adhikari, R. Rajbhandari, B. P. Pokharel, R. Ma, K. Ariga, and L. K. Shrestha. Nanoporous carbons materials from terminalia bellirica seed for iodine and methylene blue adsorption and high-performance supercapacitor applications. *Bulletin of the Chemical Society of Japan*, 96:572–581, 2023.

[22] L. K. Shrestha, S. Shahi, C. L. Gnawali, M. P. Adhikari, R. Rajbhandari, B. P. Pokharel, R. Ma, R. G. Shrestha, and K. Ariga. Phyllanthus emblica seed-derived hierarchically porous carbon materials for high-performance supercapacitor applications. *Materials*, 15:8335, 2022.

[23] S. Bhandari, K. B. Rajguru, C. L. Gnawali, and B. P. Pokharel. Influence of precursor type on activated carbon prepared by phosphoric acid-chemical activation for supercapacitor applications. *Journal of Nepal Physical Society*, 9(3):12–17, 2023.

[24] S. Mishra, R. Srivastava, A. Muhammad, A. Amit, E. Chiavazzo, M. Fasano, and P. Asinari. The impact of physicochemical features of carbon electrodes on the capacitive performance of supercapacitors: a machine learning approach. *Scientific Reports*, 13:6494, 2023.

[25] K. Zhang, S. Zhong, and H. Zhang. Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. *Environmental Science & Technology*, 54:7008–7018, 2020.

[26] M. Khan, Z. Ullah, O. Mašek, S. R. Naqvi, and M. N. A. Khan. Artificial neural networks for the prediction of biochar yield: a comparative study of metaheuristic algorithms. *BioSource Technology*, 355:127215, 2022.

[27] A. Hai, G. Bharath, M. F. A. Patah, W. M. A. W. Daud, K. Rambu, P. Show, and F. Banat. Machine learning models for the prediction of total yield and specific surface area of bio-char derived from agricultural biomass by pyrolysis. *Environmental Technology & Innovation*, 30:103071, 2023.

[28] Z. U. Haq, H. Ullah, M. N. A. Khan, S. R. Naqvi, A. Ahad, and N. A. S. Amin. Comparative study of machine learning methods integrated with genetic algorithm and particle swarm optimization for bio-char yield prediction. *BioResource Technology*, 363:128008, 2022.

[29] J. A. Okolie, S. Savage, C. C. Ogbaga, and B. Gunes. Assessing the potential of machine learning methods to study the removal of pharmaceuticals from wastewater using biochar or activated carbon. *Total Environment Research Themes*, 1:100001, 2022.

[30] S. Nanda, S. Ghosh, and T. Thomas. Machine learning aided cyclic stability prediction for supercapacitors. *Journal of Power Sources*, 546:231975, 2022.

[31] Y. W. Liew, S. K. Arumugasamy, and A. Selvarajoo. Potential of biochar as soil amendment: prediction of elemental ratios from pyrolysis of agriculture biomass using artificial neural network. *Water, Air, & Soil Pollution*, 233:54, 2022.

[32] A. Shafizadeh, H. Shahbeik, S. Rafiee, A. Moradi, M. Shahbaz, M. Madadi, C. Li, W. Peng, M. Tabatabaei, and M. Aghbashlo. Machine learning-based characterization of hydrochar from biomass: Implications for sustainable energy and material production. *Fuel*, 347:128467, 2023.

[33] J. Wang, X. Zhang, Z. Li, Y. Ma, and L. Ma. Recent progress of biomass-derived carbon materials for supercapacitors. *Journal of Power Sources*, 451:227794, 2020.

[34] K. Rahmani, A. H. Mamaghani, Z. Hashisho, D. Crompton, and J. E. Anderson. Prediction of heel build-up on activated carbon using machine learning. *Journal of Hazardous Materials*, 433:128747, 2022.

[35] W. Jiang, X. Xing, S. Li, X. Zhang, and W. Wang. Synthesis, characterization and machine learning based performance prediction of straw activated carbon. *Journal of Cleaner Production*, 212:1210–1223, 2019.

[36] S. K. Arumugasamy, A. Selvarajoo, and M. A. Tariq. Artificial neural networks modelling: gasification behaviour of palm fibre biochar. *Materials Science for Energy Technologies*, 3:868–878, 2020.

[37] L. Wu, Z. Xu, Z. Wang, Z. Chen, Z. Huang, C. Peng, X. Pei, X. Li, J. P. Mailoa, C. Y. Hsieh, and T. Wu. Machine learning accelerated carbon neutrality research using big data from predictive models to interatomic potentials. *Science China Technological Sciences*, 65(10):2274–2296, 2022.

[38] D. Tena-Gag, I. Golcarenarenji, I. Martinez-Alpiste, Q. Wang, and J. M. Alcaraz-Calero. Machine-learning-based carbon dioxide concentration prediction for hybrid vehicles. *Sensors*, 23(3):1350, 2023.

[39] M. Mądziel, A. Jaworski, H. Kuszewski, P. Woś, T. Campisi, and K. Lew. The development of co2 instantaneous emission model of full hybrid vehicle with the use of machine learning techniques. *Energies*, 15:142, 2021.

[40] J. Seo, B. Yun, J. Park, J. Park, M. Shin, and S. Park. Prediction of instantaneous real-world emissions from diesel light-duty vehicles based on an integrated artificial neural network and vehicle dynamics model. *Science of the Total Environment*, 786:147359, 2021.

[41] S. Ghosh, G. R. Rao, and T. Thomas. Machine learning-based prediction of supercapacitor performance for a novel electrode material: Cerium oxynitride. *Energy Storage Materials*, 40:426–438, 2021.

[42] Y. Y. Chia, L. H. Lee, S. Shafiabady, and D. Isa. A load predictive energy management system for supercapacitor-battery hybrid energy storage system in solar application using the support vector machine. *Applied Energy*, 137:588–602, 2015.