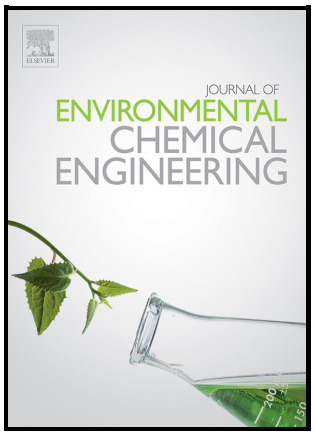

Machine learning-aided prediction of COD removal in the electrocoagulation process using a super learner model

Mhd Taisir Albaba, Mohammed Talhami, Abdullah Omar, Sumith Varghese, Rayane Akoumeh, Mohamed Arselene Ayari, Probir Das, Ali Altaee, Maryam AL-Ejji, Alaa H. Hawari

PII: S2213-3437(25)02165-7

DOI: https://doi.org/10.1016/j.jece.2025.117469

Reference: JECE117469

To appear in: *Journal of Environmental Chemical Engineering*

Received date: 13 April 2025
Revised date: 23 May 2025
Accepted date: 5 June 2025

Please cite this article as: Mhd Taisir Albaba, Mohammed Talhami, Abdullah Omar, Sumith Varghese, Rayane Akoumeh, Mohamed Arselene Ayari, Probir Das, Ali Altaee, Maryam AL-Ejji and Alaa H. Hawari, Machine learning-aided prediction of COD removal in the electrocoagulation process using a super learner model, *Journal of Environmental Chemical Engineering,* (2025) doi:https://doi.org/10.1016/j.jece.2025.117469

Machine learning-aided prediction of COD removal in the electrocoagulation process using a super learner model

Mhd Taisir Albaba [a], Mohammed Talhami [a], Abdullah Omar [b], Sumith Varghese [b], Rayane Akoumeh [c], Mohamed Arselene Ayari [a], Probir Das [d], Ali Altaee [e], Maryam AL-Ejji [c], Alaa H. Hawari [a*]

[a] Department of Civil and Environmental Engineering, College of Engineering, Qatar University, PO Box 2713, Doha, Qatar

[b] Department of Chemical Engineering, College of Engineering, Qatar University, Doha, Qatar

[c] Center of Advanced Materials, Qatar University, PO Box 2713, Doha, Qatar

[d] Algal Technologies Program, Center for Sustainable Development, College of Arts and Sciences, Qatar University, Doha 2713, Qatar

[e] School of Civil and Environmental Engineering, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia

[*]Corresponding author. a.hawari@qu.edu.qa

Abstract

A new predictive machine learning stacking model was developed to examine chemical oxygen demand (COD) removal efficiency in electrocoagulation. The model used a comprehensive dataset consisting of 379 points containing no missing data collected from different studies investigating COD removal efficiency using electrocoagulation, encompassing different wastewater types. The newly developed model included 10 input parameters, namely initial COD concentration, pH, conductivity, anode material, cathode material, inter-electrode distance, number of electrodes, current density, the ratio between the effective electrode area and the reactor volume, and

electrolysis time. The stacking model uses three ensemble models, specifically, gradient boosting regression (GBR), eXtreme Gradient Boosting (XGB), and random forest regression (RFR), as the base learners, while the meta learner is a linear regression model. The developed model has a prediction accuracy of 95.3% for the R^2 value in the test dataset. Additionally, the study used sensitivity analysis and Partial Dependence Plots (PDPs) to determine the impact of each input parameter on COD removal efficiency. The results show that the three most influential parameters are electrolysis time, inter-electrode distance, and current density.

Keywords:

Electrocoagulation; COD removal; Machine Learning; Water treatment; Super-learner model; chemical oxygen demand.

1. Introduction

Electrocoagulation (EC) is a technique that forms coagulants of suspended, dissolved, and emulsified pollutants through the application of an electric charge to the wastewater [1]. As early as 2006, studies have implemented machine learning models to produce predictive models for contaminant removal in EC [2]. Recent advances in machine learning for EC process parameter prediction have examined operational current as the output for a modified ANN model [3]. The ANN model utilized 367 datapoints collected from a water treatment plant with input parameters such as pH, conductivity, effluent and influent flow, effluent and influent turbidity, and temperature. Model interpretation was accomplished using sensitivity analysis, feature engineering, and scenario analysis, with effluent turbidity being the most impactful parameter on operational current in EC. ANN modelling was developed to investigate dye removal in a synthetic

mixture of Golden Yellow X-GL 200% dye and deionized wastewater [2]. Seven input parameters were considered and collected from performing 49 EC experiments on the synthetic mixture, namely current density, time, pH, initial dye concentration, conductivity, sludge retention time, and inter-electrode distance. The ANN model was used to find optimum values of the input parameters for efficient dye removal. ANNs have also been applied to investigate the removal of other types of dye in EC [4]. The model dataset was gathered from 25 EC experiments and used electric current, NaCl concentration, pH, and time. The mean square error of the model was relatively low in the loss function for the validation phase compared to the training phase.

Other studies investigated the removal of specific contaminants like arsenic using gradient boosting machines as their machine learning model [5]. A gradient boosting model used data collected from performing 44 experiments on groundwater samples collected from draw wells. The model used 10 input parameters, namely pH, electrode material, conductivity, electrode treatment, voltage, current density, coagulant dosage, number of electrodes, time, and inter-electrode distance. Through the use of a feature importance analysis, time was observed to be the most impactful parameter on arsenic removal in the model. One study investigated Cr(VI) removal in EC using an Artificial Neural Network (ANN) model based on 212 experimental datapoints [6]. The ANN model used current density, time, Cr(VI) initial concentration, and NaCl concentration. An optimum pH range for Cr(VI) removal of 5-8 was determined. ANN was also combined with a Sugeno-type Fuzzy Inference System (FIS) to examine phosphate removal in EC [7]. Sixty-two datapoints were collected from the literature to construct the hybrid model. The dataset contained 5 input parameters, pH, current, initial phosphate concentration, electrode type, and time. The resulting model was most impacted by electrode type, followed by initial phosphate concentration.

Other studies have focused on pollutant indicators like chemical oxygen demand (COD) removal efficiency [8]. Fifteen datapoints collected from EC experiments on a synthetic oil emulsion were used to build an ANN model based on 6 parameters, specifically pH, current density, inter-electrode distance, oil concentration, electrolyte concentration, and time. The ANN model results were compared with a polynomial model and were found to be more accurate at COD removal prediction. Similarly, COD removal from dairy wastewater was examined using an ANN model [9]. The results of 275 dairy effluent treatment trials by EC were utilized to construct the ANN model with nine input parameters, total solids, total dissolved solids, total suspended solids, turbidity, initial COD concentration, initial pH, time, current density and inter-electrode distance. The most impactful parameter on the model was found to be total dissolved solids, followed by initial COD concentration. Other studies have utilized ANN models to examine multiple output parameters including COD [10], [11]. For example, using a set of 20 experiments, an ANN model was constructed based on three input parameter, namely current density, pH, and time, along with four output parameters, COD, BOD, nitrate, and phosphate removal [10]. The resulting ANN model was accompanied by a Response Surface Methodology (RSM) model to interpret the impact of the input parameters on the output parameters. On the other hand, ANN was also applied to examine total dissolved solids (TDS), biochemical oxygen demand (BOD), COD, and Chromium removal in EC via examining 4 input parameters [11]. The input parameters used in constructing the model, namely pH, inter-electrode distance, voltage, and time, were collected from 80 experimental trials per electrode shape and material. The resulting ANN model had a 95% confidence level.

From examining the studies involved in developing predictive machine learning models for EC, it is clear that ANNs are the most common machine learning algorithm used. Despite machine

learning encompassing a myriad of diverse modelling techniques. Another observation is that most studies either use a small dataset to train the models [4], [8], [10], or use a small number of input parameters [6], [10], [11]. These limitations can hinder the models' ability to cover large ranges of data and their ability to encompass the complexity of the electrocoagulation as a system. While some studies have a large dataset to construct predictive models, they tend to focus on only a single wastewater type, as each study either performs the experiments themselves, or they gather the data from a single study. In addition, the input parameters themselves are limited. In that variables such as the ratio of electrode surface area to reactor volume (A/V) is not considered by any study examined in the literature to develop a predictive model. Also, electrode material is always assumed to be the same for cathode and anode. This is despite numerous studies using different materials for both [12], [13], [14].

In this paper, a new predictive machine learning model was developed using the stacking method, with three base models and one meta learner for COD removal efficiency in EC. The model uses ten variables as input parameters, namely electrolysis time, A/V ratio, current density, inter-electrode distance, conductivity, pH, initial COD concentration, number of electrodes, as well as cathode and anode materials. The ten input parameters will be used on Gradient Boosting Regression (GBR), eXtreme Gradient Boosting (XGB), and Random Forest Regression (RFR) as base models to then be used to train a linear regression model as the meta learner. The dataset used to develop the stacking model was collected from multiple studies on EC and features 379 experimental points. The goal of this study is to combine a comprehensive set of input parameters with a large dataset based on real experiments to produce an accurate predictive machine learning model based on the principles of model stacking. In addition, a sensitivity analysis, along with

Partial Dependence Plots, were used to interpret the model results. Where the aim is to examine the impact of each input parameter on COD removal efficiency.

This study will follow a clear outline. First, methodology of the work will be explained, including data collection, data processing, model overview, hyperparameter selection, and statistical evaluation. Second, the results and discussion section will include the model results, as well as the model interpretation, specifically PDPs and the sensitivity analysis, followed by discussing the prediction perspectives of the model. Finally, concluding remarks will be presented in the conclusion section.

2. Methodology

A total of 379 datapoints were collected from multiple studies in order to construct a predictive ML model for COD removal in EC. The datapoints included 8 numerical and two categorical input parameters, and a singular output parameter, which is COD removal efficiency. The input parameters can be further categorized as electrode configuration, wastewater characteristics, and operating conditions. Where the electrode configuration included the number of electrodes as a numerical parameter, and anode material and cathode material as categorical input parameters. While the wastewater characteristics were represented by the COD initial concentration, pH, and conductivity. The operating conditions were represented by the electrode surface area to reactor volume (A/V) ratio, inter-electrode distance, current density, and electrolysis time. The ML model used for this paper was dependent on stacking, which is a system that utilizes multiple ML models to construct a model based on the trained models [15]. This type of model is commonly referred to as a super learner [16], [17]. Ensemble models were selected for the models used as training, specifically RFR, GBR, and XGB, as they tend to perform better than single models [18], [19]. The chosen split for training and testing of the models was 80/20. Seeing

as the total size of the data was 397 points, 20% was decided as sufficient for testing the models. Hyperparameter optimization was performed on all ensemble models as well as the super learner. Following that, model performance was examined using four indications, the root mean square error (RMSE), the coefficient of determination (R^2), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). Model interpretation was then performed using PDP and sensitivity analysis.

2.1. Data collection

A total of 379 data points were collected from the studies listed in Table 1. The compiled dataset by the authors contained no missing data and covered extensive ranges for the input parameters and the output parameter. The dataset is made available in the supplementary materials for better access to the data and its distribution, as it is a controlling factor in developing machine learning models. Initial COD concentrations vary from 260 to 122000 mg/L. Conductivity ranges from 9.87 to 25000 µS/cm. While pH ranges from 3 to 11. For the electrode materials, they were split into anode and cathode materials. Where the anode material consisted of either Aluminium (Al), Copper (Cu), or Iron (Fe). While the cathode materials were either Al, Fe, or Stainless Steel (SS). The number of electrodes used in EC ranged from 2 to 7 electrodes. Although the majority used 2 electrodes. Additionally, the distance between each electrode ranged from 0.25 to 10 cm. Current density data ranged from 0.43 to 50 mA/m^2. An additional factor was calculated using effective electrode surface area and reactor volume from the data due to its impact on the EC process [20]. The factor is specifically the ratio of electrode effective surface area and reactor volume (m^2/m^3). Whereas A/V ratio varied from as low as 1.4 to as high as 26.67 m^2/m^3. The electrolysis time for the EC process in the dataset was between 2 to 240 minutes. Figure 1 shows the Pearson correlation heatmap between the input parameters and the COD removal efficiency.

The light colours of the bottom row indicate insignificant linear correlation between the input parameters and COD removal efficiency. This further proves that machine learning techniques are required to represent such a complex process.

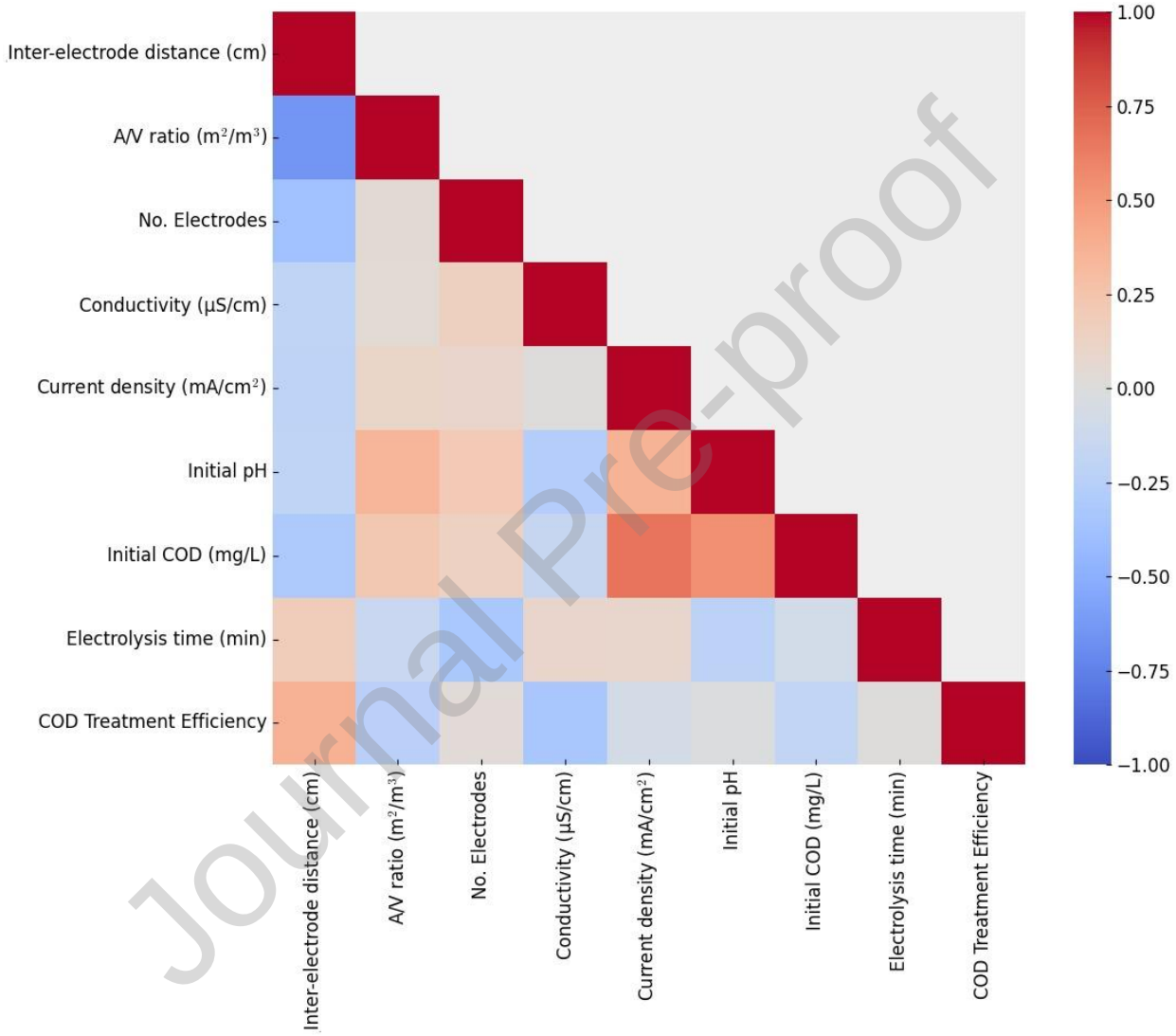

Figure 1. Heatmap between the input parameters and the output variable showing Pearson correlation coefficients.

Table 1. Data ranges for the studies used to compile the data for model development.

Initial COD concentration (mg/L)	Conductivity (μS/cm)	pH	Anode material	Cathode material	Number of electrodes	Inter-electrode distance (cm)	Current density (mA/cm^2)	A/V ratio (m^2/m^3)	Electrolysis time (min)	Reference
10000	17220	6.5	Al	Al	4	3.00	6-12	4.48	15-60	[21]
420-488	1514	3.0-9.0	Fe	Fe	4	0.25	0.43-2.50	7.50	30	[22]
11150	3980	3.0-11.0	Al	Al	2	1.00	0.50-10.00	26.67	5-60	[23]
6114.25	8074	7.24	Al	Al	6	2.00	2.98-17.86	14.00	15-60	[24]
27000-41000	680-950	5.6-7.2	Al	Al	7	4.00	20	1.4	2-10	[25]
8870	25800	7.5	Al-Fe	Al-Fe	2	1.00	12-36	10	25-240	[12]
260	2400	7	Al	Al	6	1.00	0.80-8.00	12.56	4-24	[26]
23520	5900	4.3	Al-Fe	Al-Fe	2	2	5.00-20.00	8.00-9.60	15-180	[27]
122000	11.00	10.50	Al	Al	4	1.00	10.00-50.00	15.00	2-60	[28]
670	254	4.00-10.00	Al-Fe-Cu	Al-Fe-SS	2	2.5-10	0.50-3.00	1.5	10-60	[14]
397-1000	9.87-11.00	4.00-10.00	Al-Fe-Cu	SS	2	4.00	1.33-12.00	7.5	10-60	[13]

2.2. Data processing

Data preprocessing is a crucial step before ML implementation to ensure high-quality and reliable input data. The dataset was processed using Python on Google Collaboratory, which is a popular platform enabling users to write and execute Python code within the web browser environment. The dataset was first cleaned from duplicate entries to achieve consistency and uniqueness. The compiled dataset was fully free of missing data to eliminate inaccuracies arising from the absence of complete information. The utilized dataset contains two categorical variables, namely, anode and cathode materials. The anode material categories were aluminum, iron, and copper, whereas the cathode material consisted of aluminum, iron, and stainless steel. It should be noted that to maintain data reliability and representability, any anode-cathode combination with fewer than 10 instances was excluded from the processed dataset. This categorical nature was

converted to a numerical format which can be dealt with better in machine learning. Since both categorical parameters contain more than two options, the "OneHotEncoder" was applied instead of the "LabelEncoder" function in Python. Unlike the "LabelEncoder" which can handle up to two categories, the "OneHotEncoder" function converts the categorical variable into multiple binary columns, where the number of binary columns depends on the number of categorical options. Binary columns take a value of zero or one to indicate the absence or the presence of a certain category within a parameter, respectively. Considering the total number of cathode and anode materials, six binary columns were created, three for each categorical parameter. The data was then standardized using the "StandardScalar" function from the Scikit learn module in Python, making the mean value in each parameter equal to zero with a standard deviation of one, thus preventing performance issues related to the model's sensitivity to the data scale.

2.3. Model overview

2.3.1. Random Forest Regression (RFR)

Random Forest is an ensemble machine learning technique for regression and classification that combines bagging with decision trees [29]. In bagging, the base learners are trained independently [30], as is the case for Random Forest Regression. In Random Forest Regression, multiple versions of the same model are created in the form of decision trees and are run in parallel to produce independent outputs. The final output is the mean prediction of the parallel models. This is to correct overfitting in individual trees [31]. Though the model can still struggle with overfitting issues [32]. Decision trees in Random Forest Regression work by splitting high and low values of a predictor in relation to an outcome [33]. Randomness is introduced at each tree split. Where splitting selects random input variables in the beginning, with the best variable being

chosen to complete each split. One of the main advantages of this method is its ability to deal with datasets containing a large number of variables [33].

2.3.2. Gradient Boosting Regression (GBR)

GBR is a machine learning technique that sequentially boosts weak learners into strong learners. GBR requires a loss function for optimization, a weak learner for prediction, and an adaptive model for estimation [34]. In each iteration, weak learners are assigned more weight in the sum and adapted at the opposite gradient [35]. The gradient represents the partial derivative of the loss function. And it is utilised model parameter direction to reduce error in the following iteration. In principle, the aim is to minimise the loss function through the addition of consecutive models (decision trees), which are trained using the error residuals from previous models [36]. Thereby making the training process dependent, unlike Random Forest Regression. This means that GBR is more computationally demanding. However, the end result is more accurate [37]. In addition to its accuracy, GBR has the benefit of fast estimation, allowing it to be used in real-time applications [35]. Given the following training example: $\{(x_i, y_i)\}_{i=1}^{Z}$, where $x_i \subseteq R^m$ is the input space with number of features m, and $y_i \in R^m$ is the response variable. The output of the model, the prediction, is the weighted sum of decision trees, and the $F_T(x_i)$ model is expressed in eq (1)[19]:

$$F_T(x_i) = \sum_{t=1}^{T} f_t(x_i) \quad (1)$$

Here, T is the number of decision trees and $f_t(x_i)$ is the set of decision trees. The algorithm for GBR is as follows:

1- Begin the initial training with a constant value of $F_0(x)$ using eq (2):

$$F_0(x) = \underset{c}{argmin} \sum_{i=1}^{n} L(y_i, c) \quad (2)$$

2- Obtain the negative gradient is using eq (3):

$$g_{it} = -\frac{\partial L[y_i, F_{t-1}(x_i)]}{\partial F_t - 1(x_i)} \text{ for I = 1, 2, …, n} \quad (3)$$

3- Fit a new regression tree $f_t(x_i)$ to the negative gradient descent

4- Calculate the multiplier c_t through eq (4):

$$c_t = \underset{c}{argmin} \sum_{i=1}^{n} L\big(y_i, F_{t-1}(x_i) + c f_t(x_i)\big) \quad (4)$$

5- Update the model using eq (5):

$$F_T(x) = F_{t-1}(x) + v c_t f_t(x) \quad (5)$$

6- Repeat steps 2-5 until $t = T$

7- Produce final decision tree $F_T(x)$

Where L is the loss function, and v is the rate of learning.

2.3.3. eXtreme Gradient Boosting (XGB)

XGB is a machine learning algorithm based on combining the gradient boosting framework with the ensemble method for building decision trees [38]. The algorithm uses a loss function and regularization. Where the loss function calculates the difference between actual and predicted values, and regularization controls the complexity of the model [39]. It can be used for both classification and regression. Contrary to GBR, the gradient of XGB is the second partial derivative of the loss function, as it provides additional information regarding the gradient direction [40]. Moreover, XGB parallelizes tree construction, making it faster than GBR. However, because it is still boosting, it shares the concept of iterative learning of weak learners with GBR.

One benefit to using the XGB algorithm is its ability to handle missing data within a set. This decreases the required effort and time in data preparation and collection [36]. XGB's objective function is shown in eq (6) as the sum of two parts. The first part is the loss function, while the second is the regularization parameter **Ω**. Where eq (7) shows the regularization parameter equation.

$$Obj = \sum_{i=1}^{n} L(\hat{y}_i, y_i) + \sum_{t=1}^{T} \Omega(f_t) \quad (6)$$

$$\Omega(f_t) = \Upsilon T + \frac{1}{2}\lambda \|\omega\|^2 \quad (7)$$

In eq (6) and eq (7), T is number of leaves of decision trees, Υ is complexity of the leaves, λ is the penalty term, and $\|\omega\|$ is a vector space which is made up of scores and weights on leaves.

2.3.4. Stacking (super learner) model

Stacking is an ensemble learning technique which combines models in order to produce a stronger learning model [41]. The approach uses the outputs of the base models as inputs to a higher-level model. And that model then provides the final prediction [42]. In the first stage of constructing the stacking model, the base models are trained on the training set, and the outputs of those models are stored in a new dataset as new parameters [42]. The new parameters are used to train the stacking model in the second stage. The model development process is demonstrated in Figure 2. Stacking also has some advantages over other ensemble model techniques like bagging and boosting, specifically, in its ability to combine diverse machine learning models which excel in different aspects to better capture the complex non-linear relationships between the input parameters and the output parameter [43]. Additionally, stacking helps in reducing the variance problems wrought by a limited dataset [43]. For this work, a linear regression model was used as the meta learner.

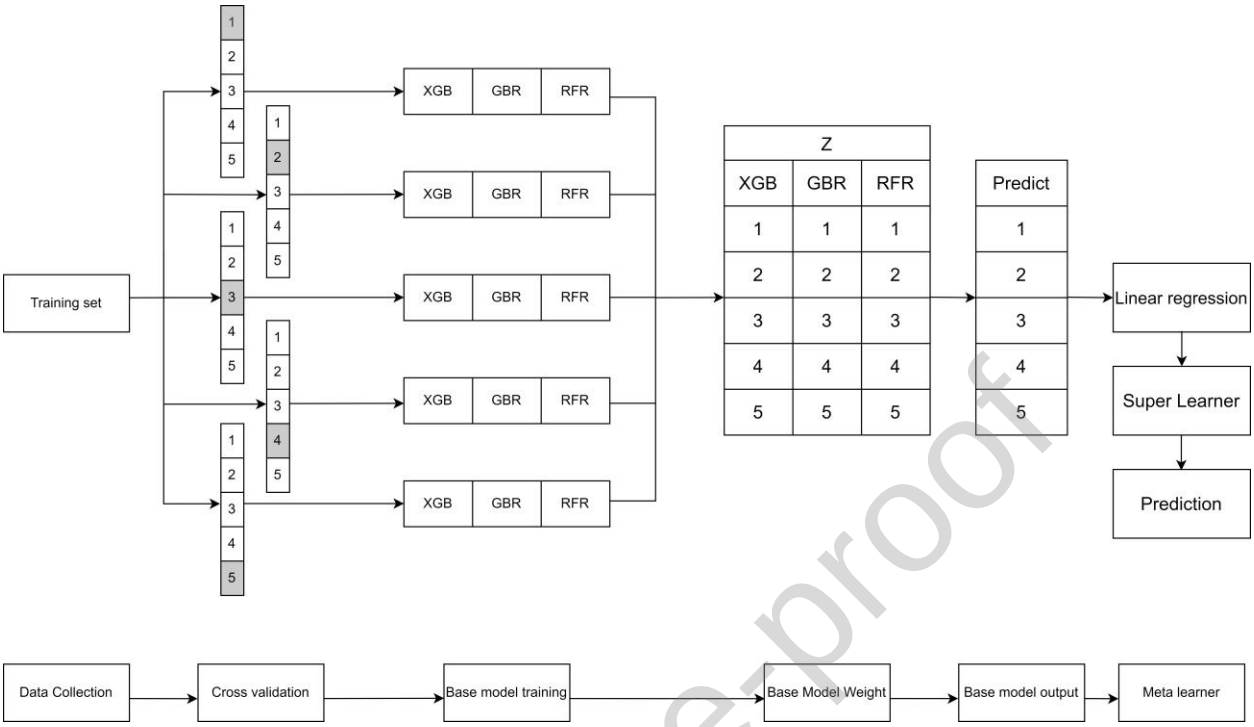

Figure 2. The training process in the stacking (super learner) model. The model employed a 5-fold cross-validation technique to enhance the model generalization. The lower chart represents the general form of the upper chart

2.4. Hyperparameter selection

To promote the performance of the ML models employed in the present study, the hyperparameters were fine-tuned using grid search in conjunction with 5-fold cross-validation. Grid search was responsible for comprehensively exploring the various hyperparameter combinations within a predefined search space for a given model. This thorough investigation of hyperparameter values helps identify the optimal configuration in terms of model accuracy and generalization capacity. On the other hand, overfitting is a very common challenge in machine learning model development [44], [45]. In order to ensure a minimal risk of overfitting, 5-fold cross-validation was utilized as part of the grid search. This technique evaluates the model performance across multiple subsets of the training data. For example, for the 5-fold cross-

validation deployed in this study, the dataset was divided into five equal folds, whereby four folds were used for training the model, and the fifth fold was employed for validation. This process was repeated five times, and the average performance across the validation folds was used for the selection of the optimal hyperparameter values. This adopted approach in hyperparameter optimization ascertains that the selected sets of hyperparameters were not biased towards specific data splits, but rather assist in determining configurations that generalize well on unseen data. It is imperative to note that the hyperparameter optimization was primarily based on minimizing the value of the root mean squared error (RMSE), and the optimal combinations, as well as search ranges for the ML models developed in this study, can be found in Table 2.

Table 2. The optimal hyperparameters for the models.

Model	Hyperparameter	Range of Grid Search (start, finish, step)	Optimal Value
Random Forest Regression (RFR)	Random State	(0, 10, 1)	8
	Number of Estimators	(50, 1000, 50)	100
	Maximum Depth	(1, 20, 1)	15
	Maximum Features	(1, 20, 1)	14
	Minimum Sample Leaf	(1, 5, 1)	1
	Minimum Sample Split	(2, 10, 1)	2
Gradient Boosting Regression (GBR)	Random State	(0, 10, 1)	8
	Number of Estimators	(50, 1000, 50)	600
	Learning Rate	(0.05, 1, 0.05)	0.35
	Maximum Depth	(1, 20, 1)	3
	Maximum Features	(1, 20, 1)	15
	Subsample	(0.05, 1, 0.05)	0.7
	Minimum Sample Leaf	(1, 5, 1)	1
	Minimum Sample Split	(2, 10, 1)	2
Extreme Gradient Boosting (XGB)	Number of Estimators	(50, 1000, 50)	100
	Learning Rate	(0.05, 1, 0.05)	0.85
	Maximum Depth	(1, 10, 1)	6
	Subsample	(0.05, 1, 0.05)	0.85
	Reg Lambda	(0, 1, 1)	1
	Reg Alpha	(0, 1, 1)	0
	Minimum Sample Split	(2, 10, 1)	4
	Gamma	(0, 1, 0.1)	0

2.5. Statistical evaluation

The evaluation of the various models implemented in the present study was performed using three primary statistical error indices, namely, root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), as well as a single correlation index, i.e., the coefficient of determination (R^2). Selecting multiple performance indicators enables achieving an exhaustive evaluation of the ML models' predictive capacity for COD removal in the EC process. RMSE is a popular performance metric that calculates the mean difference between the experimental and predicted values, thus quantifying the model's error and providing a sense of accuracy. MAE captures the magnitude of difference between the experimental and predicted values; however, this metric overlooks the direction and the relative deviation. MAPE, on the other hand, calculates the mean absolute deviation relative to the actual value and presents the error as a percentage, thus making the ML models' prediction capabilities more interpretable in practical scenarios. Finally, R^2 is a commonly employed statistical determinant of how well data is predicted by a certain model by assessing the closeness of the data to the fitted regression line. It is imperative to note that the Pearson correlation coefficient is equal to the coefficient of determination in the context of simple linear regression. However, this is not the case for multiple linear regression or machine learning models with multiple predictors. For people interested in knowing the differences between the coefficient of determination (i.e., used in this study) and the Pearson correlation coefficient, you may refer to [57]. Table 3 outlines the mathematical formulation of these performance metrics as well as the perfect match scenario for each statistic. In the mathematical formulations, y_i, x_i, $\widehat{y_i}$, and n denote the experimental COD removal (%), predicted COD removal (%), average results of experimental COD removal (%), and total number of experimental data, respectively.

Table 3. The mathematical equations of the employed statistical metrics in this study along with the perfect match value for each.

Performance Index	Mathematical Formulation	Perfect match value
Mean Absolute Error (MAE)	$\frac{\sum_{i=1}^{n}\lvert y_i - x_i\rvert}{n}$	0
Mean Absolute Percentage Error (MAPE)	$\frac{\sum_{i=1}^{n}\left\lvert\frac{y_i - x_i}{y_i}\right\rvert}{n}$	0
Root Mean Squared Error (RMSE)	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}}$	0
Coefficient of Determination (R^2)	$1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	1

3. Results and discussion

3.1. Model results

Figure 3 shows the coefficient of determination for both training and testing data on the three ensemble models and the super learner model. The training and testing data points are represented using blue and red, respectively. The plots also have a visual aid in the form of a 20% error-bound to indicate relative deviation between actual and predicted results. Here, the super learner model is shown to excel in its prediction ability compared to the three ensemble models for the testing data. Whereas for the training data, as demonstrated by Table 4, XGB shows the best results for the chosen statistical indicators. As it scored 0.193 %, 0.275 %, 0.692 %, and 99.905 % for MAE, MAPE, RMSE, and R-squared, respectively. The significant difference between the XGB model's performance on the training data relative to the other models, including the super learner is likely due to the overfitting of the training data on the XGB model observed in Figure 3. The effects of this become apparent when observing the model's performance in the testing dataset,

as it was exceeded by both the Gradient Boosting and super learner models. The fit (R^2) for the super learner based on the testing dataset was 95.260 %. While the MAE, MAPE, and RMSE results were 3.225 %, 6.332 %, and 4.144 %, respectively, clearly demonstrating how stacking the three base models can produce a significantly improved model. Even when compared to the model with the second-best performance, which was Gradient Boosting. Figure 4 shows the residual plots with percentage errors for the models, with the 20% error bound illustrated in light grey. The 20% bound is significant as an indicator for the accuracy of the four models [46].

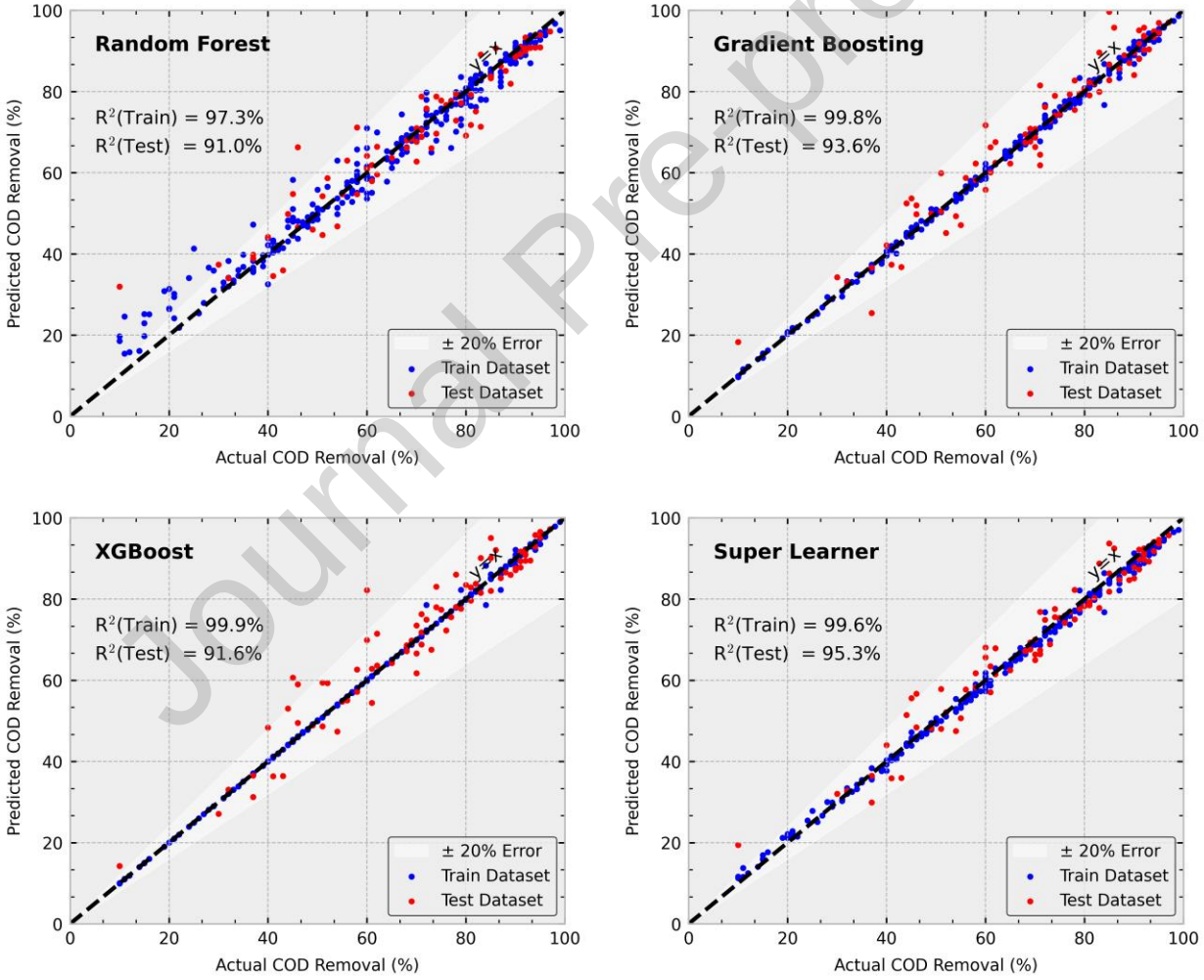

Figure 3. Predicted and actual COD removal for the three training models and the super learner

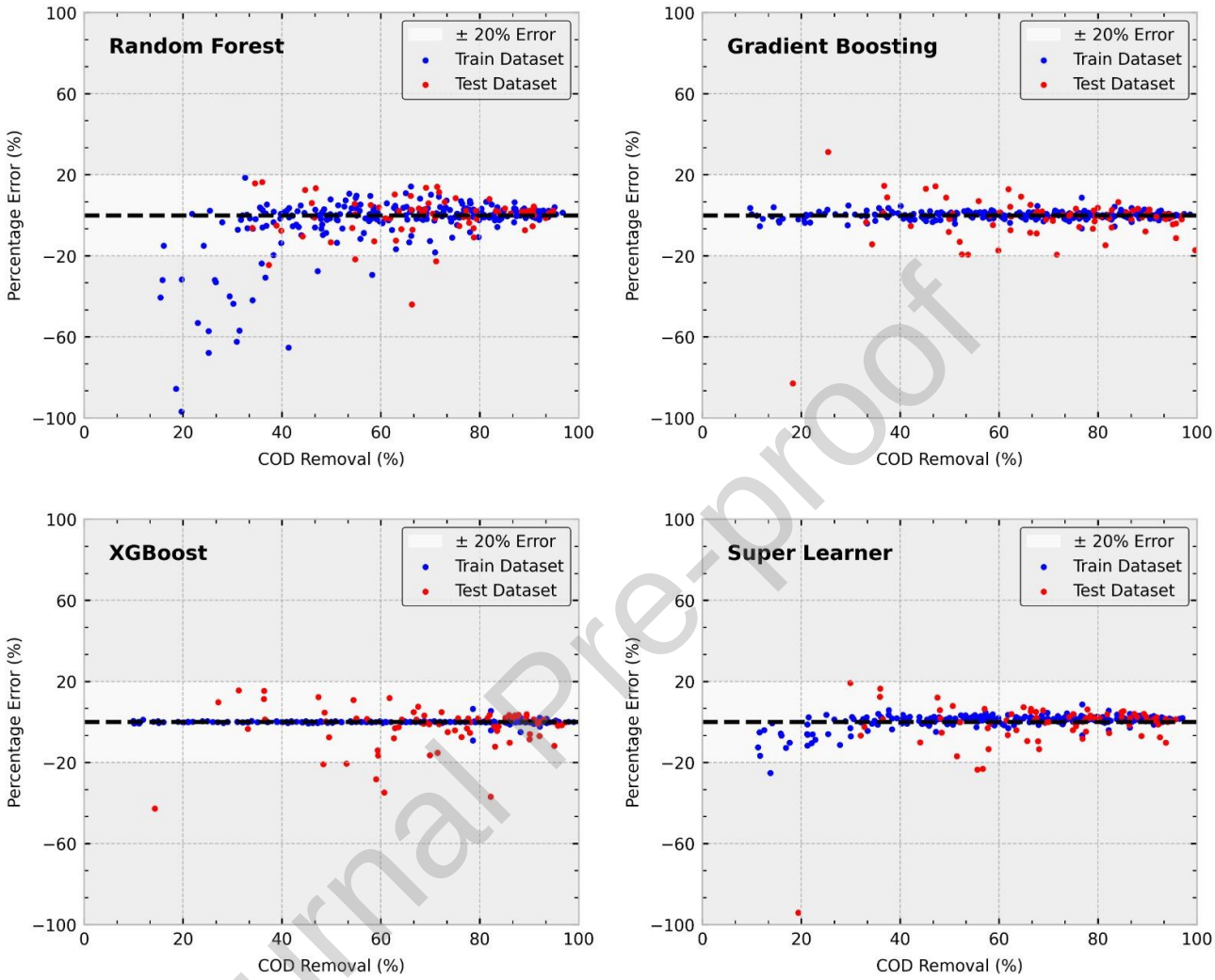

Figure 4. Residual plots showing the percentage errors for the three base models and the super learner model.

Table 4. Model results based on the chosen statistical indicators, and best performance outcomes in the training and testing stages are highlighted in bold.

Metrics	Training set				Testing set			
	RFR	GBR	XGB	Super learner	RFR	GBR	XGB	Super learner
MAE (%)	2.367	0.697	**0.193**	1.078	4.003	3.485	3.869	**3.225**

MAPE (%)	6.421	1.186	**0.275**	2.042	9.069	6.668	6.930	**6.332**
RMSE (%)	3.667	1.042	**0.692**	1.369	5.725	4.797	5.500	**4.144**
R^2 (%)	97.328	99.784	**99.905**	99.628	90.953	93.648	91.650	**95.260**

3.2. Model interpretation

3.2.1. Partial Dependence Plots (PDPs)

PDPs are used to visualize the interaction between individual input features and their respective effect on the machine learning model's output prediction, assuming all other features remain constant. These plots illustrate the impact of changing a single feature on the prediction of the dependent variable, thereby making the results of advanced machine learning algorithms more interpretable and understandable. Figure 5 shows the PDPs for each numerical input parameter employed in the current study on the COD removal in the EC process. It shows that within the range of the parameter, 1-5 cm, removal significantly increases with increasing the inter-electrode distance. The improved removal capabilities arising from increasing the inter-electrode distance result from enhancing the circulation between the anode and cathode [14]. In addition, the inter-electrode distance impacts the formation velocity of flocs and their ability to remove contaminants [47]. Figure 5 shows that A/V ratio has a relatively smaller impact on COD removal. Electrocoagulation studies tend to not consider the A/V ratio's impact on COD removal efficiency. This can explain why it is less impactful relative to other parameters in the PDPs. However, an increased impact is observed when the A/V ratio is more than 15 m^2/m^3. Figure 5 shows a positive relation with COD removal when number of electrodes is between 4 to 6 electrodes, which has been observed in the literature [48]. However, their impact is relatively static otherwise. Within the examined ranges of conductivity, it negatively impacts COD removal at a smaller scale. Experimentally, the effect of conductivity is complex, as it was observed to have a very slight positive impact for iron electrodes and a negative impact for aluminium electrodes [49]. On the

other hand, at much higher conductivity values, reaching 25,800 μS/cm, the negative correlation between conductivity and COD removal is much more pronounced. Comparatively, current density has a much stronger positive impact on COD removal from 0 to 20 mA/cm^2. This is a result of current density increasing the anode dissolving rate. Thereby improving pollutant removal efficiency [50]. The plot then plateaus afterwards due to minimal available data points after the range between 0-20 mA/cm^2. While for pH, no significant impact on COD removal through most of the examined range was observed. The large drop observed as the pH reaches 10 is likely due to how the negatively charged flocs reduce contaminant removal in alkaline conditions [14]. The initial concentration of COD shows a negative relation with COD removal at the examined range. This is due to the formed flocs being insufficient in removing high concentrations of pollutants [51]. Though the extremely large values of initial COD do not follow this trend, likely due to minimal available data at the higher ranges. Whereas electrolysis time significantly improves COD removal up to a point, after which it plateaus. This was also observed in experimental studies under different ranges [23], [25], [50]. Thus, electrolysis time can be concluded to have an optimum value beyond which it bears no significant impact on COD removal efficiency. For the case of the anode and cathode materials, despite being categorical parameters, they were introduced to the model as numerical parameters using the OneHotEncoder function. However, introducing them as numerical parameters of 0s and 1s only hinders the process of producing PDPs for the two categorical parameters. For that reason, the two parameters were only included in the sensitivity analysis section.

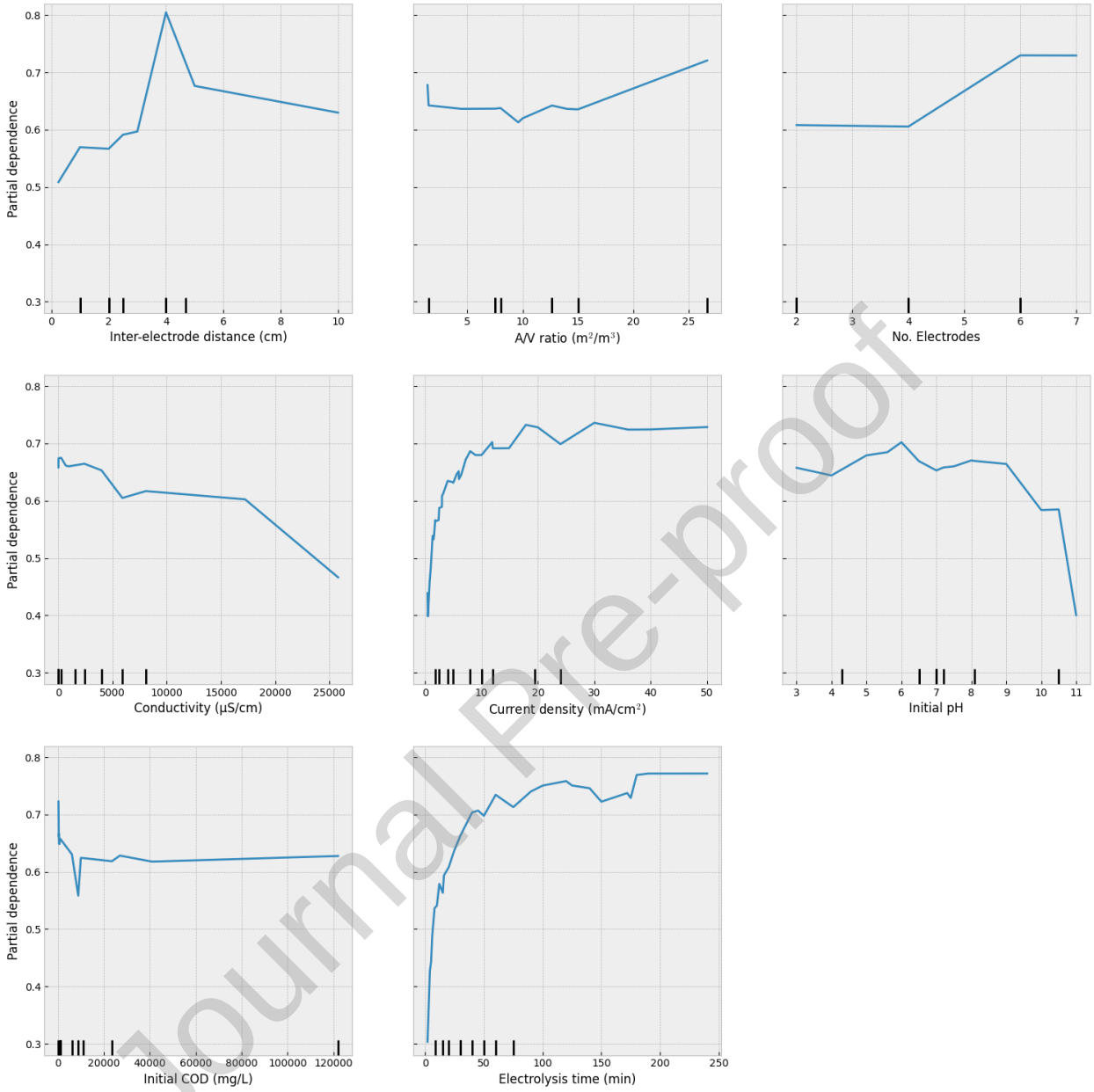

Figure 5. PDPs for the 8 numerical input parameters based on the super learner.

3.2.2. Sensitivity Analysis

This section assessed the influence of individual input features on the super learner model's prediction of the output variable (i.e., COD removal) through a univariate sensitivity analysis. To

achieve that, the values for each feature varied systematically across the observed range while keeping all the other EC process-related inputs held constant. Then, the model predictions were computed for each modified input, and the mean absolute deviation from the base predictions was calculated. The base predictions here refer to the super learner model's predicted outputs using the original dataset, before changing that specific input. This approach assists in quantifying the relative importance of each feature based on its effect on the model's predictions. Figure 6 shows the sensitivity analysis results for all eight numerical input parameters as well as the two categorical parameters, where the y-axis shows the mean sensitivity of each parameter. As the mean sensitivity approaches zero, the impact of the corresponding parameter is lower. The three most important parameters being the electrolysis time, inter-electrode distance, and the current density. This is in agreement with the PDPs of the same parameters. As all three impact COD removal significantly over a large portion of the data range. Whereas conductivity is only slightly less impactful. The sensitivity analysis result of conductivity is also comparable to its PDP. As conductivity is effective at impacting COD removal but at a slightly smaller rate than parameters like electrolysis time and inter-electrode distance. On the other hand, the number of electrodes, pH, and initial COD concentration all affect COD removal efficiency in a small portion of their ranges, leading to their lower sensitivity analysis scores. This is likely due to most studies focusing on investigating parameters like inter-electrode distance and current density, as they are much easier to manipulate compared to indicators of wastewater condition. Whereby the resulting dataset has a more equal distribution of values for operational parameters like current density and electrolysis time compared to wastewater parameters like pH and initial COD concentration. As most studies use either one or at maximum two wastewater types for their experiments. Similarly, A/V ratio is also impacted by this. where most studies used for constructing the super learner did

not consider it. and most used the same reactor volume and effective electrode surface area throughout their experiments with only one exception [27]. Comparing the two categorical input parameters, it is apparent that the cathode material impacts the model's behaviour more significantly compared to the anode material. In fact, anode material is the least impactful input parameter on COD removal efficiency in EC to the model. One of the studies used in developing the super learner examined the impact of anode materials on COD removal efficiency [13]. In that study, the COD removal did not change significantly between the different anode materials used. Which can explain why the anode material parameter has a small impact on COD removal efficiency.

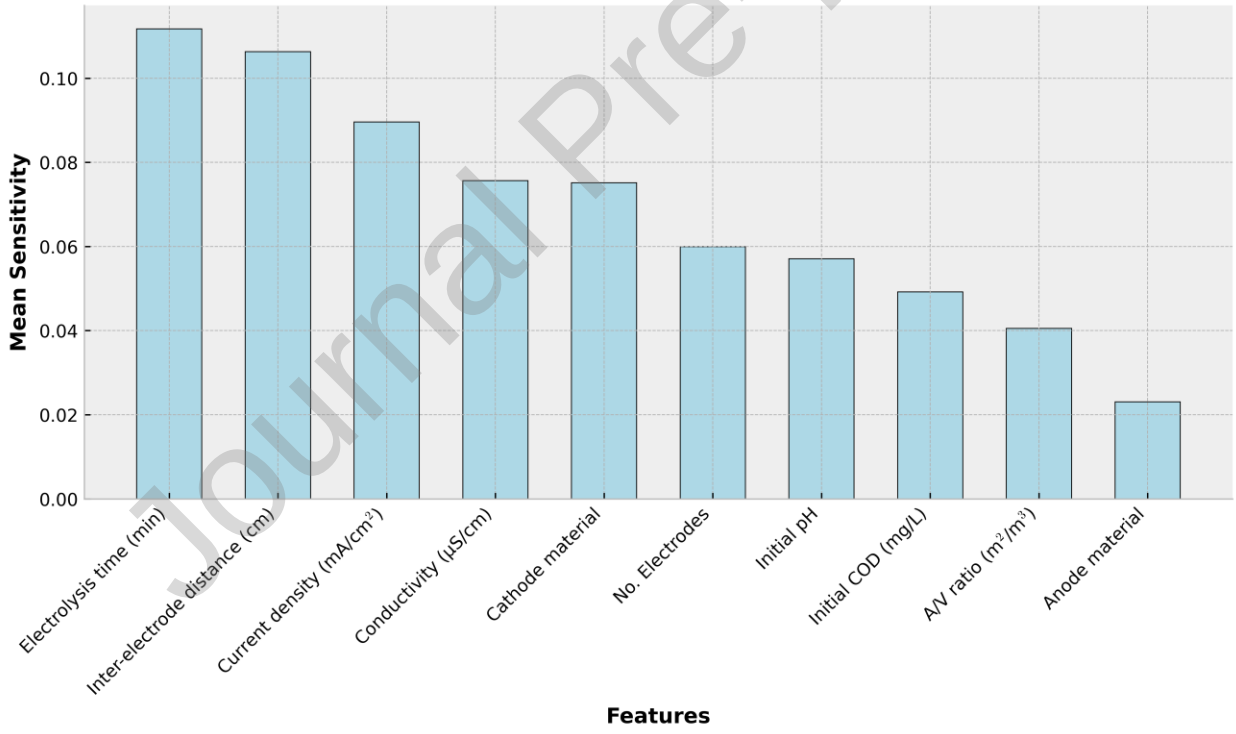

Figure 6. Sensitivity analysis plot for all ten input parameters used in developing the super learner model.

4. Conclusion

The study produced a highly effective machine learning model to predict COD removal using electrocoagulation based on a comprehensive dataset collected from the literature. For that purpose, three base models, consisting of RFR, GBR, and XGB were trained and tested at a ratio of 80:20 on 379 data points from 11 research papers with no missing data. Following that, a linear regression model was trained on parameters produced by the base models as a meta learner. This modelling approach provides much needed variety in the modelling techniques used in investigating COD removal efficiency in EC. In order to assess the integrity of the super learner, 4 statistical indicators were chosen, which were MAPE, MAE, RMSE, and R^2. The resulting super learner model had a predictive accuracy of 95% (R^2) in the test dataset. Finally, the results of the model were interpreted using a combination of PDPs for the numerical parameters and sensitivity analysis for all input parameters. Combining PDPs and the sensitivity analysis allowed for accurate interpretation of input parameter impact on COD removal efficiency irrespective of whether the parameter was numerical or categorical. The results from both show that the most impactful parameters on COD removal within the model were electrolysis time, inter-electrode distance, and current density. The results can be utilized for developing larger scale EC systems through the inclusion of the A/V ratio as an input parameter. Due to the parameter's flexibility when examining EC process scalability. The interpretability also determined the most significant parameters that influence the process, which can assist significantly in future implementations of EC systems. Additionally, the collected dataset by the authors serves as an important resource for researchers to develop their own unique machine learning models by utilizing the extensive dataset.

Acknowledgement

The authors gratefully acknowledge the financial support provided by Qatar Research Development and Innovation (QRDI), research grant (MME03-1015-210003). Additionally, one of the authors would like to acknowledge the support received through the Graduate Sponsorship Research Award (GSRA10-L-1-0602-23075) from QRDI. The statements made herein are solely the responsibility of the authors.

References

[1] S. K. Patel *et al.*, "State of the art review for industrial wastewater treatment by electrocoagulation process: Mechanism, cost and sludge analysis," *Desalination Water Treat*, vol. 321, p. 100915, Jan. 2025, doi: 10.1016/j.dwt.2024.100915.

[2] N. Daneshvar, A. R. Khataee, and N. Djafarzadeh, "The use of artificial neural networks (ANN) for modeling of decolorization of textile dye solution containing C. I. Basic Yellow 28 by electrocoagulation process," *J Hazard Mater*, vol. 137, no. 3, pp. 1788–1795, Oct. 2006, doi: 10.1016/j.jhazmat.2006.05.042.

[3] B. Li, C. Lu, J. Zhao, J. Tian, J. Sun, and C. Hu, "Operational parameter prediction of electrocoagulation system in a rural decentralized water treatment plant by interpretable machine learning model," *J Environ Manage*, vol. 333, p. 117416, May 2023, doi: 10.1016/j.jenvman.2023.117416.

[4] M. Akoulih *et al.*, "Electrocoagulation-based AZO DYE (P4R) Removal Rate Prediction Model using Deep Learning," *Procedia Comput Sci*, vol. 236, pp. 51–58, 2024, doi: 10.1016/j.procs.2024.05.003.

[5] K. Z. Tenodi, S. Tenodi, J. Nikić, E. Mohora, J. Agbaba, and S. Rončević, "Optimizing arsenic removal from groundwater using continuous flow electrocoagulation with iron and aluminum electrodes: An experimental and

modeling approach," *Journal of Water Process Engineering*, vol. 66, p. 106082, Sep. 2024, doi: 10.1016/j.jwpe.2024.106082.

[6] S. Aber, A. R. Amani-Ghadim, and V. Mirzajani, "Removal of Cr(VI) from polluted solutions by electrocoagulation: Modeling of experimental results using artificial neural network," *J Hazard Mater*, vol. 171, no. 1–3, pp. 484–490, Nov. 2009, doi: 10.1016/j.jhazmat.2009.06.025.

[7] M. Gholami Shirkoohi, R. D. Tyagi, P. A. Vanrolleghem, and P. Drogui, "A comparison of artificial intelligence models for predicting phosphate removal efficiency from wastewater using the electrocoagulation process," *Digital Chemical Engineering*, vol. 4, p. 100043, Sep. 2022, doi: 10.1016/j.dche.2022.100043.

[8] A. G. Merma, B. F. Santos, A. S. C. Rego, R. R. Hacha, and M. L. Torem, "Treatment of oily wastewater from mining industry using electrocoagulation: Fundamentals and process optimization," *Journal of Materials Research and Technology*, vol. 9, no. 6, pp. 15164–15176, Nov. 2020, doi: 10.1016/j.jmrt.2020.10.107.

[9] G. F. S. Valente, R. C. S. Mendonça, J. A. M. Pereira, and L. B. Felix, "Artificial neural network prediction of chemical oxygen demand in dairy industry effluent treated by electrocoagulation," *Sep Purif Technol*, vol. 132, pp. 627–633, Aug. 2014, doi: 10.1016/j.seppur.2014.05.053.

[10] P. Patel, S. Gupta, and P. Mondal, "Electrocoagulation process for greywater treatment: Statistical modeling, optimization, cost analysis and sludge management," *Sep Purif Technol*, vol. 296, p. 121327, Sep. 2022, doi: 10.1016/j.seppur.2022.121327.

[11] P. B. Bhagawati *et al.*, "Prediction of electrocoagulation treatment of tannery wastewater using multiple linear regression based ANN: Comparative study on plane and punched electrodes," *Desalination Water Treat*, vol. 319, p. 100530, Jul. 2024, doi: 10.1016/j.dwt.2024.100530.

[12] D. Marmanis, A. Thysiadou, V. Diamantis, A. Christoforidis, and K. Dermentzis, "Performance of Electrocoagulation Processes for the Removal of COD and Ammonia from High Salinity Landfill-leachate using Iron or Aluminum Electrodes," *Journal of Engineering Science and Technology Review*, vol. 14, no. 4, pp. 105–109, 2021, doi: 10.25103/jestr.144.14.

[13] K. Zaher, A. Elawwad, and R. Nadeem, "Wastewater treatment by electrocoagulation: A comparative study using different anode materials," in *World Congress on Civil, Structural, and Environmental Engineering*, Avestia Publishing, 2019. doi: 10.11159/iceptp19.152.

[14] M. Priya and J. Jeyanthi, “Removal of COD, oil and grease from automobile wash water effluent using electrocoagulation technique.,” *Microchemical Journal*, vol. 150, Nov. 2019, doi: 10.1016/j.microc.2019.104070.

[15] M. Koopialipoor, P. G. Asteris, A. Salih Mohammed, D. E. Alexakis, A. Mamou, and D. J. Armaghani, “Introducing stacking machine learning approaches for the prediction of rock deformation,” *Transportation Geotechnics*, vol. 34, p. 100756, May 2022, doi: 10.1016/j.trgeo.2022.100756.

[16] Y. Song, J. Park, M.-S. Suh, and C. Kim, “Prediction of Full-Load Electrical Power Output of Combined Cycle Power Plant Using a Super Learner Ensemble,” *Applied Sciences*, vol. 14, no. 24, p. 11638, Dec. 2024, doi: 10.3390/app142411638.

[17] Y. Li *et al.*, “Using the super-learner to predict the chemical acute toxicity on rats,” *J Hazard Mater*, vol. 480, p. 136311, Dec. 2024, doi: 10.1016/j.jhazmat.2024.136311.

[18] M. Talhami *et al.*, “Modeling of flat sheet-based direct contact membrane distillation (DCMD) for the robust prediction of permeate flux using single and ensemble interpretable machine learning,” *J Environ Chem Eng*, vol. 13, no. 2, p. 115463, Apr. 2025, doi: 10.1016/j.jece.2025.115463.

[19] M. Talhami *et al.*, “Single and ensemble explainable machine learning-based prediction of membrane flux in the reverse osmosis process,” *Journal of Water Process Engineering*, vol. 57, p. 104633, Jan. 2024, doi: 10.1016/j.jwpe.2023.104633.

[20] B. Khaled, B. Wided, H. Béchir, E. Elimame, L. Mouna, and T. Zied, “Investigation of electrocoagulation reactor design parameters effect on the removal of cadmium from synthetic and phosphate industrial wastewater,” *Arabian Journal of Chemistry*, vol. 12, no. 8, pp. 1848–1859, Dec. 2019, doi: 10.1016/j.arabjc.2014.12.012.

[21] M. Arbabi, S. Shafiei, S. Mehraban, A. Khodabakhshi, A. Abdoli, and A. Arbabi, “Electrocoagulation process using aluminum electrodes for treatment of baker’s yeast industry wastewater,” *Int J Environ Health Eng*, vol. 11, no. 1, Jan. 2022, doi: 10.4103/ijehe.ijehe_28_20.

[22] S. Farhadi, B. Aminzadeh, A. Torabian, V. Khatibikamal, and M. Alizadeh Fard, “Comparison of COD removal from pharmaceutical wastewater by electrocoagulation, photoelectrocoagulation, peroxi-electrocoagulation and peroxi-photoelectrocoagulation processes,” *J Hazard Mater*, vol. 219–220, pp. 35–42, Jun. 2012, doi: 10.1016/j.jhazmat.2012.03.013.

[23] J. Lu *et al.*, “Treatment of wastewater from adhesive-producing industries by electrocoagulation and electrochemical oxidation,” *Process Safety and Environmental Protection*, vol. 157, pp. 527–536, Jan. 2022, doi: 10.1016/j.psep.2021.10.035.

[24] E. Bazrafshan, H. Moein, F. Kord Mostafapour, and S. Nakhaie, “Application of electrocoagulation process for dairy wastewater treatment,” *J Chem*, 2013, doi: 10.1155/2013/640139.

[25] W. Khanitchaidecha, K. Ratananikom, B. Yangklang, S. Intanoo, K. Sing-Aed, and A. Nakaruk, “Application of Electrocoagulation in Street Food Wastewater,” *Water (Switzerland)*, vol. 14, no. 4, Feb. 2022, doi: 10.3390/w14040655.

[26] C. J. Nawarkar and V. D. Salkar, “Solar powered Electrocoagulation system for municipal wastewater treatment,” *Fuel*, vol. 237, pp. 222–226, Feb. 2019, doi: 10.1016/j.fuel.2018.09.140.

[27] Y. Ozay and N. Dizge, “The effect of pre-treatment methods on membrane flux, COD, and total phenol removal efficiencies for membrane treatment of pistachio wastewater,” *J Environ Manage*, vol. 310, May 2022, doi: 10.1016/j.jenvman.2022.114762.

[28] W. Pantorlawn, W. Khanitchaidecha, T. Threrujirapapong, D. Channei, and A. Nakaruk, “Electrocoagulation for spent coolant from machinery industry,” *Journal of Water Reuse and Desalination*, vol. 8, no. 4, pp. 497–506, Dec. 2018, doi: 10.2166/wrd.2017.057.

[29] J. Zhou, P. G. Asteris, D. J. Armaghani, and B. T. Pham, “Prediction of ground vibration induced by blasting operations through the use of the Bayesian Network and random forest models,” *Soil Dynamics and Earthquake Engineering*, vol. 139, p. 106390, Dec. 2020, doi: 10.1016/j.soildyn.2020.106390.

[30] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera, “A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities,” *Information Fusion*, vol. 64, pp. 205–237, Dec. 2020, doi: 10.1016/j.inffus.2020.07.007.

[31] B. Grillone, S. Danov, A. Sumper, J. Cipriano, and G. Mor, “A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings,” *Renewable and Sustainable Energy Reviews*, vol. 131, p. 110027, Oct. 2020, doi: 10.1016/j.rser.2020.110027.

[32] H. Zhang *et al.*, “A generalized artificial intelligence model for estimating the friction angle of clays in evaluating slope stability using a deep neural network and

Harris Hawks optimization algorithm," *Eng Comput*, vol. 38, no. S5, pp. 3901–3914, Dec. 2022, doi: 10.1007/s00366-020-01272-9.

[33] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst Appl*, vol. 134, pp. 93–101, Nov. 2019, doi: 10.1016/j.eswa.2019.05.028.

[34] H. Afzaal *et al.*, "Artificial neural modeling for precision agricultural water management practices," in *Precision Agriculture*, Elsevier, 2023, pp. 169–186. doi: 10.1016/B978-0-443-18953-1.00005-2.

[35] A. Miller, J. Panneerselvam, and L. Liu, "A review of regression and classification techniques for analysis of common and rare variants and gene-environmental factors," *Neurocomputing*, vol. 489, pp. 466–485, Jun. 2022, doi: 10.1016/j.neucom.2021.08.150.

[36] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst Appl*, vol. 244, p. 122778, Jun. 2024, doi: 10.1016/j.eswa.2023.122778.

[37] A. Malik, Y. T. Javeri, M. Shah, and R. Mangrulkar, "Impact analysis of COVID-19 news headlines on global economy," in *Cyber-Physical Systems*, Elsevier, 2022, pp. 189–206. doi: 10.1016/B978-0-12-824557-6.00001-7.

[38] W. Zhang, X. Gu, L. Hong, L. Han, and L. Wang, "Comprehensive review of machine learning in geotechnical reliability analysis: Algorithms, applications and further challenges," *Appl Soft Comput*, vol. 136, p. 110066, Mar. 2023, doi: 10.1016/j.asoc.2023.110066.

[39] H. Ahmetoglu and R. Das, "A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions," *Internet of Things*, vol. 20, p. 100615, Nov. 2022, doi: 10.1016/j.iot.2022.100615.

[40] H. Belyadi and A. Haghighat, "Supervised learning," in *Machine Learning Guide for Oil and Gas Using Python*, Elsevier, 2021, pp. 169–295. doi: 10.1016/B978-0-12-821929-4.00004-4.

[41] M. Lu *et al.*, "A Stacking Ensemble Model of Various Machine Learning Models for Daily Runoff Forecasting," *Water (Basel)*, vol. 15, no. 7, p. 1265, Mar. 2023, doi: 10.3390/w15071265.

[42] A. Ghasemieh, A. Lloyed, P. Bahrami, P. Vajar, and R. Kashef, "A novel machine learning model with Stacking Ensemble Learner for predicting emergency

readmission of heart-disease patients," *Decision Analytics Journal*, vol. 7, p. 100242, Jun. 2023, doi: 10.1016/j.dajour.2023.100242.

[43] L. Li, L. Zuo, G. Wei, S. Jiang, and J. Yu, "A stacking machine learning model for predicting pullout capacity of small ground anchors," *AI in Civil Engineering*, vol. 3, no. 1, p. 11, Dec. 2024, doi: 10.1007/s43503-024-00032-8.

[44] P. G. Asteris, A. D. Skentou, A. Bardhan, P. Samui, and K. Pilakoutas, "Predicting concrete compressive strength using hybrid ensembling of surrogate machine learning models," *Cem Concr Res*, vol. 145, p. 106449, Jul. 2021, doi: 10.1016/j.cemconres.2021.106449.

[45] D. J. Armaghani and P. G. Asteris, "A comparative study of ANN and ANFIS models for the prediction of cement-based mortar materials compressive strength," *Neural Comput Appl*, vol. 33, no. 9, pp. 4501–4532, May 2021, doi: 10.1007/s00521-020-05244-4.

[46] P. G. Asteris *et al.*, "Predicting uniaxial compressive strength of rocks using ANN models: Incorporating porosity, compressional wave velocity, and schmidt hammer data," *Ultrasonics*, vol. 141, p. 107347, Jul. 2024, doi: 10.1016/j.ultras.2024.107347.

[47] J. F. Martínez-Villafañe *et al.*, "Interelectrode Distance Analysis in the Water Defluoridation by Electrocoagulation Reactor," *Sustainability*, vol. 14, no. 19, p. 12096, Sep. 2022, doi: 10.3390/su141912096.

[48] I. Amri, Z. Meldha, S. Herman, D. Karmila, Mhd. Fadilah Ramadani, and Nirwana, "Effects of electric voltage and number of aluminum electrodes on continuous electrocoagulation of liquid waste from the palm oil industry," *Mater Today Proc*, vol. 87, pp. 345–349, 2023, doi: 10.1016/j.matpr.2023.03.621.

[49] S. Manikandan and R. Saraswathi, "Electrocoagulation technique for removing Organic and Inorganic pollutants (COD) from the various industrial effluents: An overview," *Environmental Engineering Research*, vol. 28, no. 4, pp. 220231–0, Sep. 2022, doi: 10.4491/eer.2022.231.

[50] S. Boinpally, A. Kolla, J. Kainthola, R. Kodali, and J. Vemuri, "A state-of-the-art review of the electrocoagulation technology for wastewater treatment," *Water Cycle*, vol. 4, pp. 26–36, 2023, doi: 10.1016/j.watcyc.2023.01.001.

[51] D. Tibebe, Y. Kassa, and A. N. Bhaskarwar, "Treatment and characterization of phosphorus from synthetic wastewater using aluminum plate electrodes in the electrocoagulation process," *BMC Chem*, vol. 13, no. 1, p. 107, Dec. 2019, doi: 10.1186/s13065-019-0628-1.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Highlights

- A comprehensive dataset of 379 COD removal results was compiled from electrocoagulation studies.
- The dataset considered 10 distinct input parameters and was free of missing data.
- The stacking model was developed using 3 ensemble models with linear regression as a meta-learner.
- The stacking (super learner) model exhibited high prediction capacity with an R^2 of 95.3% in the test dataset.
- The super learner model was interpreted using partial dependence plots and sensitivity analysis.
- The most impactful parameters on COD removal were electrolysis time, inter-electrode distance, and current density.