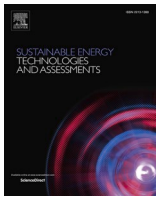
Sustainable Energy Technologies and Assessments 53 (2022) 102608

Contents lists available at ScienceDirect

Sustainable Energy Technologies and Assessments

journal homepage: www.elsevier.com/locate/seta



Sustaining struvite production from wastewater through machine learning based modelling and process validation

Krishnamoorthy Nageshwari [a], Vimaladhasan Senthamizhan [b], Paramasivan Balasubramanian [a,*]

[a] Department of Biotechnology & Medical Engineering, National Institute of Technology Rourkela, Odisha 769008, India
[b] Centre for Biotechnology, Anna University 600025, India

ARTICLE INFO

Keywords:
Struvite
Phosphorus recovery
Ammonium recovery
Machine learning method
eXtreme Gradient Boosting

ABSTRACT

The looming scarcity of phosphorus rock and intensification of its extraction for fertilizing applications has triggered the researchers to work upon a potential alternative such as struvite precipitation from wastewaters. Struvite production at commercial scale requires the support of novel prediction tools to smoothen the planning and execution processes. The present work aims at predicting the struvite recovery using several machine learning algorithms such as linear regression model, polynomial regression model, random forest regression model and eXtreme Gradient Boosting (XGB) regression model. Datasets for ten significant process parameters such as pH, temperature, concentrations of phosphate, ammonium and magnesium, stirring speed, reaction and retention time, drying temperature and time of various wastewater sources were collected for predicting the recovery. To minimize the loss function, extensive grid search hyperparameter tuning was performed to optimize the model. XGB was found to be the most robust method for prediction of nutrient recovery as struvite. The highest regression coefficient (R^2) of 0.9683 and 0.9483 were achieved for phosphate and ammonium recoveries, respectively. The key influencing factors on target output were studied using SHapley Additive exPlanations (SHAP) plots that depicts the interactive effect of each of the input parameters on phosphate and ammonium recovery. Experimental validation was carried out to further support the model predictions.

Introduction

Wastewater produced from domestic, municipal, commercial and industrial areas contains significant concentration of phosphate and nitrogen. In most cases, these valuable nutrients are left untreated causing algal blooms and eutrophication of the natural aquatic reservoirs. High concentration of ammonia in waterbodies can cause toxicity while preventing excretion of toxicant by aquatic animals, leading to accumulation in body tissues and blood, followed by death [47,28]. In the pressing scenario where phosphate minerals are facing extinction, recycling and utilization of the minerals in a sustainable way is the need of the hour [24,27]. The recovered nutrients can find use as fertilizers in the agricultural sectors. There are several nutrient recovery technologies such as adsorption, thermochemical and biological processes. However, there are many disadvantages associated with these technologies like requirement of a desorption process, specific adsorbents, downstream processes, high energy and chemical input etc. Struvite ($MgNH_4PO_4{\cdot}6H_2O$) precipitation is emerging to be one of the promising technologies for simultaneous recovery of phosphate (PO_4^{3-}) and ammonium (NH_4^+). Struvite recovery through chemical precipitation is simple, cost-effective, spontaneous and doesn't involve excessive downstream processes or energy [49,26]. As per the standard methods of American Public Health Association (APHA), the theoretical value of struvite was estimated to be 12.6 % phosphate, 9.9 % magnesium and 5.7 % ammonium [43]. However, from a practical and commercial point of view, the nutrient value might vary depending on several parameters that influence struvite recovery like pH, temperature, ion composition, stirring rate etc. Hence, maintenance of such optimum process conditions is necessary to attain better and consistent yields.

Struvite crystallization depends on various environmental and physicochemical conditions. Sustaining an alkaline pH in the medium is the most indispensable requirement for struvite formation. Similarly, as the bacterial activity in wastewater plays a vital role in determining its ion composition, maintaining a suitable pH also becomes important [25]. PO_4^{3-}, NH_4^+ and magnesium ions (Mg^{2+}) are the primary reactive species involved in struvite recovery and are theoretically said to from in an equimolar ratio [21,27]. Drying process also enhances the storage capacity of struvite to make them marketable. Wastewaters have also

* Corresponding author.
E-mail address: biobala@nitrkl.ac.in (P. Balasubramanian).

https://doi.org/10.1016/j.seta.2022.102608
Received 13 February 2022; Received in revised form 10 May 2022; Accepted 22 July 2022

known to be rich in bacterial population which can precipitate along with struvite hindering their direct application in agricultural field [58,29]. Once the precipitation reaction is complete, the crystals are dried to get rid of certain undesirable effects such as bacterial activity and odour, making them suitable for long-term storage [7,46].

Commercialization of struvite is blooming and various reactor strategies such as fluidized bed reactors, stirred tank reactors, ion exchange reactors and bioelectrochemical systems are constructed to serve the productivity [5,15,23]. Parallelly, several prediction models like polynomial, chemical equilibrium model, thermodynamic model for nucleation and kinetics are established to predict phosphate and ammonium recovery with suitable input variables [23,45]. However, recently developed machine learning algorithms possess many advantages with respect to efficiency and accuracy than conventional models. The most important advantage of machine learning is that the model does not make any assumptions regarding the data [8]. In the cases of other models, may it be polynomial, chemical or thermodynamic, the prediction model assumes some property or behaviour regarding the system, which might not exactly replicate the real-world process. Since machine learning is purely data-driven, it can obtain the best approximation.

Model selection is a process that can be applied both across different types of models (*e.g.* Logistic Regression, Support Vector Machines, K-Nearest Neighbor, etc.) and across models of the same type configured with different model hyperparameters (*e.g.* different kernels in a Support Vector Machines) [37]. Machine learning, more specifically the field of predictive modelling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables. Linear regression is a linear model, *e.g.,* a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, y can be calculated from a linear combination of the input variables (x). When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression [39].

Polynomial regression is a special case of linear regression where one fit a polynomial equation on the data with a curvilinear relationship between the target variable and the independent variables. In a curvilinear relationship, the value of the target variable changes in a non-uniform manner with respect to the predictor [3]. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and techniques called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees [42,60].

A gradient descent procedure is used to minimize the loss when adding trees. Traditionally, gradient descent is used to minimize a set of parameters, such as the coefficients in a regression equation or weights in a neural network. After calculating error or loss, the weights are updated to minimize that error. Instead of parameters, weak learner sub-models or more specifically decision trees are available. After calculating the loss, to perform the gradient descent procedure, a tree is added to the model that reduces the loss (i.e., follow the gradient). This is done by parameterizing the tree, then modifying the parameters of the tree and moving them in the right direction for reducing the residual loss. Generally, this approach is called functional gradient descent or gradient descent with functions [36]. eXtreme Gradient Boosting (XGB or XGBoost) algorithm uses this approach for prediction implementation.

To interpret the outcomes of the developed model, SHAP (SHapley Additive exPlanations) values are studied. SHAPley values are derived from a game theoretic approach to elucidate the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions [44]. SHAP provides the following major benefits: Global interpretability — The collective SHAP values can show what proportion each predictor contributes, either positively or negatively, to the target variable. This is like the variable importance plot but it will be able to show the positive or negative relationship for each variable with the target. Local interpretability — Each observation gets its own set of SHAP which greatly increases its transparency. The prediction and therefore the contributions of the predictors can be easily explained. Traditional variable importance algorithms only show the results across the entire population but not on each individual case. The local interpretability enables us to pinpoint and contrast the impacts of the factors [34,35].

Previous models developed for struvite such as chemical equilibrium model, thermodynamic model and logistic model have focused on either phosphorus or ammonium recovery from a particular source of wastewater or with conditions specific to a reactor design. This study aims to (1) develop robust machine learning model to predict phosphate (PO_4^{3-}) as well as ammonium (NH_4^+) recovery as struvite by comparing the performance of several well-established models, (2) establish wide applicability of the models by involving data collected from published articles across several wastewater types rich in nutrients for struvite crystallization, (3) deduce the performance efficiency of the model using robust statistical metrics, (4) validate the model predictions with training and testing aspects and experimental values and (5) to comprehend the feature importance of the input variables and their impact on the targeted outcome using SHAP dependence and summary plots. To the best of author's knowledge, this is the foremost machine learning study in this carried out for prediction of struvite recovery.

Materials and methods

Data collection and pre-processing

The database for the development of a machine learning model was constructed by collecting suitable data from manuscript, figures, tables and supplementary data of published literature. The period opted for data collection is 1997–2019 and renowned databases such as Scopus, Web of Science and Google scholar were used. The articles were chosen based on search analysis with keywords such as struvite, phosphate recovery etc. (articles on medical aspects of struvite were neglected) and information available on the influential parameters under study. It consisted of 100 datasets from 85 scientific documents on struvite precipitation from wastewater (**Table S1**). The input variables under consideration were pH, reaction temperature, concentration of phosphate, ammonium and externally added magnesium, stirring rate, reaction time, retention time, drying temperature and drying time. The target variables of the model were phosphate and ammonium recovered as struvite. The articles were chosen based on the information available on the influential parameters under study.

The collected data were compiled into a structured table for further study. During initial exploratory data analysis (EDA), it was found that there were significant number of missing values in the dataset. There are three main problems that missing data can cause: a) missing data can introduce a substantial amount of bias, b) make the handling and analysis of the data more laborious, and c) create reductions in efficiency [6]. Hence, the technique of imputation was employed to handle the missing data. In statistics, imputation is the process of replacing missing data with substituted values. Imputation handles the aforementioned bottlenecks by replacing missing data with an estimated value based on other available information. Once all missing values have been imputed, the data set can then be analysed using standard techniques for complete data. *sklearn* package offers imputation tools to be used on incomplete datasets [42]. The 'most frequent' imputer mode was used where the imputer class fills the missing data points with the

arithmetic mode of the feature set in a fast and easy manner [38;64]. Though imputation with mode involves limitations such as over-representation of data and bias, using this method, one could ensure the complete dataset within the same distribution of the population [54]. Following the pre-processing, the dataset is directed to the next step of machine learning, Model Building.

Machine learning model selection

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset. There are two main classes of models: Probabilistic measures and Resampling methods. Probabilistic measures involve analytically scoring a candidate model using both its performance on the training dataset and the complexity of the model, whereas resampling methods seek to estimate the performance of a model (or more precisely, the model development process) on out-of-sample data. In this study, models belonging to resampling methods are used since they are robust to changes in the data and reduce the inherent bias that may exist in the distribution differences of the training and test sets.

The present study inherits the cohort of models mainly consist of linear regression models and tree-based models. Complex models were avoided as the main focus of the manuscript was to achieve good model explainability as much as model accuracy. SVM and deep learning architectures do not offer model interpretability even though they have a better chance of providing higher scoring models [19]. The candidate models used in this study are: linear regression model, polynomial regression model, random forest regression model and XGB regression model. All these models were implemented using *sklearn* 0.24.1 version.

Linear regression model

The developed database has more than one feature and hence a multiple linear regression was employed. Ordinary least squares were used as the loss function which seeks to minimize the sum of the squared residuals. Linear regression will often make more reliable predictions if the input variables are rescaled using standardization or normalization. Hence, the dataset was scaled to the range of 0–1 using *MinMaxScaler* class from *sklearn* package.

Polynomial regression model

The implementation of polynomial regression is a two-step process. First, the data was transformed into a polynomial using the *Polynomial Features* function from *sklearn* package and linear regression was used to fit the parameters to data. Here, a 2-degree polynomial was chosen based on the relationship between target and predictor. The 1-degree polynomial is a simple linear regression; therefore, the value of degree must be greater than 1. With the increasing degree of the polynomial, the complexity of the model also increases. Therefore, the value of n must be chosen precisely. If this value is low, then the model won't be able to fit the data properly and if high, the model will overfit the data [3]. Like linear regression, the loss function used here was Ordinary Least Squares.

Random forest regression model

Random forest (RF) has multiple decision trees as base learning models. Row and feature sampling can be randomly performed from the database forming sample datasets for every mode through bagging technique. In this study, RF regression with 100 trees were bootstrapped and aggregated to predict final output. This model uses bootstrap replicas, meaning it subsamples the input data with replacement. In case of selection of cut points, RF chooses the optimum split to split the nodes, (i.e.) the RF model splits the data into subsets of overlapping samples and trains a decision tree on each of these subsets. The nodes on the decision tree are split optimally in such a way that the classification capability between the nodes is maximum. Once all decision trees are trained, RF will aggregate the individual labels provided by each tree and output the label with the most votes [20,60].

Extra trees regression model

As an extension of the RF regression model, extra trees (ET) regression model was also developed with the same number of trees used for the random forest model. ET uses the entire original sample. In the ET *sklearn* implementation, there is an optional parameter that allows users to bootstrap replicas, but by default, it uses the entire input sample. This may increase variance because bootstrapping makes it more diversified. This model chooses the cut points randomly to split nodes. However, once the split points are selected, RF and ET algorithms choose the best one between all the subset of features. Therefore, ET adds randomization while ensuring optimization [41].

XGBoost model

XGB or XGBoost (eXtreme Gradient Boosting) is a powerful approach for building supervised regression models that uses gradient descent technique. The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e., how far the model results are from the real values. A benefit of the gradient boosting framework is that a new boosting algorithm does not have to be derived for each loss function that may want to be used, instead, it is a generic enough framework that any differentiable loss function can be used. The most common loss function in XGBoost for regression problems is ordinary least squares, and that for binary classification is logistic regression [2,40,59].

The advantage of XGB is that many input variables can be studied at once and the accuracy and capacity of the model were shown to be better than previous recorded approaches. XGB is a state-of-the-art algorithm that belongs to the set of classification and regression trees (CART) of tree ensemble techniques. It splits the node after every iteration in an additive way and the predictions are made by addressing the mistakes in the previous trees [48,50]. The final prediction is made by considering all the combined improvement rendered by the existing trees. The model possesses some unique features such as parallelization and regularization that helps prevent overfitting, making the prediction better compared to other models.

Model validation

Learning curve

A learning curve is a plot of model learning performance over experience or time. Learning curves are a widely used diagnostic tool in machine learning for algorithms that learn from a training dataset incrementally. The model was evaluated using the training dataset with a holdout validation dataset after each update during training. The plots of the measured performance were created to show learning curves. Reviewing learning curves of models during training can be used to diagnose problems with learning, such as an underfit or overfit model, as well as whether the training and validation datasets are suitably representative.

Learning curve for the most robust model was plotted to study the progress of the model's performance over a varied number of trees employed (**Fig. S1**). The model was run for 100 iterations, where the number of trees used was equal to the iteration number. A robust statistical metric, Root mean square error (RMSE) was recorded for each iteration, and it was plotted against the number of trees. The RMSE was calculated using **Eq. (1)** [32,59]. Such curves show how the model performs in terms of an articular hyperparameter (in this case, number of trees) and if the model converges at a certain score.

$$RMSE = \sqrt{\left[\frac{\sum_{i=1}^{N}\left(Y_i^{pred} - Y_i^{exp}\right)^2}{N}\right]} \quad (1)$$

where, Y_i^{pred} and Y_i^{exp} are predicted model values and experimental values and N is the number of data points.

True vs predicted plot

The model's performance was also evaluated by plotting the corresponding true vs predicted value pairs in a scatter plot and calculate the correlation coefficient (R^2) of the plot. The higher the R^2 value, the better the model performance. The R^2 can be calculated using **Eq. (2)**.

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{N}\left(Y_i^{exp} - Y_i^{pred}\right)^2}{\sum_{i=1}^{N}\left(Y_i^{exp} - Y_{ave}\right)^2}\right] \quad (2)$$

where, Y_i^{exp} and Y_i^{pred} are actual (experimental) and predicted values of the model. $Y_{ave}$is the average of the experimental values.

Experimental validation

For analysing the credibility of the model and extrapolation to the real-time application, experimental validation of the model generated input–output variables was carried out. Ten datasets containing values of the input variables were generated for both phosphate and ammonium recovery predictions. As the recovery depends on the nutrient composition and wastewater type, the datasets were chosen such that the range of all recovery percentages for both phosphate and ammonium are covered, which would help in better assessing the model performance. Synthetic urine was prepared according to the dataset concentration values of phosphate, ammonium and magnesium using sodium dihydrogen phosphate dihydrate, ammonium chloride and magnesium hydroxide, respectively. The pH was adjusted using 1 N NaOH and 1 N HCl solutions followed by required stirring and struvite settling time. The resultant phosphate and ammonium concentration in the supernatant solution was estimated by Murphy and Riley stannous chloride spectrophotometric method and phenol-hypochlorite method, respectively, using UV–Vis Spectrophotometer (Model 2230, Systronics, India). All the experiments were carried out in triplicates and the phosphate and ammonium recoveries were calculated using **Eq, 3** and **Eq. (4)**, respectively [25].

$$Phosphaterecovery(\%) = \frac{P_I - P_R}{P_I} \quad (3)$$

$$Ammoniumrecovery(\%) = \frac{A_I - A_R}{A_I} \quad (4)$$

where, $P_I$and A_I are initial concentrations of phosphate and ammonium (mg L^{-1}) and P_R and A_R are resultant concentrations of phosphate and ammonium (mg L^{-1}), respectively.

The performance of the model was assessed using a scatter plot with model predictions and experimental outcomes on the $\times$ and y axis, and estimating the R^2 using **Eq. (2)**.

Feature importance and SHAP plots for model explainability

The relationship or correlation between the input and target variables were studied using the feature importance plots generated by provision built-in the XGB algorithm and computed with SHapley Additive exPlanations (SHAP) values. The SHAP values are often calculated for any tree-based model, while other methods use rectilinear regression or logistic regression models as their surrogates [35]. In this study, *TreeExplainer* class from SHAP package was used to plot the dependence plot for each feature and summary plot. The summary plot combines feature importance with feature effects. Each point on the summary plot may be a SHAP value for an input feature in the model. SHAP dependence plots show average effects of feature on the outcome and the variance on y-axis. Especially in case of interactions, the SHAP dependence plot will be much more dispersed in the y-axis.

Results and discussion

Statistical analysis of input variables

Numerical values of ten input process parameters that are considered significant were collected to build a robust machine learning model for the prediction of phosphate and ammonium recovery of struvite. The input variables include pH, reaction temperature, concentrations of phosphate, ammonium and magnesium, reaction time, retention (precipitation time), stirring speed, drying temperature and drying time. Few of the input parameters (such as storage conditions, supersaturation ratio) and output parameters (such as struvite yield, purity and crystal size) were neglected due to insufficiency in the data that eventually might affect the accuracy of the model. The magnesium concentration values were obtained from the initial phosphate concentrations and PO_4^{3-}: Mg^{2+} molar ratios. The extreme data points (outliners) were included in this study as they represent the processing conditions of various wastewater sources, assuring wide application and suitability of the model. The different types of wastewaters taken into consideration in this study includes, swine wastewater, swine lagoon liquid, urine (undiluted and hydrolysed), semiconductor wastewater, livestock excreta, animal wastewater, cochineal insects processing wastewater, 7-aminocephalosporanic acid wastewater, coke oven wastewater, nylon wastewater, rare-earth wastewater, wastewater from leather tanning industry, abattoir wastewater or meat packing industry, dairy manure, poultry manure, municipal landfill leachate, textile printing industry, fertilizer industry, yeast industry, oil and gas industry, synthetic wastewater, beverage waste, effluent from anaerobic treatment of landfill leachate, baker's yeast industry, domestic wastewater, molasses-based industrial wastewater, sludge from sewage treatment plants, centrifugation sludge of nutrient removal plants and biologically treated opium alkaloid wastewater (**Fig. S4**). The primary reactive ion concentrations (phosphate and ammonium) will vary with each type of wastewater and hence it should be considered as a significant parameter for the prediction of struvite recovery. For instances, human urine, swine wastewater and cochineal insects processing wastewater contained high phosphate concentration, whereas anaerobic digestate effluents from opium alkaloid wastewater, baker's yeast industry and domestic wastewater had very low phosphate concentration. In case of ammonium, animal wastewater, human urine and municipal sewage wastewater consisted of high concentration and swine wastewater and effluent from opium alkaloid wastewater possess low concentration.

The statistical analysis of the data collected is shown in Table 1. The count indicates the number of datasets of each variable under consideration. The mean and standard deviation denotes the range of datasets of each input variable. The minimum and maximum values show the extreme data points among the distribution. The various quartile values provided will help in better comprehension of the data spread pattern. The phosphate and ammonium concentrations are the highest in source-separated human urine, making them one of the best and easily available sources for struvite recovery [14,29]. Also, cochineal insects processing wastewater has shown to contain high amount of phosphate with a 100 % recovery efficiency [9]. The PO_4^{3-}:Mg^{2+} ratio mostly lies between 1 and 2 and it has been reported that it is positively correlated with struvite yield. The pH for struvite crystallization is alkaline and often in the range between 8.0 and 10.0. Beyond this range, there are evidence for the precipitation of struvite analogues such as *K*-struvite ($MgKPO_4.6H_20$). The potassium ions in wastewater replaces the ammonium ions in struvite at pH greater than 10, forming magnesium potassium phosphate as secondary crystallized phase. K-struvite can also be used for agricultural applications, especially for soils deficient in potassium [22,33]. The reaction and retention times can be seen to vary from 0.5 to 360 min and 3 to 1600 min, respectively. Similarly, stirring of the medium can facilitate nucleation and has been ranged in literature between 20 and 5280 rpm. The drying temperature and time are also equally significant whose deviation can result in structural variations.

Table 1
Statistical analysis of the input and output variables included in developing a machine learning model for phosphate and ammonium recovery as struvite.

	pH	Reaction Temperature (°C)	PO_4^{3-} concentration (mg L^{-1})	NH_4^+ concentration (mg L^{-1})	Mg^{2+} concentration (mg L^{-1})	Stirring speed (rpm)	Reaction time (min)	Retention time (min)	Drying temperature (°C)	Drying time (hr)	PO_4^{3-} recovery (%)	NH_4^+ recovery (%)
Count	97.00	51.00	84.00	79.00	84.00	46.00	61.00	50.00	47.00	35.00	73.00	54.00
Mean	9.16	25.30	265.77	2834.32	81.89	535.21	34.91	177.08	45.82	43.65	0.90	0.83
Standard Deviation	0.65	4.32	446.44	11524.30	155.50	1109.66	58.80	345.15	20.74	68.03	0.09	0.16
Minimum	8.00	20.00	0.30	33.10	0.00	20.00	0.50	3.00	24.00	3.00	0.55	0.12
25 %	9.00	23.00	57.50	230.00	7.40	152.50	10.00	30.00	35.00	12.00	0.85	0.80
50 %	9.00	25.00	125.00	695.00	26.91	200.00	20.00	60.00	40.00	48.00	0.93	0.87
75 %	9.40	25.00	350.50	1390.00	100.58	300.00	30.00	180.00	50.00	48.00	0.97	0.92
Maximum	12.00	38.00	3490.00	101018.40	892.98	5280.00	360.00	1600.00	110.00	420.00	1.00	1.00

The temperature differed between 24 and 110 °C for a time scale of 3–420 hrs. However, it is reported in few articles that drying temperature beyond 50 °C can lead to loss of water molecules from struvite [4,7]. For all the above-mentioned conditions, the phosphate and ammonium recovery varied between 55 and 100 % and 12–100 %, respectively.

Data correlation is a basic quantity of any modelling technique that will help in understanding the dependency of multiple attributes or prediction of one attribute from another. Fig. 1 provides the Pearson correlation plot between the input and output variables of interest. The sign of the coefficients determines the positive or negative effect on the output. The magnitude of the effect can be interpreted by the variation in the colour spectrum as shown in Fig. 1. A positive correlation means, an increase in the value of a parameter will ultimately increase the other and vice versa. In this context, drying time is shown to adversely affect the phosphate recovery efficiency, whereas, pH has a slightly positive effect on both phosphate and ammonium recoveries. It can also be seen that phosphate and magnesium concentration, reaction temperature and retention time, drying time and phosphate concentration, and drying time and magnesium concentration have strong positive correlations among one another. The relation between phosphate and magnesium concentration could be because of the binding interaction between them for the formation of struvite due to which magnesium is generally added at a particular molar ratio with phosphate [26]. The interaction of drying time with phosphate and magnesium concentration could be because of the variation in struvite structure with higher drying time and temperature [7]. However, a direct correlation between reaction temperature and retention time hasn't been reported in the literature.

Comparison and selection of machine learning models

To build a robust machine learning model for prediction of PO_4^{3-} and NH_4^+ recovery, a number of well-established regression algorithms were explored. Regression models have proved to be better than empirical models in several circumstances for the investigation of relationship between a given dependent and independent variable [13]. In this study, the performance and accuracy of linear regression, polynomial regression, extra-trees (ET), random forest (RF), and XGBoost (XGB) regression were tested with a commonly used and robust statistical indicator, root mean squared error (RMSE). Linear and polynomial regressions are an elementary form of predictive analysis which form a base for machine

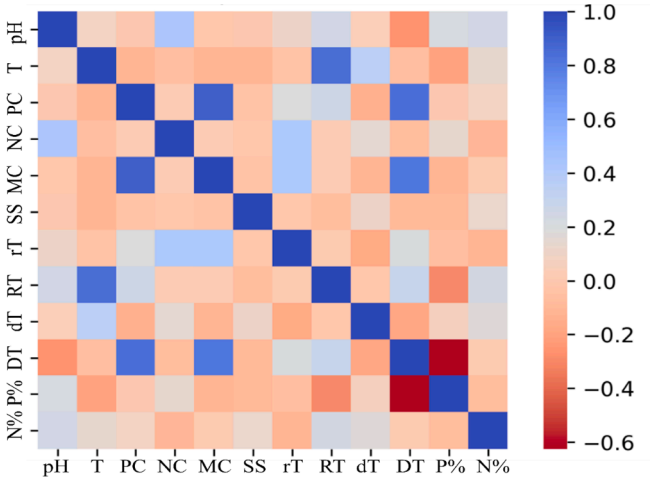


Fig. 1. Pearson correlation matrix depicting relationship between the input and output features of PO_4^{3-} and NH_4^+ recovery. (T: Temperature; PC: Phosphate concentration; NC: Ammonium concentration; MC: Magnesium concentration; SS: Stirring speed; rT: Reaction time; RT: Retention time; dT: Drying temperature; DT: Drying time; P%: Phosphate recovery; N%: Ammonium recovery).

learning. Linear regression can be used in case of continuous datasets that tend to have linear relationship with each other to fit in an equation for prediction. Similarly, polynomial regression can be applied when the variables have non-linear relationship and fit well in a polynomial equation. Following this, tree-based ensemble algorithms such as ET, RF and XGB were analysed. The RMSE values for each model were calculated using the equation. It can be seen in the Fig. 2 **a & b** that among the five models studied, XGB is the most preferable with least RMSE (almost close to zero) for prediction of both phosphate and ammonium recovery. It is clear that the features selected do not share linearity and hence application of supervised machine learning algorithm that beholds the non-linearity and interaction is required. The order of model ranks with respect to mean squared errors in both the cases is as follows: XGB < RF < ET < Polynomial < Linear. RF performed better than ET despite using original dataset is because that RF uses an optimal split which is effective in case of fewer trees (n = 100). ET might outdo RF if the number of trees is higher, but the training time will be higher which is undesirable. Thus, of all the models used in the study, XGB was able to outperform all other models. This is due to its robustness and generalizability in presence of high variance in data. Another advantage of using XGB is that feature importance can be obtained for the feature set used in the data.

Development and validation of eXtreme gradient Boosting (XGBoost) for PO_4^{3-} and NH_4^+ recovery

In this study, XGB has outperformed all the other machine learning models by securing the least RMSE value. However, with such high number of input variables, the model is prone to overfitting and requires pruning of statistical attributes, whose optimum combination can help in improvisation and optimization of the model. Some of the most important hyperparameters considered are maximum number of iterations, maximum depth and learning rate [40]. The *number of rounds or iterations* is the optimal number of trees in the model. The mean squared error values of the training curve decreased exponentially with increase in the number of trees and became almost zero at the 100th iteration (**Fig. S1**). The result remained the same at higher iterations and hence the number of trees was fixed to be 100. A 10-fold cross validation was opted which also might determine the iteration factor. The *maximum depth* determines the depth or size of the decision tree. A very high depth can make the model less conservative and hence was set to 5, (i.e.) trees that run deep before finishing the complete set of iterations tend to overfit and as the dataset is small and generalizability is the main priority, depth was set to 5, which was thought to be optimal. *Learning rate (or shrinkage)* is the rate at which the algorithm adapts itself to the growing model. It was set to 0.01 as a small positive number, which is said to be more robust to overfitting. All other parameters such as *sub_sample*, *reg_alpha* and *reg_lambda* was fixed at their default values of 1, 0 and 1, respectively. After the fine tuning, the entire dataset was divided as training and testing data in a 75:25 ratio.

Once the model is trained, the performance was evaluated and validated with a commonly used statistical measure, correlation coefficient (R^2). The scatter plot between actual (experimental) and values predicted by the model for the training and testing datasets is shown in the Fig. 2. R^2 indicates the level of similarity between × and y-axis as calculated by **Eq. (2)**. The R^2 for training datasets is 0.9971 and 0.9722 for PO_4^{3-} and NH_4^+, respectively **(**Fig. 2 **c & d)**. The R^2 values for testing datasets of PO_4^{3-} and NH_4^+ recovery after a 10-fold cross validation is 0.9683 and 0.9483, respectively **(**Fig. 2 **e & f)**. The experimental validation also showed promising results with R^2 values of 0.8095 and 0.8704 for PO_4^{3-} and NH_4^+, respectively **(**Fig. 2 **g & h)**. To the best of author's knowledge, this is so far the first machine learning model developed for the prediction of PO_4^{3-} and NH_4^+ recovery as struvite. The mechanistic (struvite equilibrium model and Crystallizer model v.2.0) and artificial neural network (NeuStruvite v.1.0) models were earlier developed by Forrest et al. [16]**,** which was mainly applied for the prediction of phosphate concentrations in the effluents with only the reacting struvite species and pH as the input parameters. Hence, the results of this study are compared and discussed with other theoretical models established in the past years. The polynomial and logistic kinetic models were chosen by Fang et al. [15] to predict PO_4^{3-} recovery efficiency by estimating struvite settling percentage (settleability index) as a key dependant variable. The logistic model proved to be better with a R^2 of 0.9661; however, the conditions preferred confines the model applicability to fluidized bed reactor. The probability of presence of coexisting ions such as calcium, carbonate and heavy metals were also taken into consideration in this model. Warmawadenthi and Lu [55] applied a chemical equilibrium model to predict ammonium recovery from anaerobic digester effluent and achieved an efficiency of 99.5 %. Similarly, a thermodynamic model was built using PHREEQC (an Interactive software) for PO_4^{3-} and NH_4^+ recovery from semiconductor wastewater. This model gave additional information on the feasibility of formation of struvite analogues such as bobierrite subject to alterations in PO_4^{3-}:Mg^{2+} molar ratio.

Evaluation of feature importance

The feature importance plot generated through the XGB algorithm provides information on the most influential parameters that have a significant effect of the required output. Once the algorithm constructs all the boosted trees, the importance scores are calculated explicitly for each decision tree indicating every input attribute in the model. In this study, the feature importance graph for PO_4^{3-} and NH_4^+ recovery as shown in Fig. 3 are plotted using the importance metrics called total gain. It is one the most appropriate metrics that can be used to understand the relative importance of every feature. The scores thus generated imply the sum of the gains attained by the relative contribution of each attribute in every single tree in the model. The higher the score, the more is the significance of the feature in making decisions with the decision trees.

For PO_4^{3-} recovery, Mg^{2+} concentration is shown to be the most valuable attribute in constructing the boosted decision trees for the corresponding model. According to literary references, PO_4^{3-} and Mg^{2+} concentrations are highly associated because of the following two prime reasons: 1) The amount of magnesium is relatively low in wastewaters due to which an external supply is required; 2) As the theoretical struvite formation occurs in the presence of equimolar concentration of the reactive ions, magnesium is supplied in levels relative to PO_4^{3-} molar concentration [30]. In several articles, the Mg^{2+}:PO_4^{3-} molar ratio is maintained to be 1:1 and a higher ratio is said to improve the PO_4^{3-} recovery as well as the struvite yield. The latter is applied when the high concentration of coexisting ions present in wastewater has to be tackled, which is prevalent in most of the real-time cases [1,53]. However, such increased addition has been reported to affect the purity by formation of compounds other than struvite crystals. For example, molar ratio greater than 2.5 leads to the formation of bobierrite [25;57].

The second-most significant parameter is the reaction temperature at which the precipitation is carried out. Though crystallization takes place at room temperature in general, the effect of temperature was studied by several researchers and found that a slightly higher temperature can improve PO_4^{3-} recovery as struvite by facilitating the solubilisation of insoluble PO_4^{3-} in effluents [57]. In addition, higher temperature has been reported to improve ureolysis, resulting in elevated ammonium formation in case of urine [10;53]. Stirring speed is equally important and aids in enhancing nucleation and crystal growth. An optimal mixing rate (150–300) can generate crystals of bigger sizes, while higher rates (greater than1000 rpm) disrupt the bonding [26]. Drying temperature is relatively the least significant as per the model predictions. However, drying temperatures also can cause an adverse effect on the structural integrity of the struvite crystals; yet the condition holds true only when struvite is exposed to temperatures greater than 40 °C for a prolonged time. Hence, drying temperature is not as influential as other input variables due to which the model was not able to classify it as a

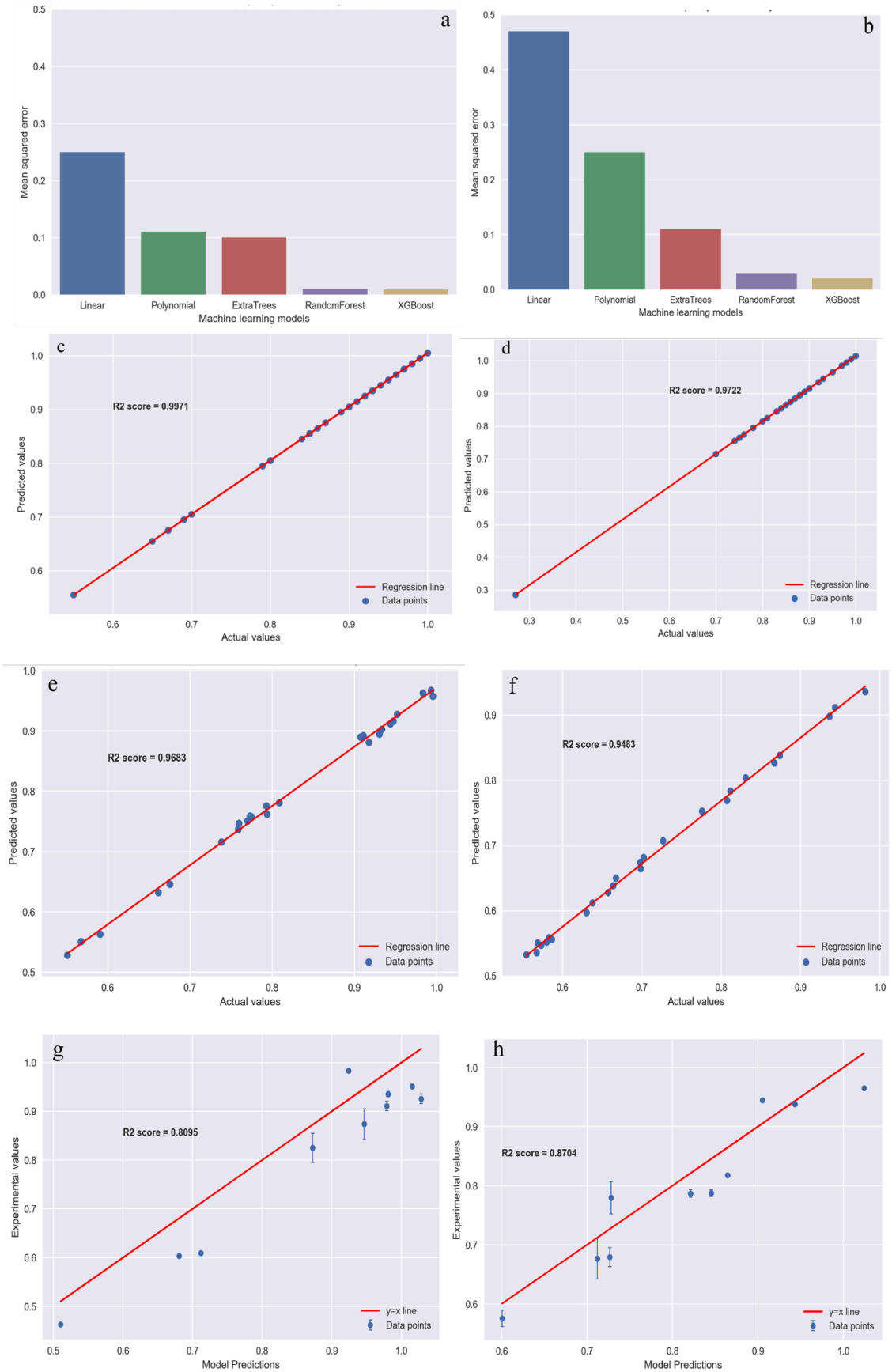


Fig. 2. Various machine learning models developed for a) PO_4^{3-} and b) NH_4^+ recovery as struvite. XGB model validation using correlation coefficient (R^2) of training datasets for c) PO_4^{3-} & d) NH_4^+ recovery and testing datasets for e) PO_4^{3-} & f) NH_4^+ recovery. Experimental validation using correlation coefficient (R^2) for g) PO_4^{3-} & h) NH_4^+ recovery.

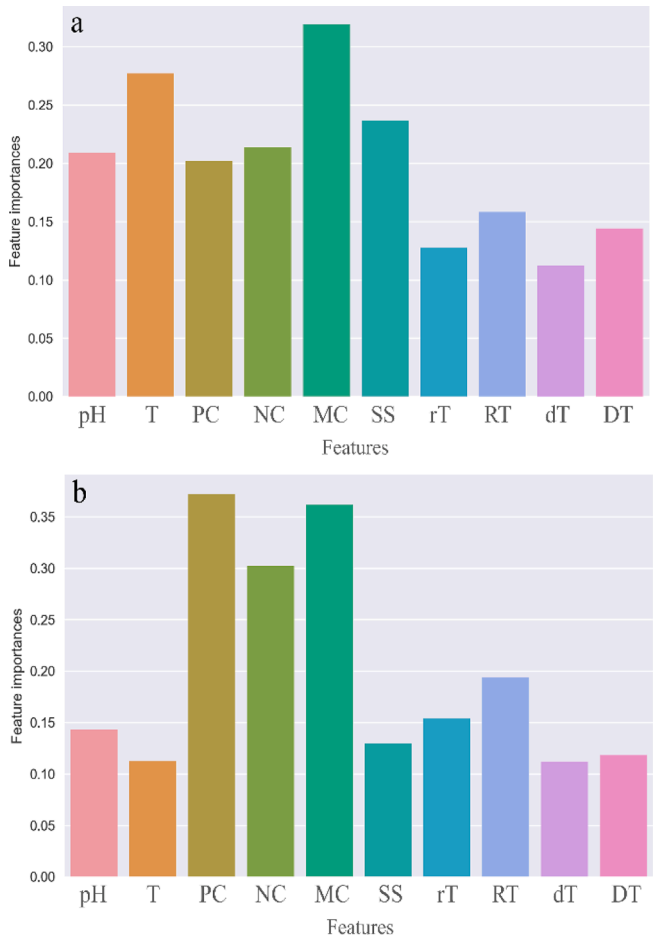


Fig. 3. Feature importance plots for a) phosphate and b) ammonium recovery as struvite developed by XGB model. (T: Temperature; PC: Phosphate concentration; NC: Ammonium concentration; MC: Magnesium concentration; SS: Stirring speed; rT: Reaction time; RT: Retention time; dT: Drying temperature; DT: Drying time).

significant parameter [7,46].

In case of NH_4^+ recovery, the concentrations of equivalent reactive species (PO_4^{3-} and Mg^{2+}) for struvite formation are comparatively more significant in the model. In studies where recovery of NH_4^+ is of major concern, external supply of phosphorus and magnesium in respective molar ratios becomes necessary for struvite recovery. In addition, NH_4^+ concentration is important due to the reason that almost 40 % is lost as ammonia gas and measures to prevent this deed is still under research [18,56]. Hence, presence of a higher concentration of ammonium will guarantee better struvite yield. Further measures on reducing conversion to ammonia or utilization for other purposes will contribute to the economic feasibility of struvite recovery. The influence of drying temperature is the same as in case of PO_4^{3-} recovery; However, the effect of reaction temperature is in the contrary. A better understanding on how these attributes affect recovery efficiencies can be gained in the following section.

Dependence of input parameters on PO_4^{3-} and NH_4^+ recovery

The SHapley Additive exPlanations (SHAP) dependence plots used to study the interaction of important features and effect on the target variables are show in Figs. 4 & 5**.** SHAPley values were mathematically calculated for each feature to analyse how their value affects the prediction. The *x*-axis (horizontal) contains actual values of the input attribute and *y*-axis (vertical) contains the corresponding SHAP values. While each dot represents a dataset, their trend helps gain insights on the model interpretability and the collective SHAP values portray the extent of positive or negative effect, the parameter has on the target variable. Positive SHAP value represents that the input attribute positively favours PO_4^{3-} or NH_4^+ recovery or vice versa. The interpretation of the plots was made based on majority of the datasets (dots) and plots with scarce data points can be ignored as they may not be accurate enough for prediction. Therefore, fruitful explanation from SHAP plots for all the parameters may not be likely. As per the relative importance, the model by default generates a secondary *y*-axis showing another input variable that the parameter in discussion interacts most with. The colour spectrum can be correlated with the magnitude of the secondary variable [35,40]. The pair plot displaying the relationship between each input variable with respect to PO_4^{3-} or NH_4^+ recovery is shown in **Fig. S2 & Fig. S3**.

The SHAP dependence plots for the input variables on PO_4^{3-} recovery are shown in Fig. 4 **(a-j)**. The increase in pH above 9 increases the PO_4^{3-} recovery. pH is one of the most crucial factors for struvite recovery and this effect is shown to be closely correlated with the PO_4^{3-} concentration in wastewater. The reaction temperature at about 25 °C (room temperature) has positive effect on PO_4^{3-} recovery irrespective of pH. PO_4^{3-} concentration in the range of 100–500 mg L^{-1} contributes well to the recovery, alongside NH_4^+ concentration between 100 and 1000 mg L^{-1}. However, PO_4^{3-} recovery decreases at low concentrations of NH_4^+ and PO_4^{3-}.

In case of dependence plot for Mg^{2+} concentration, a linear decreasing trend in PO_4^{3-} recovery with increase in Mg^{2+} concentration can be observed **(**Fig. 4**e)**. This can be attributed to the formation of magnesium salts like $Mg(OH)_2$ instead of struvite at higher dosages [52,61]. The data points for stirring rate were relatively lesser; yet speed between 100 and 200 rpm favoured PO_4^{3-} recovery. Higher reaction times have negative influence on recovery **(**Fig. 4**g),** due to the chances of struvite crystal deposition at an earlier stage. This parameter is said to be highly dependent on the initial concentration of PO_4^{3-}. Retention time required for struvite precipitation contributes to PO_4^{3-} recovery up to 150 min and is found to be correlated with pH in the range of 9.0–10.0. The recovery efficiency decreases linearly with increase in drying temperature. Temperatures beyond 40 °C have been reported to evaporate the water molecules present in struvite, leading to structural instability and formation of other compounds [7,12,46]. In case of drying time, the overall effect seems to be positive **(**Fig. 4**j).**

SHAPley values of pH and reaction temperature can be seen to have a negative influence on NH_4^+ recovery (Fig. 5 **(a-j)**). Though pH and temperature aid in increasing the conversion of urea to ammonium, after reaching a particular level, the ammonium ions in the solution get converted to ammonia, thereby decreasing its availability for recovery [11,56]. The temperature dots are all aggregated in the same region due to insignificant variation between the temperature data points collected from literature **(**Fig. 5**b)**. Lower PO_4^{3-} concentrations have undesirable effect on recovery and vice versa, irrespective of NH_4^+ concentration. This is because, certain amount of PO_4^{3-} would be utilized in the formation of phosphate salts such a calcium phosphates and magnesium phosphates, limiting its presence for ammonium recovery [52,61]. This is similar to SHAPley values of NH_4^+ concentration, where an increase in concentration positively correlates with the recovery **(**Fig. 5**d)**. Addition of magnesium up to 100 mg L^{-1} when the medium PO_4^{3-} concentration is between 100 and 400 mg L^{-1} contributes well to the recovery of ammonium. Beyond the suggested range, there are possibilities for compounds other than struvite to be formed. NH_4^+ recovery efficiency increases with mixing, at any given PO_4^{3-} concentration. With all the reactive ions of struvite being present, stirring can initiate nucleation process within a short time span and assist crystal formation [17,51]. Reaction and retention time have close interaction with NH_4^+ concentration of the medium. Reaction time up to 20 min promotes NH_4^+ recovery, beyond which it might have adverse effects on the molecular bonding and crystal size. However, the retention time has negative influence on the recovery **(**Fig. 5**h)**. Struvite formation is a spontaneous

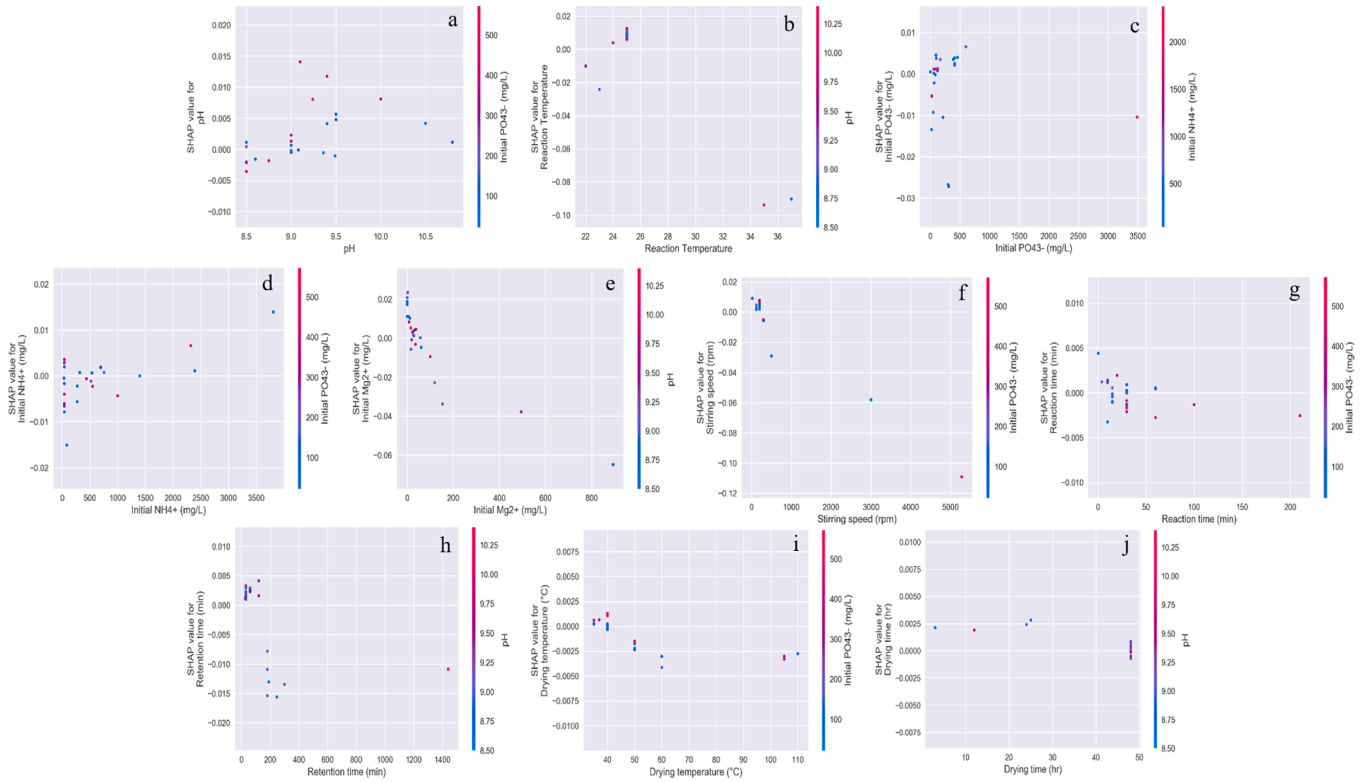


Fig. 4. (a-j): SHapley Additive exPlanations (SHAP) dependence plots for the interactive effect of all the input attributes on PO_4^{3-} recovery.

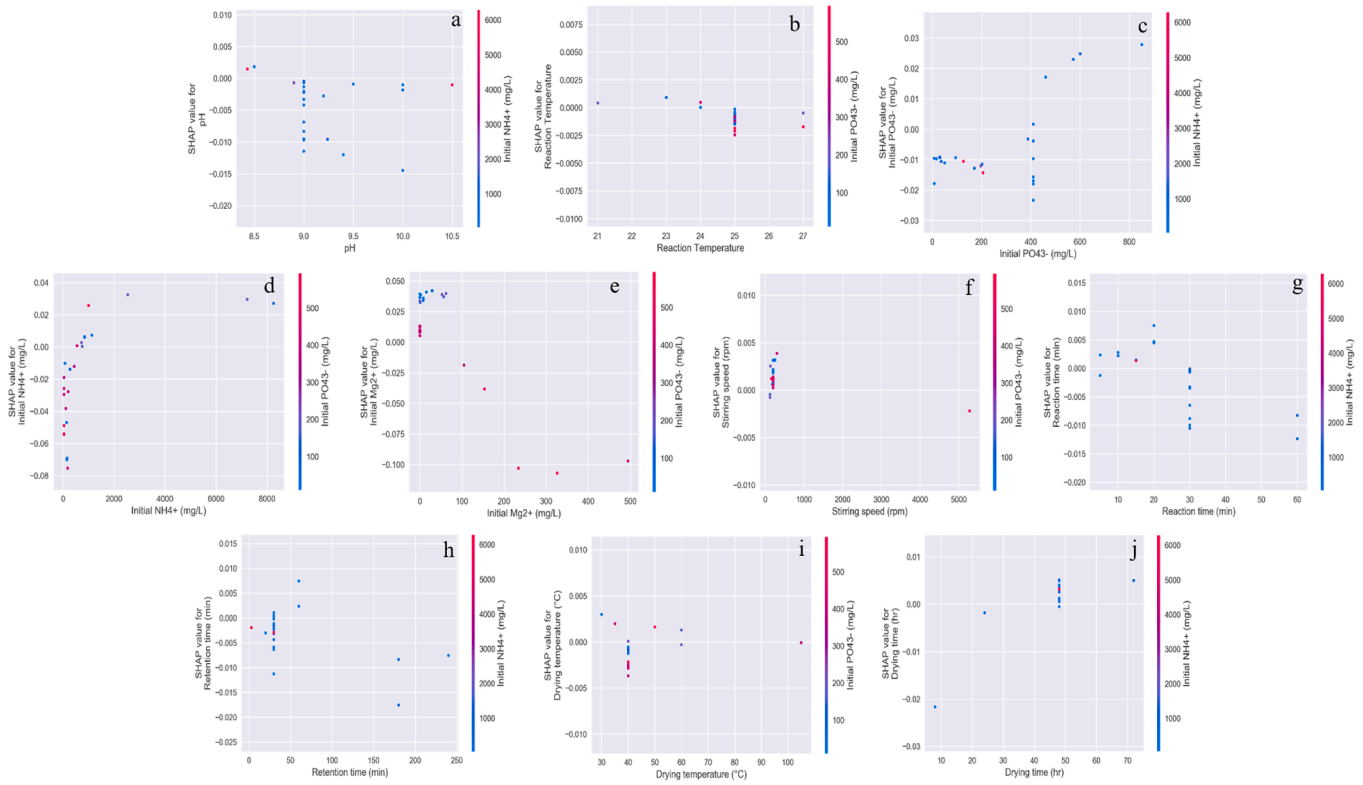


Fig. 5. (a-j): SHapley Additive exPlanations (SHAP) dependence plots for the interactive effect of all the input attributes on NH_4^+ recovery.

reaction and holding up the medium for a long time can lead to ammonia volatilization [56]. While drying time positively impacts NH_4^+ recovery, drying temperature has harmful consequence of the structural integrity of struvite crystals. Almost all the data points at 40 °C negatively affect recovery. However, it is generally considered as an ideal temperature to dry struvite, after which the water molecules in struvite are lost [7,31].

This disagreement can be due to insufficiency in data concerning the drying aspects of struvite. The pair plots showing the relationship between input variables for PO_4^{3-} recovery and NH_4+ recovery as struvite is shown in **Fig S2** and **Fig. S3**.

Summary plots for input parameters on PO_4^{3-} and NH_4^+ recovery

Summary plots are a combined way of visualizing feature importance and SHAP dependence plots. The *y*-axis denotes the list of input variables in descending order based on its importance to the model output. *x*-axis is determined by the SHAPley values of the input parameters. The colour bar indicates the values of features from low (blue) to high (pink). Summary plots convey-four set of information such as feature importance, impact on the prediction efficiency (higher or lower), original value of the varibles (colour bar) and correlation of features based on feature value and x-axis. In general, they provide insights on the relationship between feature values and their influence on the target variable [35,40].

As per the plot, ranked on top is the reaction temperature, the most influential factor governing phosphate recovery (Fig. 6**a**). A low to medium temperature range (20–30 °C) can positively favor recovery. Similarly, low magnesium concentrations serve better recovery efficiencies, further reaffirming the inference from feature importance plots. Recovery of PO_4^{3-} can also be improved with low stirring rates. On the contrary, high pH has positive SHAPley values, thereby can help achieve the required output. Reaction time, drying temperature and time almost have negligible effect on phosphate recovery. In case of ammonium recovery (Fig. 6**b**), magnesium concentration has the most impact and lower concentrations can favor better output. Higher ammonium concentrations have a positive effect of NH_4^+ recovery and vice versa. Similarly, wastewaters having high phosphate concentrations can improve ammonium recovery. Lower PO_4^{3-} concentrations have their SHAP values on the negative side of plot. Contrary to the conditions of PO_4^{3-} recovery, higher pH values have negative impact on NH_4^+ recovery due to enhanced ammonia volatilization [56]. The data points of other operating conditions such as drying temperature, stirring speed and reaction temperature are close to zero and signifies that they have minimal effect on the recovery of ammonium.

It can be observed that, out of ten parameters chosen to predict phosphate recovery, reaction temperature, magnesium concentration, stirring speed, phosphate concentration and retention time are the top five key parameters that should be optimized before struvite precipitatation for better recovery of phosphate. Similarly, in order to obtain better ammonium recovery, concentrations of magnesium, ammonium and phosphate, pH and retention time are the most influential parameters to be considered for optimization and better yield predictions.

Fig. 6. Summary plots for various input variables on a) phosphate and b) ammonium recovery as struvite. (T: Temperature; PC: Phosphate concentration; NC: Ammonium concentration; MC: Magnesium concentration; SS: Stirring speed; rT: Reaction time; RT: Retention time; dT: Drying temperature; DT: Drying time).

Conclusion

Struvite precipitation is a sustainable strategy introduced to handle environmental issues concerning circular bioeconomy. The development of a machine learning model is essential in facilitating proper planning and application of struvite technology in real-time scenario. An XGB model was built using 100 input–output databases from approximately 85 different articles on struvite precipitated from various wastewater sources. Analysis revealed that, out of the ten input variables, reaction temperature and magnesium concentrations are the most important features that demands optimization for enhanced phosphate and ammonium recovery, respectively. Experimental validation showed 80.95 % and 87.04 % similarity between the predicted and experimental values. This model offers deeper insights on understanding the effect of process conditions on struvite crystallization.

Availability of data and material

All data generated or analysed during this study are included in this published article (and its supplementary information files). E-supplementary data for this work can be found in e-version of this paper online.

CRediT authorship contribution statement

Krishnamoorthy Nageshwari: Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Vimaladhasan Senthamizhan:** Methodology, Data curation, Formal analysis, Investigation. **Paramasivan Balasubramanian:** Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition, Resources, Supervision.

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.

Acknowledgement & Funding

The authors thank the Department of Biotechnology and Medical Engineering of National Institute of Technology Rourkela for providing the necessary research facilities. The authors greatly acknowledge the Science and Engineering Research Board (SERB) of the Department of Science and Technology (DST), Government of India (GoI) for sponsoring the PhD fellowship through ASEAN-India Science Technology and Innovation Cooperation [File No. IMRC/AISTDF/CRD/2018/000082] and research grant through Technology Development Programme [File No. DST/TDT/Agro43/2020].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.seta.2022.102608.

References

[1] Acelas NY, Flórez E, López D. Phosphorus recovery through struvite precipitation from wastewater: effect of the competitive ions. Desalin Water Treat. 2015;54(9):2468–79. https://doi.org/10.1080/19443994.2014.902337.
[2] Agarwal AK, Wadhwa S, Chandra S. Diagnosis of tuberculosis–newer tests. J. Association Phys. India 1994;42(8):665.
[3] Peckov A. A machine learning approach to polynomial regression. Work 2012;156.
[4] Ali MI, Schneider PA. A fed-batch design approach of struvite system in controlled supersaturation. Chem Eng Sci 2006;61(12):3951–61. https://doi.org/10.1016/j.ces.2006.01.028.
[5] Almatouq A, Babatunde AO. Concurrent hydrogen production and phosphorus recovery in dual chamber microbial electrolysis cell. Bioresour Technol 2017;237:193–203. https://doi.org/10.1016/j.biortech.2017.02.043.
[6] Barnard J, Meng XL. Applications of multiple imputation in medical studies: From AIDS to NHANES. Stat Methods Med Res 1999;8(1):17–36. https://doi.org/10.1191/096228099666230705.
[7] Bischel HN, Schindelholz S, Schoger M, Decrey L, Buckley CA, Udert KM, et al. Bacteria inactivation during the drying of struvite fertilizers produced from stored urine. Environ Sci Technol 2016;50(23):13013–23. https://doi.org/10.1021/acs.est.6b03555.
[8] Bishop CM, Nasrabadi NM. Pattern recognition and machine learning, 4(4), 738. New York: Springer; 2006.
[9] Chimenos JM, Fernández AI, Villalba G, Segarra M, Urruticoechea A, Artaza B, et al. Removal of ammonium and phosphates from wastewater resulting from the process of cochineal extraction using MgO-containing by-product. Water Res 2003;37(7):1601–7. https://doi.org/10.1016/S0043-1354(02)00526-2.
[10] Corona F, Hidalgo D, Martín-Marroquín JM, Antolín G. Study of the influence of the reaction parameters on nutrients recovering from digestate by struvite crystallisation. Environ Sci Pollut Res 2020;28(19):24362–74.
[11] Darestani M, Haigh V, Couperthwaite SJ, Millar GJ, Nghiem LD. Hollow fibre membrane contactors for ammonia recovery: Current status and future developments. J Environ Chem Eng 2017;5(2):1349–59. https://doi.org/10.1016/j.jece.2017.02.016.
[12] Decrey L, Udert KM, Tilley E, Pecson BM, Kohn T. Fate of the pathogen indicators phage ΦX174 and Ascaris suum eggs during the production of struvite fertilizer from source-separated urine. Water Res 2011;45(16):4960–72. https://doi.org/10.1016/j.watres.2011.06.042.
[13] Dellar M, Topp C, Pardo G, del Prado A, Fitton N, Holmes D, et al. Empirical and dynamic approaches for modelling the yield and N content of European grasslands. Environ Modell Software 2019;122:104562.
[14] Etter B, Tilley E, Khadka R, Udert KM. Low-cost struvite production using source-separated urine in Nepal. Water Res 2011;45(2):852–62. https://doi.org/10.1016/j.watres.2010.10.007.
[15] Fang C, Zhang T, Jiang R, Ohtake H. Phosphate enhance recovery from wastewater by mechanism analysis and optimization of struvite settleability in fluidized bed reactor. Sci Rep 2016;6:1–10. https://doi.org/10.1038/srep32215.
[16] Forrest AL, Fattah KP, Mavinic DS, Koch FA. Application of artificial neural networks to effluent phosphate prediction in struvite recovery. J Environ Eng Sci 2007;6(6):713–25. https://doi.org/10.1139/S07-023.
[17] Galbraith SC, Schneider PA, Flood AE. Model-driven experimental evaluation of struvite nucleation, growth and aggregation kinetics. Water Res 2014;56:122–32. https://doi.org/10.1016/j.watres.2014.03.002.
[18] Gethke K, Herbst H, Montag D, Bruszies D, Pinnekamp J. Phosphorus recovery from human urine. Water Practice and Technology 2006;1(4):1–6. https://doi.org/10.2166/wpt.2006.070.
[19] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In: Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics; 2019. https://doi.org/10.1109/DSAA.2018.00018.
[20] Gkerekos C, Lazakis I, Theotokatos G. Machine learning models for predicting ship main engine Fuel Oil Consumption : A comparative study. Ocean Eng 2019;188:106282. https://doi.org/10.1016/j.oceaneng.2019.106282.
[21] Hu L, Yu J, Luo H, Wang H, Xu P, Zhang Y. Simultaneous recovery of ammonium, potassium and magnesium from produced water by struvite precipitation. Chem Eng J 2019;382:123001. https://doi.org/10.1016/j.cej.2019.123001.
[22] Huang H, Zhang D, Wang W, Li B, Zhao N, Li J, et al. Alleviating Na+ effect on phosphate and potassium recovery from synthetic urine by K-struvite crystallization using different magnesium sources. Sci Total Environ 2019;655:211–9. https://doi.org/10.1016/j.scitotenv.2018.11.259.
[23] Jia G, Zhang H, Krampe J, Muster T, Gao B, Zhu N, et al. Applying a chemical equilibrium model for optimizing struvite precipitation for ammonium recovery from anaerobic digester effluent. J Cleaner Prod 2017;147:297–305. https://doi.org/10.1016/j.jclepro.2017.01.116.
[24] Krishnamoorthy N, Paramasivan B. Evolution of struvite research and the way forward in resource recovery of phosphates through scientometric analysis. J Cleaner Prod 2022;131737.
[25] Krishnamoorthy N, Dey B, Arunachalam T, Paramasivan B. Effect of storage on physicochemical characteristics of urine for phosphate and ammonium recovery as struvite. Int Biodeterior Biodegrad 2020;153:105053. https://doi.org/10.1016/j.ibiod.2020.105053.
[26] Krishnamoorthy N, Dey B, Unpaprom Y, Ramaraj R, Maniam GP, Govindan N, et al. Engineering principles and process designs for phosphorus recovery as struvite: A comprehensive review. *Journal of Environmental*. Chem Eng 2021;9(5):105579.
[27] Krishnamoorthy N, Zaffar A, Arunachalam T, Unpaprom Y, Ramaraj R, Maniam GP, et al. Municipal Wastewater as a Potential Resource for Nutrient Recovery as Struvite. In: Urban Mining for Waste Management and Resource Recovery. CRC Press; 2021. p. 187–215.
[28] Kumari S, Jose S, Tyagi M, Jagadevan S. A holistic and sustainable approach for recovery of phosphorus via struvite crystallization from synthetic distillery wastewater. J Cleaner Prod 2020;254:120037.
[29] Lahr RH, Goetsch HE, Haig SJ, Noe-Hays A, Love NG, Aga DS, et al. Urine Bacterial Community Convergence through Fertilizer Production: Storage, Pasteurization, and Struvite Precipitation. Environ Sci Technol 2016;50(21):11619–26. https://doi.org/10.1021/acs.est.6b02094.
[30] Le Corre KS, Valsami-Jones E, Hobbs P, Parsons SA. Phosphorus recovery from wastewater by struvite crystallization: A review. In *Critical Reviews in Environmental Science and Technology* 2009;39(6):433–77. https://doi.org/10.1080/10643380701640573.
[31] Lind BB, Ban Z, Bydén S. Nutrient recovery from human urine by struvite crystallization with ammonia adsorption on zeolite and wollastonite. Bioresour Technol 2000;73(2):169–74. https://doi.org/10.1016/S0960-8524(99)90157-8.

[32] Liu J, Li Q, Chen W, Yan Y, Qiu Y, Cao T. Remaining useful life prediction of PEMFC based on long short-term memory recurrent neural networks. Int J Hydrogen Energy 2019;44(11):5470–80. https://doi.org/10.1016/j.ijhydene.2018.10.042.
[33] Liu X, Wen G, Hu Z, Wang J. Coupling effects of pH and Mg/P ratio on P recovery from anaerobic digester supernatant by struvite formation. J Cleaner Prod 2018;198:633–41.
[34] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2019). Explainable AI for trees: From local explanations to global understanding. *ArXiv preprint arXiv:1905.04610 2*. https://doi.org/10.1038/s42256-019-0138-9.
[35] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 2017;4766–4775. arXiv preprint arXiv:1705.07874.
[36] Mason L, Baxter J, Bartlett P, Frean M. Boosting algorithms as gradient descent. Advances in Neural Information Processing Systems 2000;12:512–8.
[37] Murphy, K. P. (2012). Machine Learning - A Probabilistic Perspective - Table-of-Contents. *The MIT Press*, 1049.
[38] Nishanth KJ, Ravi V. Probabilistic neural network based categorical data imputation. Neurocomputing 2016;218:17–25.
[39] Nunno L. Stock market price prediction using linear and polynomial regression models. Albuquerque, NM, USA: Computer Science Department, University of New Mexico; 2014. p. 1–6.
[40] Pathy A, Meher S, P, b.. Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. Algal Research 2020;50:102006. https://doi.org/10.1016/j.algal.2020.102006.
[41] Pavlov YL. Random forests. Walter de Gruyter GmbH & Co KG 2019;1–122. https://doi.org/10.1201/9780429469275-8.
[42] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011;12:2825–30.
[43] Prabhu M, Mutnuri S. Cow urine as a potential source for struvite production. International Journal of Recycling of Organic Waste in Agriculture 2014;3(1):1–12. https://doi.org/10.1007/s40093-014-0049-z.
[44] Priyadharshini T, Nageshwari K, Vimaladhasan S, Parag Prakash S, Balasubramanian P. Machine learning prediction of SCOBY cellulose yield from Kombucha tea fermentation. Bioresource Technology Reports 2022;18:101027.
[45] Rahaman MS, Mavinic DS, Meikleham A, Ellis N. Modeling phosphorus removal and recovery from anaerobic digester supernatant through struvite crystallization in a fluidized bed reactor. Water Res 2014;51:1–10. https://doi.org/10.1016/j.watres.2013.11.048.
[46] Schürmann B, Everding W, Montag D, Pinnekamp J. Fate of pharmaceuticals and bacteria in stored urine during precipitation and drying of struvite. Water Sci Technol 2012;65(10):1774–80. https://doi.org/10.2166/wst.2012.041.
[47] Sena M, Seib M, Noguera DR, Hicks A. Environmental impacts of phosphorus recovery through struvite precipitation in wastewater treatment. J Cleaner Prod 2021;280:124222.
[48] Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM. Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. J Chem Inf Model 2016;56(12):2353–60. https://doi.org/10.1021/acs.jcim.6b00591.
[49] Siciliano A, Limonti C, Curcio GM, Molinari R. Advances in struvite precipitation technologies for nutrients removal and recovery from aqueous waste and wastewater. Sustainability 2020;12(18). https://doi.org/10.3390/su12187538.
[50] Song R, Chen S, Deng B, Li L. eXtreme gradient boosting for identifying individual users across different digital devices. In International Conference on Web-Age Information Management 2016;43–54. https://doi.org/10.1007/978-3-319-39937-9.
[51] Tansel B, Lunn G, Monje O. Struvite formation and decomposition characteristics for ammonia and phosphorus recovery: A review of magnesium-ammonia-phosphate interactions. Chemosphere 2018;194:504–14. https://doi.org/10.1016/j.chemosphere.2017.12.004.
[52] Tilley E, Atwater J, Mavinic D. Effects of storage on phosphorus recovery from urine. Environ Technol 2008;29(7):807–16. https://doi.org/10.1080/09593330801987145.
[53] Uysal A, Demir S, Sayilgan E, Eraslan F, Kucukyumuk Z. Optimization of struvite fertilizer formation from baker's yeast wastewater: Growth and nutrition of maize and tomato plants. Environ Sci Pollut Res 2014;21(5):3264–74. https://doi.org/10.1007/s11356-013-2285-6.
[54] Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw 2011;45:1–67.
[55] Warmadewanthi, & Liu, J. C. (2009). Recovery of phosphate and ammonium as struvite from semiconductor wastewater. *Separation and Purification Technology*, *64*(3), 368–373. https://doi.org/10.1016/j.seppur.2008.10.040.
[56] Wu X, Modin O. Ammonium recovery from reject water combined with hydrogen production in a bioelectrochemical reactor. Bioresour Technol 2013;146:530–6. https://doi.org/10.1016/j.biortech.2013.07.130.
[57] Xiao D, Huang H, Jiang Y, Ding L. Recovery of phosphate from the supernatant of activated sludge pretreated by microwave irradiation through chemical precipitation. Environ Sci Pollut Res 2017;24(35):26901–9.
[58] Yee RA, Leifels M, Scott C, Ashbolt NJ, Liu Y. Evaluating microbial and chemical hazards in commercial struvite recovered from wastewater. Environ Sci Technol 2019;53(9):5378–86.
[59] Yuan X, Suvarna M, Low S, Dissanayake PD, Lee KB, Li J, et al. Applied Machine Learning for Prediction of CO2 adsorption on biomass waste-derived porous carbons. Environ Sci Technol 2021;55(17):11925–36.
[60] Zhang K, Zhong S, Zhang H. Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. Environ Sci Technol 2020;54(11):7008–18.
[61] Zhang T, He X, Deng Y, Tsang DCW, Jiang R, Becker GC, et al. Phosphorus recovered from digestate by hydrothermal processes with struvite crystallization and its potential as a fertilizer. Sci Total Environ 2020;698:134240. https://doi.org/10.1016/j.scitotenv.2019.134240.
[62] Zhang Z. Missing data imputation: focusing on single imputation. Annals Of Translational Medicine 2016;4(1):9.