

RESEARCH

Open Access



Machine learning model for the optimization and kinetics of petroleum industry effluent treatment using aluminum sulfate

Victor Ifeanyi Ugonabo¹, Prosper Eguono Ovuoraye^{1,2*} , Ahmad Chowdhury³ and Endrit Fetahi⁴

*Correspondence:
pe.ovuoraye@gmail.com

¹ Department of Chemical Engineering, Nnamdi Azikiwe University, P.M.B. 5025, Awka, Nigeria

² Department of Chemical Engineering, Federal University of Petroleum Resources, P.M.B. 1221, Effurun, Nigeria

³ Imaging Consulting Technical Services, LLC, Dayton, NJ, USA

⁴ Department of Computer Science, University of Prizren, Prizren, Kosovo

Abstract

Small-scale preliminary studies are necessary to determine the feasibility of the machine learning (ML) algorithm and time-evolution kinetics to meet the design specification of the treatment unit. The train and test datasets were obtained from jar test experimentation on the petroleum industry effluent (PIE) sample using aluminum sulfate (AS) as the coagulant. The ML algorithm from scikit-learn was employed to determine the optimum operating condition for the removal of colloidal particles, causing turbidity in the PIE. The predictive capacity of four ML models was compared based on their statistical metrics for clean discharge. The predicted optimum condition corresponds to pH (10), dosage (0.1 g/L), and settling time (30 min) which transcends to residual turbidity ≤ 10 NTU and translates to 95% removal efficiency. The second-order AS-sweep flocculation kinetic showed that at the predicted optimum conditions, modeled rate constant of 1.33×10^{-3} L/g.min and flocculation period of 1.2 min reduced the combination of the monomer, dimmer, and trimmer class colloids from an initial 570 mg/L concentration to the residual counts of 24 mg/L corresponding to residual turbidity ≤ 10 NTU under the mixing regime $14 \text{ s}^{-1} \leq G \leq 164 \text{ s}^{-1}$ satisfied the EPA standard for clean effluent discharge. It incorporated the selected ML output with time-evolution and aggregation kinetics to define sedimentation tank geometry for cleaner discharge. The findings from the design-driven optimization recommended a flow rate ($1000 \text{ m}^3\text{s}^{-1}$), coefficient of kinematic viscosity (0.841 mm/s), and the required detention time (30–60 min) to define the sedimentation tank geometry.

Keywords: Petroleum effluent, ML algorithm, Time-evolution kinetics, Sedimentation tank geometry, Colloidal particles, Coagulation-flocculation

Introduction

Among the challenges facing the human race is the growing demand for energy needed for power generation, automobile energy for transportation, food processing, and electricity. Petroleum serves as feedstock for several consumer goods [1]. The primary source of energy around the globe remains petroleum and natural gas. According to the Environmental Impact Assessment (EIA) reports, petroleum accounts for 90% of the transportation sector's energy consumed in 2020, which amounts to 23 quadrillion BTU [2]. As of 2021, the EIA estimated about 607,000 BPD of refined petroleum is produced

across the US state of Texas at Port Arthur [2, 3]. The oil and gas journal reported that the world's largest refinery in India produces 1.2 million barrels per day of refined petroleum, with about 940,000 BPD and 350,000 BPD being produced across Venezuela and Algeria [2].

In 2020, total US primary energy consumption was equal to about 93 quadrillions BTU as reported in the oil and gas journal publication [2, 3]. A recent survey across sub-Saharan Africa reported evidence of the growing demands, and production of refined petroleum is spurred by the 650,000 BPD target of the Dangote Refinery in Nigeria currently under construction [3]. The adverse effects of the increasing energy demands for petroleum products and natural gas account for the production of high volumes of petroleum industry effluent, posing a severe threat to source water and environmental pollution.

The petroleum refining and processing industry wastewater is vastly one of the major sources of point-blank environmental pollutants across the globe. Petroleum industry effluents (PIE) generated from upstream, midstream, and downstream operations account for the pollution of groundwater. PIE contains heavy metals, the presence of hydrocarbons with high volatility, and concentrations of hazardous polycyclic aromatic compounds [4, 5]. These contaminants have a high tendency to pollute source water, leading to a scarcity of quality water available for domestic use. The contaminants can also render arable land and water bodies unfit for agricultural production. The contaminants such as suspended and dissolved solids are present in the PIE in form of colloidal particles [4]. The colloids are extremely small particles with low sedimentation speed. They can affect water clarity and transfer pollutants such as phenol present in the PIE over long distances into surface water. A high concentration of colloid particles present in discharges can affect soil fertility by depleting acidity or alkalinity [6]. Colloidal particles can also block membranes [7]. Depending on the compositions of the PIE, the colloids can also stick soil surfaces together and cause water bodies to become unfit for aquatic life.

The effective removal of colloidal particles from PIE is important for environmental sustainability. The colloid particles do not settle easily and are difficult to remove until coagulants are added to destabilize the particles in the water [8]. Treatment via coagulation-flocculation is the best approach that allows the tiny colloidal particles to stick together and agglomerate into heavier floc [8–11] that settles easily and can be removed by filters. There are various traditional coagulants have been applied to treat wastewater, which includes sulfates and chloride of aluminum and iron [11, 12]. The finding from the literature shows that the effectiveness of the sources of coagulants is largely dependent on several factors such as pH modification, mixing speed, the configuration of the coagulants, and the physical characteristics of the wastewater [13]. The application of aluminum sulfate to wastewater is dispersed by rapid mixing to remove hydrophilic colloids that have been tested over time [14–16]. Aluminum sulfate is usually offered at a low price per pound [14] and has been proven to be very effective in wastewater treatment operations [16–20]. Various parameters such as the effluent stability and pH balance will determine the achievable quality of the finished water at optimum levels and will be cost-effective [13, 14].

Treatment units must be optimized to reduce costs while increasing efficiency. The optimal value of each experimental factor and their levels of significance on the

wastewater treatment process can be determined. Several methods of optimization of wastewater and effluent treatment processes are available in the literature. This includes the application of analytical techniques, response surface methodology (RSM), and machine learning (ML) algorithms. The traditional methodology of optimizing more than one experimental parameter by taking one factor at a time that is followed by keeping other variables constant is very complicated [21]. The difficulties associated with the application of the analytical method of optimizing the operating conditions in the design process can be overcome by employing more statistical approaches such as machine learning algorithms.

Although the RSM is very instrumental to engineering applications, it is concerned with predictions and studying the relationship between variables to create a better understanding of how the production system works [20]. The RSM has been applied across wastewater treatment operations, modeling, and prediction performances in the design of treatment units [20, 21]. It has been shown successful when incorporated with batch-type laboratory procedures to estimate optimum settling time, dosage control, pH modification, and solid removal rate [20, 22]. One advantage of machine learning (ML) is that it reduces the uncertainty associated with determining optimum features on which kinetics and mechanism can be based [23]. The kinetics, destabilization mechanism [22, 24–26], and optimization of the treatment process are of paramount interest to wastewater researchers and environmental technology experts. The applications of RSM have become increasingly common. The RSM also comes with limited robustness and requires repetitive experimental validation based on the design of experiments (DOE) [23]. In terms of handling large and complex sets of data, the RSM cannot be ascertained with the adequacy of precision [21, 23]. The application of the nonlinear ML model has become imperative to industrial wastewater and effluent treatment processes. Nonlinear regression in machine learning (ML) includes the decision tree (DT) model, support vector machine (SVM), polynomial regression, and random forest (RF) among other models. The application of the ML model is complex, whereby a limited amount of data is available. One advantage of predicting and interpreting optimum conditions in wastewater treatment operations using ML is due to its adequacy of precisions in dealing with large sets of data without having to test reliability using the DOE [23]. The machine learning model and algorithm are designed to investigate the predicted response convergence levels with the observed values obtained from the experimentation. The random forest (RF), decision tree (DT), polynomial model (PM), and support vector machine (SVM) have been reportedly applied in making comparative prediction outcomes in wastewater treatment processes, and their performances were reported to be substantially feasible [23].

Sedimentation basin geometry has been estimated via traditional mathematical modeling and empirical perspectives. In the theoretical computation of sedimentation tank design, particles are assumed to be spherical. The design of the type 2 sedimentation tank takes into account the settling of particles [27] in low concentration with flocculation to increase the strength and mass of floc formed. The addition of chemical coagulant forms and causes heavier flocs to settle rapidly [20]. The settling efficiency of suspended particles is dependent on tank geometry (depth, tank plan, and area), rather than detention time [27–29]. The kinetics parameters and time evolution and aggregation of colloid

particles are consequential to the design of the treatment unit [20]. In recent times, the geometry of the sedimentation tank has been found to depend on temperature [28], initial and final particle concentrations, and flow velocity [27]. The application of design equations based on alternative sedimentation theory [29] enables the design of type 2 sedimentation tank geometry. It involves the application of the conservation law, Newton's second law, and the effect of particle concentration on the sedimentation rate of a suspension of uniform particles [29–31].

In this study, the performances of four machine learning (ML) models for predicting optimum operating conditions (pH, dosage, and settling time) required to minimize the concentration of the colloid particles causing turbidity in petroleum industry effluent (PIE) will be examined. The output of the selected ML will be integrated with the 2nd-order coagulation-flocculation and time-evolution kinetics using the criterion (turbidity ≤ 10 NTU) for cleaner effluent discharge. The kinetic output will be incorporated with the alternative sedimentation theory and adapted for the optimal design of type 2 sedimentation tank geometry. It is believed the selected ML algorithm will serve as a decision-making tool regarding design specifications for environmental sustainability.

Methods

Field sampling and characterization

The authors collected the petroleum industry effluent sample in an attempt to investigate the optimum operating conditions (pH, the dosage of coagulant, and settling time) for the treatment of the petroleum industrial effluent to satisfy discharge standards. We collected the industrial effluent sample from the facility of the Warri Refinery and Petrochemical Company (WRPC), Ekpan, Nigeria. The conventional coagulant used for the PIE treatment consisted of 5 kg of industrial-grade aluminum sulfate (AS) coagulant with the specifications $\text{Al}_2(\text{SO}_4)_3 \cdot 16\text{H}_2\text{O}$, $MW = 666.42$ g/mol. Consisting of 51–60%, Al_2SO_4 and pH (2.5–5.5) were bought from a local vendor Kincel Excel Nigeria Limited [14].

The sampling and preservation of the PIE were carried out following the standard method of examination of wastewater handling reported in the work of Greenberg et al. [32]. The chemical oxygen demand (COD) and biological oxygen demand (BOD) contents in the effluent were examined following ASTM and A.P.H.A method of water examination [32–34].

Data collection

In this research work, the authors used a coagulation-flocculation experimental dataset obtained from a standard nephelometric test procedure for the investigation of the PIE treatment process. The dosage of the coagulant was varied from 0.1 g to 0.5 g/L. The pH of the medium was adjusted from 2 to 10 at 1 atm and 28 at °C. The stirring speed (120 rpm) was administered on the sample for 2 min and reduced to 10 rpm maintained for 20 min. The magnetic stirrer was stopped after 20 min to allow for the studying of settling characteristics of the colloidal particles at varying settling time of 3, 6, 10, 20, and 30 min. The residual turbidity present in the effluent medium after the coagulation-flocculation treatment was measured in the nephelometric turbidity unit and converted to concentrations (mg/L) using a calibration method. The residual turbidity and removal

efficiency were calculated following Eqs. 1–2. The data were stored in a data repository ready for the optimization studies following the ML algorithm shown in Fig. 1.

$$\text{Residual turbidity (NTU)} = N_0 - N_t \tag{1}$$

$$\text{Removal Efficiency(\%)} = \frac{C_0 - C_t}{C_0} \times 100 \tag{2}$$

where N_0 and N_t are the values of the initial and final turbidities in NTU caused by the colloidal particles present in suspension expressed at time t (minutes), C_0 (mg/L) is the initial concentration of the colloid particles present in the industrial effluent before

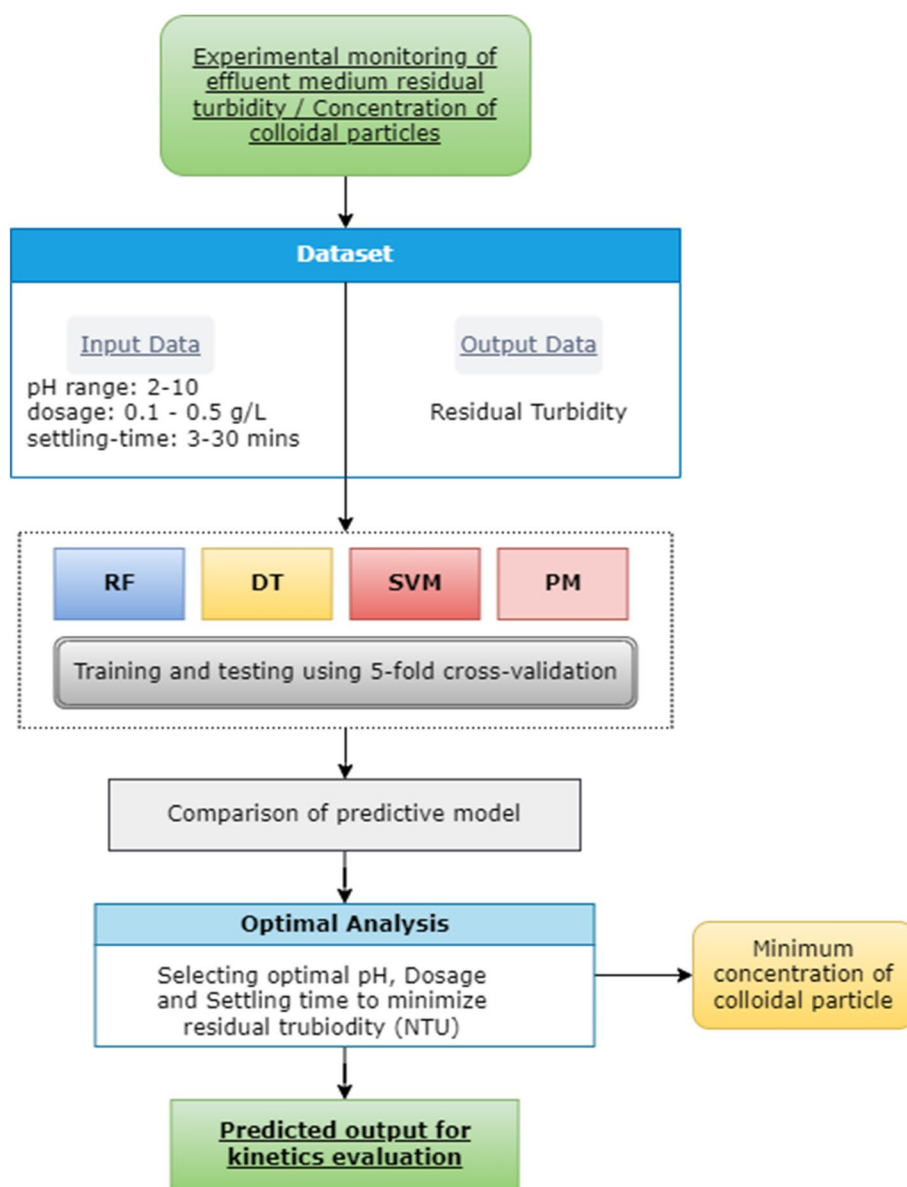


Fig. 1 Sequence of workflow for ML-driven optimization algorithm

treatments, and C_i is the concentration of the i -th numbers of colloidal particles causing turbidity in the effluent in expressed in (mg/L) at time t (minute).

Experimental data training and testing methodology

In this study, the interactions of three variables pH, coagulant dosage, and settling time for the treatment of the PIE were each varied to 5 levels in a total of 125 experimental runs which were stored in a repository with their corresponding residual turbidities, were stored in a repository in the form of a central composite design matrix. The experimental dataset was optimized using several machine learning models that include the random forest (RF), decision tree (DT), support vector machine (SVM), and polynomial regression model (PM). To overcome the over-fitting problem, 20% of the training data were used for validation and 80% for training the model. Additionally, to test the robustness of each model, fivefold cross-validation was used and evaluated using the following performance metrics: mean absolute error (MAE), root-mean-square error (Rmse), adjusted R-square (R^2), and standard deviation (std.). The performances of the models used in the study are shown in Table 1. The scikit-learn library written in Python codes was used to conduct all ML optimizations, and NCSS-PASS software version 20.0 was used for data visualization. In the current study, all ML implementations were executed on a system with a 64-bit Intel Core i5 processor, and 12 GB of RAM. The workflow of the ML optimization process is shown in Fig. 1.

Development of the machine learning algorithm

The machine learning (ML) optimization algorithm was executed using the scikit-learn library in python programming. The sequence of sorting, training, testing, and validation of the datasets obtained from the pilot scale treatment of petroleum industry effluent follows the order described in Fig. 2. In the current research, four (4) ML models were selected based on their proven reliability in the literature. The optimization procedures were implemented via an iterative approach involving a stepwise process of sorting out the experimental data, training and testing datasets, model validation and evaluation using statistical tools. The data train-test procedure for each model follows the steps represented in flowchart Fig. 2. The RF model, DT, PM, and SVM models were executed using the scikit-learn library in python programming by employing a set of parametric codes containing special regressor functions. The normalization, splitting, shuffling, testing, and fitting of the model and final evaluation, validation metrics and discard of

Table 1 Fivefold design developed for ML models

Type	Output shape	Parameter numbers
Dense	(None, 3)	12
Dense	(None, 64)	256
Dense	(None, 128)	8320
Dense	(None, 64)	8256
Dense	(None, 1)	65

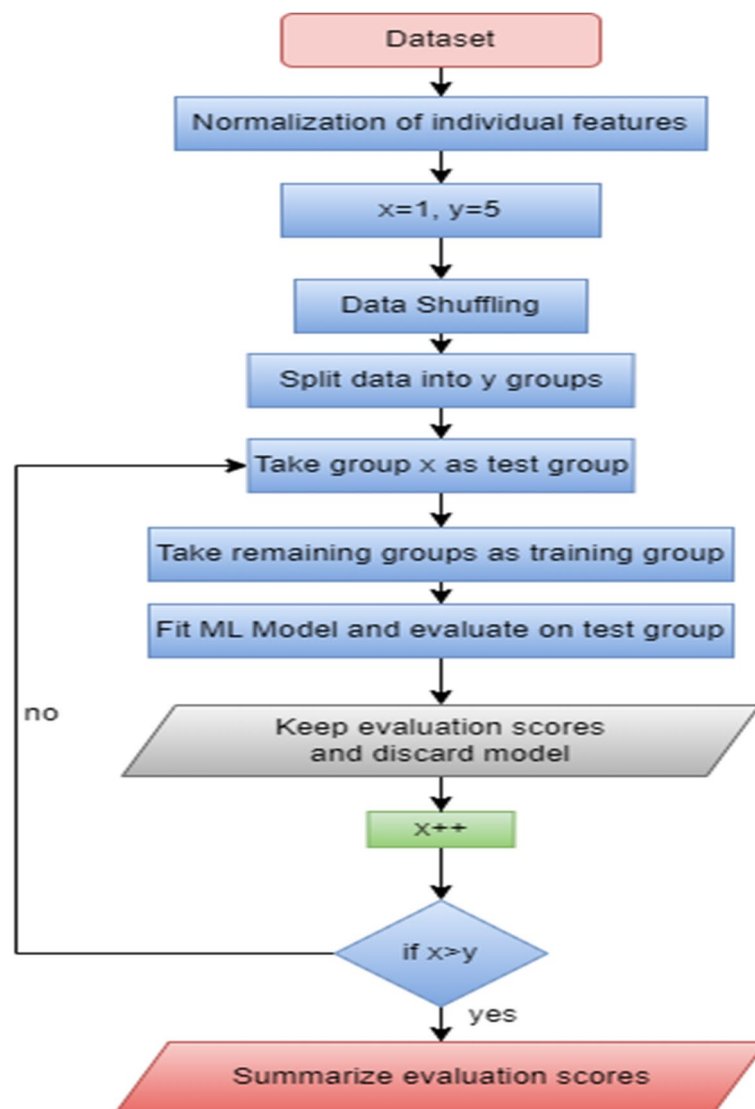


Fig. 2 Schematic of the machine learning model algorithm

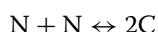
the optimization procedures are summarized in the flowchart in Fig. 1 (“[Experimental data training and testing methodology](#)” section).

The dataset is divided into a y number of sections/folds, with each fold serving as a testing set at some point. In this research work, five fold cross-validation was employed, and the dataset is divided into fivefold ($y=5$) as presented in Table 1. The first fold was used to test the model (20% of a dataset as testing data), while the remaining dataset was used to train it (80% of a dataset as training data) in the first iteration (denoted by x in the flowchart). The second iteration uses the second fold as the testing set and the rest as the training set. This procedure is repeated until each of the fivefold has been utilized as a test set. The predictive outputs across all the selected ML models were compared based on statistical tools and model evaluation metrics (MAE, AAD, RMSE, and adjusted R^2).

where y is the number of folds/section in which the testing data sets were divided to, and x denotes the numbers of iterations used to develop the prediction. The table shows that the ML model implemented with the machine codes in Python (scikit-learn library) was based on fivefold with minimum and maximum parameter number corresponding to 12 and 8320 with an outshape dense (none, 3) and (none, 128) respectively at varying numbers of iterations.

Time-evolution and particle distribution kinetics theory

Assuming N numbers of colloidal particles are present in suspension per cubic meter, the reversible coagulation reaction based on the removal of varying concentrations (C) mg/L of the colloids causing turbidity can be represented by the following:



The removal of colloidal particles causing turbidity from water by coagulation-flocculation has been established to follow 2nd-order rate law [27, 35]. The second-order rate of removal of $2N$ number of particles per cubic meter from the suspension can be described by $r_i = \frac{1}{2} \frac{dC}{dt}$ which is further expressed as follows:

$$\frac{dC^2}{dt} = k_i C^2 \quad (3)$$

Integrating both sides of Eq. 3, to arrive at the following:

$$\frac{1}{C_t} = k_i t + \frac{1}{C_0} \quad (4)$$

Equation 4 can be rearranged to arrive at the following:

$$k_i = \frac{\frac{1}{C_t} - \frac{1}{C_0}}{t} \quad (5)$$

where k_i is the coagulation-flocculation constant at the time (t) expressed in (L/g.min), C_0 is the initial concentrations of the colloidal particles present in the water before treatment, and C_t represents the concentrations of the studied parameter present in PIE at a particular time (t) in minutes. The flocculation period (τ) following the kinetics of the removal of colloidal particles causing turbidity in the PIE can be determined by rearranging Eq. 4 above to arrive at Eq. 6 [24, 26] given by the following:

$$\tau = \frac{1}{C_0 k} \quad (6)$$

Equation 4 can be rewritten to arrive at the following:

$$C_t = \frac{C_0}{1 + \frac{t}{(1/C_0 K)}} \quad (7)$$

Substituting Eq. 6 into Eq. 7 to obtain the following:

$$N_t = \frac{C_0}{1 + \frac{t}{\tau}} \tag{8}$$

For a coag-flocculation period (τ), where the total number of particles concentrated per cubic meter is halved, the average number of colloidal particles present per cubic meter in the petroleum industry effluent at a given time (t). The kinetic expression [22, 24] is given by the following:

$$N_t = \frac{N_0}{2} \text{ and } \tau \rightarrow \frac{t}{2}, \text{ such that } \frac{\tau}{2} = \frac{1}{2C_0K}$$

Equation 8 can be rewritten so that the general expression for the n -th order of the particle aggregation [22, 35] is given by the following:

$$\frac{N_n(t)}{C_0} = \frac{\left(\frac{kC_0t}{2}\right)^{n-1}}{\left(1 + \frac{kC_0t}{2}\right)^{n+1}} \tag{9}$$

For the monomers or singlet particle in water medium ($n=1$), the general expression [22, 24, 35] for the aggregation of the colloidal particles causing turbidity in PIE becomes the following:

$$\frac{N_1(t)}{C_0} = \frac{1}{\left(1 + \frac{kC_0t}{2}\right)^{1+1}} \tag{10}$$

The aggregation of the monomers can be rewritten in terms of C_0 and k_i [22] given by the following:

$$N_1(t) = \frac{C_0}{\left(1 + \frac{kC_0t}{2}\right)^2} \tag{11}$$

Substituting Eq. 6 into Eq. 11, to obtain the following:

$$N_1 = \frac{C_0}{\left(1 + \frac{t}{\tau}\right)^2} \tag{12}$$

For dimmers or doublet particles substitute ($n=2$) into Eq. 9 to obtain the following:

$$N_2(t) = \frac{C_0\left(\frac{kC_0t}{2}\right)^2}{\left(1 + \frac{kC_0t}{2}\right)^3} \tag{13}$$

Substituting Eq. 6 into Eq. 13 to obtain the following:

$$N_2 = \frac{C_0\left(\frac{t}{\tau}\right)^2}{\left(1 + \frac{t}{\tau}\right)^3} \tag{14}$$

Similarly, in terms of trimmers of triplet particles ($n=3$), the aggregation at any given time (t) is given by the following:

$$N_3 = \frac{C_0 \left(\frac{t}{\tau}\right)^3}{\left(1 + \frac{t}{\tau}\right)^4} \quad (15)$$

The overall or sum of all colloidal particles ($n=4$) forming aggregates be it monomers, dimmers, and trimmers can thus be expressed by Eq. 16 [22, 35].

$$\sum N_i = \frac{C_0}{\left(1 + \frac{t}{\tau}\right)^4} \quad (16)$$

The flocculation stirring depth at any applied time was calculated based on the design model of Eq. 16 formulated by Argaman and Kaufman et al. [36] for batch experiments in terms of the stirring (N) rpm. The model indicated the influence of the velocity gradient (G) s^{-1} on hydrodynamics in terms of the mixing speed (rpm) in a substantial flocculation process is expressed in Eq. 17 [37].

$$G = 0.12(S)^{1.54} \quad (17)$$

Empirical theory of sedimentation tank geometry

The optimum flocculation conditions help to understand the unique cause of any sub-optimal performances associated with the treatment process. The authors applied the principle of conservation of mass, the quantity of wastewater containing varying concentrations of solids C (mg/L); following the assumptions made by Sadyrbek et al. [38], the quantity of water entering the system can be expressed as a product of the density and volume of water given by Eq. 18 [31, 38]:

$$m = \Delta\rho \cdot w = (\rho_s - \rho_w)w \quad (18)$$

where ρ_w is density of the heavy particles, ρ_s is the density of the lighter fractions, and w is the volume of the wastewater. Expressing Eq. 18 in terms of change in density and volume of wastewater before and after treatment m_I and m_E to arrive at the following:

$$m_I = (\rho_f - \rho_I)w \quad (19)$$

$$m_E = (\rho_f - \rho_E)w \quad (20)$$

where ρ_f is the final density of the water after treatment, ρ_i is the initial density of water before treatment, and w is the volume of the wastewater [35]. Modifying the model terms in Eqs. 19 and 20, by taking $C_I = m_I/\rho_I$ and $C_E = m_E/\rho_E$ and expressing the density parameters in Eqs. 19 and 20 in terms of concentrations of influent C_I (mg/L) and effluent C_E (mg/L), to arrive at the following:

$$\rho_I = (\rho + C_I) \quad (21)$$

$$\rho_E = (\rho + C_E) \quad (22)$$

Substituting Eqs. 21 and 22 into Eq. 18 to arrive at the following:

$$m_I = (\rho + C_I) - (\rho + C_E) \tag{23}$$

$$m_E = [(\rho + C_I) - (\rho + C_E)]w \tag{24}$$

Expanding the brackets, to arrive at the following:

$$m_E = (C_I + C_E)w \tag{25}$$

Consequently, design equation for the sedimentation basin sludge height [38] becomes the following:

$$H_0 = \sqrt{\frac{v \cdot t}{\ln\left(\frac{C_I}{C_I - C_E}\right)}} \tag{26}$$

The Eq. 26 can be rewritten in terms of sedimentation tank sludge depth H_0 , influent and effluent concentrations C_I and C_E , kinematic viscosity (ν), and theoretical settling time (t). The sedimentation depth is divided into two zones, the settling zone and the sludge zone. Sadyrbek et al. [38] gave the thickness of the boundary for laminar flow regime is given by $2\sqrt{\frac{\nu \cdot L}{\mu}}$, such that movement around the thin plate can be represented by Eq. 27 given by the following:

$$\delta = 5.2 * \sqrt{\frac{\nu \cdot L}{\mu * \varnothing}} \tag{27}$$

The height of the settling zone H_s becomes the following:

$$H_s = 5.2 \left(\frac{\nu \cdot L}{\mu \cdot \ln\left(\frac{C_I}{C_I - C_E}\right)} \right)^{0.5} \tag{28}$$

Sedimentation depth depends on flow velocity, assuming the tank is divided into two zones, H_s settling zone and sludge zone H_0 . To determine the depth of the sedimentation zone H , we use the following:

$$H = H_s + H_0 \tag{29}$$

The height of settling zone H_s in practice can be obtained as a function of the sedimentation depth in accordance to the research work of Sadyberk et al. [38] given by the following:

$$H_s = 0.12H_0 \tag{30}$$

The width of the sedimentation tank was found to depend on the height of the tank and flow through the tank [31, 38] given by the following:

$$W = \frac{Q}{3.6\nu} \tag{31}$$

Expressing length of the tank (L) in terms of settling depth H (m), flow velocity U (m/s), and concentration profile, to arrive at the following:

$$L = \frac{H^2U}{27.04v} \ln\left(\frac{C_I}{C_1 - C_E}\right) \quad (32)$$

Equation 32 can be used to establish the sedimentation tank length.

Results and discussion

Characteristics of the effluent

The analysis of the PIE characteristics presented in Table 2 shows that the total suspended solids and dissolved solid content present in the industrial effluent correspond to 60 mg/L and 10 mg/L. The total suspended solid (TSS) content is > 30 mg/L, indicating the effluent does not satisfy the criterion for industrial effluent discharge [22, 39, 40]. This characteristic suggests the presence of residual contamination of the effluent. The turbidity composition of the PIE was 220 NTU, corresponding to 517 mg/L in concentrations of colloidal particles. The value of the turbidity is > 10 NTU, suggesting the PIE does not satisfy the EPA turbidity standard for clean effluent discharge in Ovuoraye et al. (2021). Consequently, the authors reasoned that the concentration level of the colloid particles causing turbidity in the PIE does not satisfy the baseline for residual colloid concentration required for environmental sustainability. The calculated COD-BOD ratio of the PIE is equal to 18. This proportion is greater than 3.5, suggesting that the inert fraction of the PIE is prevalent, eliminating the option for a biological method of

Table 2 Characteristics of the petroleum industrial effluent

Wastewater parameters	Reference	Values
pH	APHA 4500-H	7.08
Turbidity	ASTM-D1889	220 NTU
Total dissolved solids	ASTM D-1888	10 mg/L
Total suspended solids	APHA 2540-C	60 mg/L
Chemical oxygen demand (BOD)	APHA 5210-B	266 mg/L
Biological oxygen demand (COD)	APHA 5220-D	< 14.3 mg/L
Electrical conductivity, s/cm	APHA 2510-C	Nil
Zinc	ASTM-D4691	0.45 mg/L
Calcium	ASTM-D-3511-09	25.64 mg/L
Silicon	ASTM-D4691	1.87 ppm
Ammonia	ASTM-D1426-03	3.92 mg/L
Arsenic	ASTM-D4691	4.82 mg/L
Aluminum	ASTM-D4691	9.72 mg/L
Iron	ASTM-D4691	1.15 mg/L
Mercury	ASTM-D3223-17	1.50 mg/L
Lead	ASTM-D4691	0.63 ppm
Nitrate	ASTM-D3867-09	5.94 mg/L
Phenol	ASTM-D258-06	0.59 mg/L
Nickel	ASTM-D4691	0.02 ppm
Sulfate	A.P.H.A-4500-SO42	287.70 mg/L
Phosphate	A.P.H.A-4500P	0.01 mg/L
Phosphorus	ASTM-D515	2.80 mg/L
Copper	ASTM-D1688-12	0.04 ppm
Sodium	A.P.H.A-3500	0.13 mg/L

treatment [41]. The finding supports the fact that the coagulation-flocculation treatment method is sufficient for treating PIE. The concentration of phenol is ≤ 0.59 mg/L. It can also be observed from the effluent characteristics presented in Table 1 that, at the initial pH of 7.04, the concentrations of ammonia, sulfate, and nitrate content are ≤ 28 mg/L, with calcium, iron, arsenic, mercury, nickel, and lead concentrations ≤ 5 mg/L. It can be concluded from the PIE characterization result that the medium contains predominantly heavy metals. This outcome confirmed the presence of hydrophilic colloids causing turbidity in the PIE, as a significant factor that affects water clarity [9, 14] with a strong tendency to pose difficulty to unit operations [8, 20]. It can also be inferred from the effluent characteristics that further treatment is needed to be administered to improve upon the discharge quality for cleaner production that guarantees environmental sustainability.

Machine learning optimization and evaluation metrics

The optimization of the aluminum sulfate-driven coagulation-flocculation treatment of the petroleum industry effluent (PIE) was executed via the application of a machine learning (ML) algorithm using the scikit-learn library in the python program. In this study, a polynomial regression model (PM), random forest (RF), decision tree (DT), and support vector machine (SVM) were implemented for the determination of the optimum operating conditions for the removal of colloidal particles precipitating turbidity in the PIE. The ML model and algorithm were applied to minimize the residual concentration of the colloids, causing turbidity in the effluent. The selected ML result was adopted considering its proven flexibility and robust application to reducing the uncertainty associated with determining optimum features on large datasets [23, 42, 43] from which kinetic parameters and design can be based. The statistical metrics across the selected models shown in Table 3 were adopted as a criterion to describe the reliability of the ML algorithm. The assessment of the ML performance output was based on the minimal qualities of the validation metrics identified with their prediction outputs.

Table 3 shows the results obtained across the ML algorithm in terms of the statistical metrics (adjusted R^2 , Rmse, AAD, and MAE values). The results from the evaluation metrics proved that the SVM yielded $MAE=55$, with an AAD value corresponding to 77.30. The output from this model indicated a very low adequacy of precision, resulting in a high Rmse value of 90.72. The findings suggest that the SVM assumptions are probably violated due to noise [42]. Consequently, the model may require a large amount of training and testing datasets to increase the adequacy of the signal-to-noise ratio, thereby reducing the associated model constraint [43]. The statistics obtained from the PM correspond to $MAE=44.67$, with an AAD of 65 and Rmse equal to 44. The model

Table 3 ML model statistical evaluations using fivefold cross-validation

fivefold validation metrics					
ML model	MAE	RMSE	pred- R^2	AAD	Std. dev
Random forest regression (RF)	26.21	40.92	0.90	27.10	40.97
Decision tree regression (DT)	26.23	41.15	0.81	29.10	41.11
Support vector machine (SVM)	55.10	77.37	0.33	72.29	77.42
Polynomial regression (PM)	44.68	60.89	0.58	50.97	60.92

statistical output indicated that the error associated with the PM metrics is lower than that of the SVM. The output indicates a slight deviation of the predicted values from the actual observation obtained from the experimentation, although the PM yielded lower statistical error metrics compared to the SVM. The findings also indicated that the model assumptions are associated with a lower confidence interval [23, 43]. Hence, the predicted outputs are less significant and require some degree of stability [42]. To obtain better model stability, the adequacy of precision which measures the signal-to-noise ratio of the model [20] can be improved upon by increasing the magnitude of training datasets, thereby reducing constraints and complexity associated with the lack of fit [23, 42]. The outputs from the SVM and PM performance indicated a poor correlation of their predicted output with the actual observation practicable. Consequently, optimization analysis based on the SVM and PM will be ignored due to the low reliability resulting from their constraint variance.

However, the DT model recorded an MAE of 26.23 and an Rmse of 41.10. The adjusted R^2 of the DT corresponds to 0.81. The evaluation metrics of the DT model indicated a better performance output than the SVM and PM. The MAE value equal to 26.21, and an Rmse of 40.90 was recorded for the RF model with an adjusted R^2 corresponding to 0.90. It can be inferred from the outputs of the DT and RF model that the statistical metrics associated with both ML were ≤ 42 indicating more reliable adequacy of precision, and a higher confidence level is associated with the performance output of both ML [23, 43, 44].

A comparative analysis of the statistical metrics across the selected machine learning models shows that lower Rmse, AAD, and MAE outputs correspond to better performance of the ML algorithm [23, 44]. Lower values of the DT and RF model statistical metrics can be attributed to a higher degree of precision [42, 44] confirming there is a probability of a higher confidence level associated with the RF and DT model assumption [20, 42, 43]. The ML evaluation metrics proved that the DT and RF model fits the ML optimization algorithm compared to SVM and PM. It can be concluded from the model summary and statistical outputs presented in Table 3 that the performance of the RF model yielded the most significant prediction outputs with minimal errors and improved validation metrics. The reliability of the RF model performance is consistent with the findings reports from the research work of Dong et al. [44] and Hong et al. [43].

Optimization process and model performance validation

The objective of the ML optimization is to minimize the residual colloidal particles causing turbidity in the PIE at the predicted optimum conditions (pH, dosage, and settling time). The validation of the ML model performance was expressed in terms of the standard deviation (std) of the predicted residual concentration from the actual observation practicable.

Table 4 shows the variation of the predicted outputs from the ML algorithm with the actual observation obtained from experiments. The results proved that the performances of the selected machine learning vary intermittently across the predictive capacities of the models implemented with the ML algorithm shown in (Fig. 2) in “[Development of the machine learning algorithm](#)” section.

Table 4 ML model performance in terms of predicted optimum

ML model	Dosage (g/L)	pH	Settling time (Minutes)	Residual conc (mg/L)	Residual turbidity (NTU)
Actual observation	0.1	10	30	21	9
Support vector machine	0.1	10	30	374	159
Random forest	0.1	10	30	24	10
Decision tree	0.1	10	30	33	14
Polynomial regression	0.1	10	30	101	43

It can be observed from Table 4 that the validation of the predicted optimum conditions across the selected ML models was consistent with experimentally determined optimum values following the jar test procedure. The experimentally determined optimum operating conditions correspond to pH of 10, the dosage of 0.1 g/L, and a settling time of 30 min, respectively, while the actual observation translates to residual turbidity of 9 NTU. The results presented in Table 4 confirmed the predicted residual turbidity (NTU) and colloid concentration (mg/L) obtained at the optimum conditions vary across selected models implemented with the ML algorithm. The authors reasoned that the disparity in the predicted residuals across the ML might be due to the relative differences observed across the model evaluation statistics.

The plots in Fig. 3 were drawn to show the distribution of the predicted outputs of the ML versus actual observation obtained from coagulation-flocculation experimentation. Figure 3 a–d shows the correlation between the predictive performances of the selected ML model versus the actual observation practicable.

Figure 3 a–b represents the predictive outcome of the RF and DT model. The outline of the plots indicated a normal distribution of the predicted outputs along the regression line. The findings confirmed that the outputs from the RF and DT models correlated with the actual observation practicable. The results transcend to adjusted R^2 values of 0.90 and 0.80, respectively. The performance of the models confirmed that a minimal magnitude of statistical errors and standard deviation is associated with the model probability [20, 23]. The predictive capacities of the RF and DT model recorded residual turbidity of 10 NTU and 14 NTU, respectively. The predicted outputs transcend to modeled concentrations of 24 mg/L and 33 mg/L of residual colloid particles in the PIE after the coagulation-flocculation treatment. The modeled performances translate to the removal efficiency of 95%, and 93%, corresponding to standard deviations of 40.97 and 41.11 for the RF and DT, respectively.

Figure 3 c–d shows the correlations of the predicted outputs of the PM and SVM versus the actual observation obtained from experimentation. The outline of Fig. 3 c–d indicated a weak correlation between the actual observations practicable and predicted outputs obtained from the PM and SVM. The plots in Fig. 3 c–d indicated a considerable deviation in their predicted residuals from the normal distribution line. The model output corresponds to low adjusted R^2 of 0.58 and 0.33 for the SVM and PM, respectively. The PM and SVM performance outputs translate to predicted residual turbidities of 43 NTU and 159 NTU, respectively, with a corresponding modeled concentration of 101 mg/L and 374 mg/L of the colloidal particle at predicted

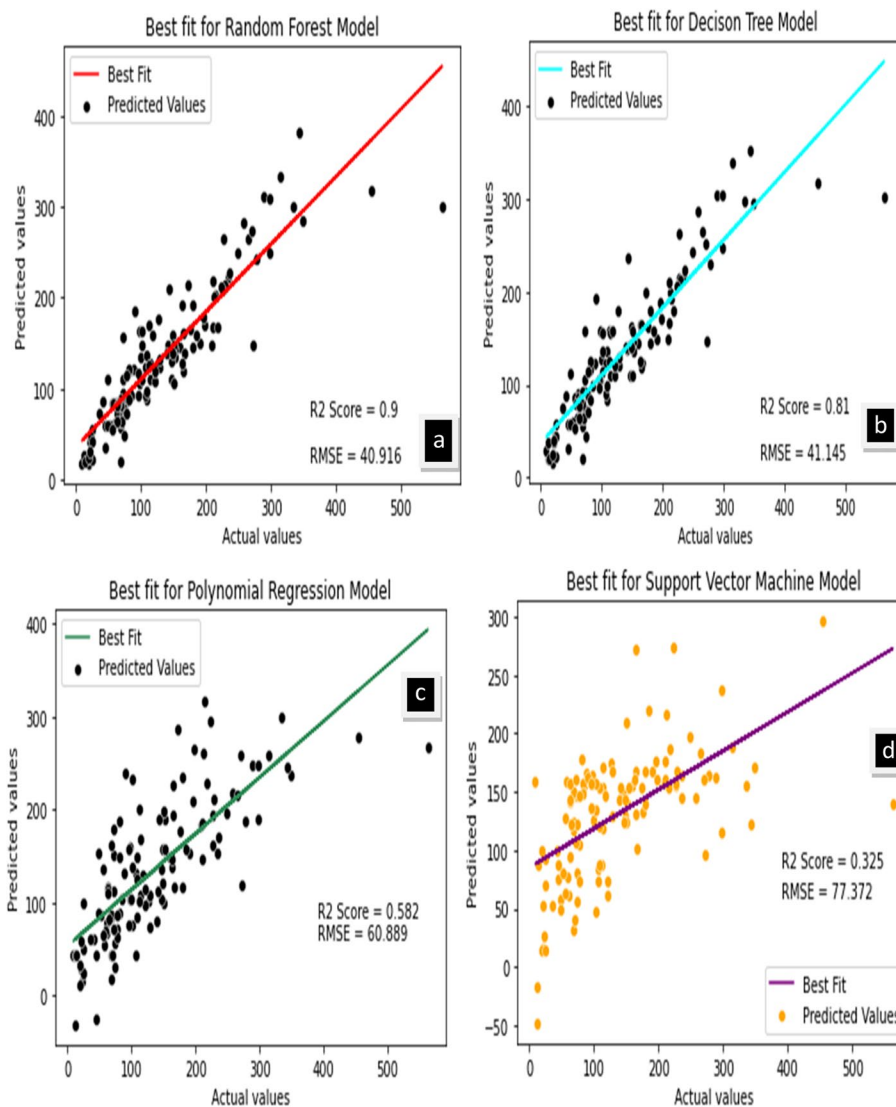


Fig. 3 **a** Comparisons between experimental based output (actual) versus random forest (RF) model predictions, **b** comparisons between experimental-based output (actual) versus decision tree (DT) model, **c** polynomial regression model (PM), and **d** support vector machine (SVM)

optimum conditions pH (10), dosage (0.1 g/L), and settling time (30 min). The predicted residuals of the PM and SVM transcend to a removal efficiency of 54% and 45% indicating standard deviations of 60.92 and 72.42 from the actual observation practicable. The results proved that the PM and SVM performance yielded relatively low removal efficiencies and translates to residual turbidity ≥ 40 NTU. The outcome suggests the predicted outputs do not align well with the actual observation, resulting from the low degree of stability. Consequently, the model performance does not guarantee residual concentrations of colloidal particles for clean effluent discharge.

Analysis of Fig. 4 shows the hierarchy-based overall performance evaluation of the selected ML model applied to minimize residual turbidity following the treatment of the petroleum industry effluent (PIE). It can be observed from Fig. 4 that the

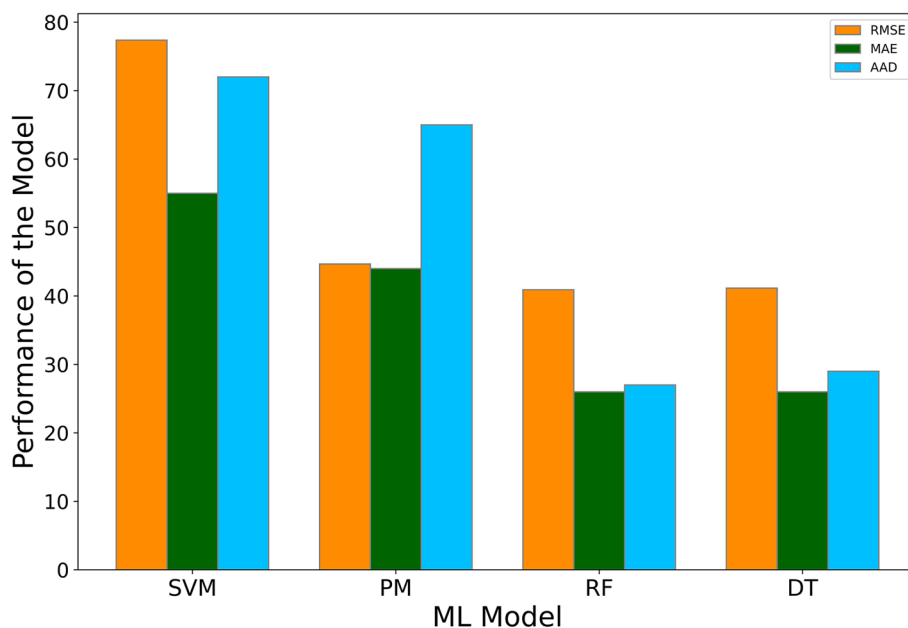


Fig. 4 Statistical evaluation metrics across the selected ML algorithm

predictive capacities across the selected ML algorithm used for the interpretation of the AS-driven coagulation-flocculation treatments of the PIE in terms of their performances and evaluation statistics follow this respective order:

RF > DT > PM > SVM

The criterion used for the ML optimization output was based on the maximum residual turbidity (≤ 10 NTU) recommended by the environmental protection agency (EPA) for clean water discharge [24]. The results based on the selected criterion showed that the SVM (374 mg/L, 43 NTU), PM (101 mg/L, 43 NTU), and DT (33 mg/L, 14 NTU) do not satisfy the EPA standard for industrial effluent discharge. The RF model yielded the best performance with a minimum residual that satisfied the EPA criterion (turbidity ≤ 10 NTU) for clean discharge. Consequently, the authors reasoned that the predicted residual concentration of 24 mg/L obtained from the RF model is substantially feasible to achieve environmental sustainability.

However, the predicted optimum pH of 10 is consistent with the results reported with the application of AS in the treatment of cosmetics wastewater [14]. The maximum removal performance transcends to the reduction of the turbidity present in PIE from an initial 220 NTU to a residual of turbidity ≤ 10 NTU. The removal efficiency transcends 210 NTU of turbidity removed from the PIE. This optimum removal (210 NTU) recorded in the present study is consistent with the results (200 NTU) removed from brewery industry wastewater using aluminum sulfate reported in previous research work [13]. The predicted optimum dosage (0.1 g/L) is considerably lesser than 0.5 g/L, suggesting the finished effluent is stable, with a negligible tendency to blind filters [26, 45]. The optimum dosage recorded in the present study is consistently lesser than 2000 mg/L reported by Elsayed et al. [46]. The optimization results translate to a reduction of the colloidal particles present in PIE from an initial concentration of 520 mg/L to a residual

of 24 mg/L at the predicted operating conditions. There is currently no threshold specified for colloid counts’ industrial discharge. The authors reasoned that the value of the residual turbidity is ≤ 10 NTU is the maximum recommended by the EPA [39] for environmentally viable effluent discharge. The predicted residual (24 mg/L) corresponding to 10 NTU recorded at optimum dosage of 0.1 g/L indicates that the predicted colloid concentration for the finished PIE is feasible for clean water discharge without compromising coagulant wastage and environmental sustainability. The predicted pH of 10 is prevalently alkaline [14, 20], and floc reached optimum size [13, 18, 37]. The optimum settling time (30 min) was consistent with the results reported by Nomantodazo et al. [15] and Ovuoraye et al. [20].

Coagulation-flocculation kinetics evaluation and validation of ML model performance

Microscale kinetics model validation proposed for the removal of colloidal particles from suspension by coagulation was based on the perikinetic model [20, 22, 24]. The values of the $C_0 = 520$ mg/L, and predicted values of residual concentration (C_t), at optimum settling time $t = 30$ min were obtained from the ML optimization result presented in Table 3 (see “Optimization process and model performance validation” above) and were substituted into Eq. 5 to obtain values for rate constant k_t (L/g.min).

The aluminum sulfate-driven coagulation-flocculation outcome based on the selected ML performance is shown in Fig. 5. The kinetics results confirmed that, at the second-order coagulation-flocculation rate, constant k_t increased as the residual colloid concentrations after the treatment decreased. The rate constant $k = 1.33 \times 10^{-3}$ L/g.min and 4.32×10^{-3} L/g.min were recorded for the RF (*std.* = 27.10) and DT (*std.* = 29.00), while rate constants and 9.52×10^{-4} L/g.min and $k = 5.8 \times 10^{-5}$ L/g.min were recorded for the PM (60.90) and SVM (*std.* = 77.42), respectively. The rate constant decreased consistently from $k = 1.33 \times 10^{-3}$ L/g.min to $k = 5.8 \times 10^{-5}$ L/g.min as the AAD metrics increased from 27 to 120. The outcome indicates that the higher the rate constant, the more stability of the finished water to meet the requirement for effluent discharge. The

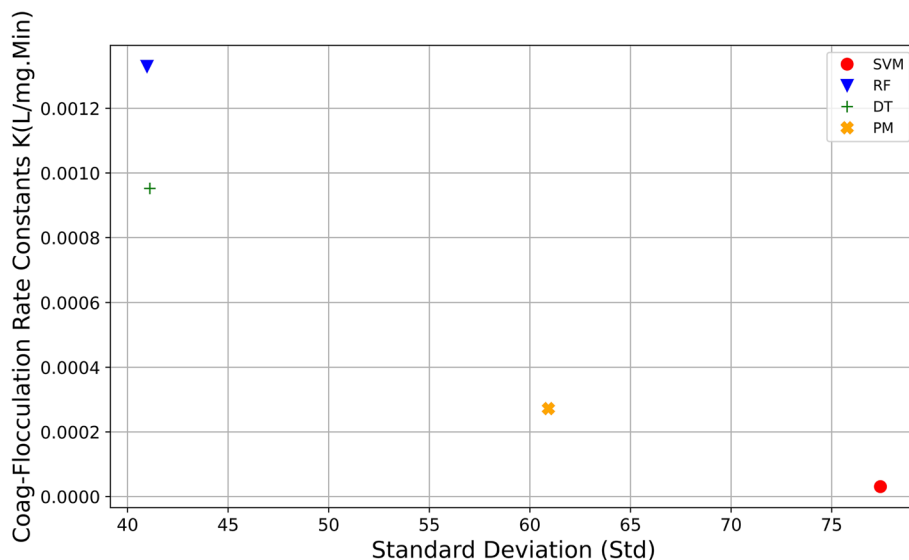


Fig. 5 Modeled coagulation-flocculation rate constant across selected ML

maximum value of $k = 1.33 \times 10^{-3}$ L/g.min was obtained at the minimum concentration of 24 mg/L, corresponding to 10 NTU residual turbidity.

The authors reasoned that the higher value of rate constants conforms to the better performance of the ML model. The best performance was reported for the RF model corresponding to residual 24 mg/L. The least value of $k = 5.8 \times 10^{-5}$ L/g.min was obtained for the SVM. The output of the RF proved that 0.1 g/L of the aluminum sulfate (AS) aligns well with the colloidal particles present in the water [14, 20], inducing the very tiny colloids to floc into aggregates. At a dosage > 0.1 g/L, there is a tendency for AS to cause scaling in the medium [14, 45]. The findings from the predictive capacity of the RF model suggest superb aggregation, and mild floc breakage occurred [13, 20] leading to a low residual ≤ 10 NTU, which accounted for the high removal efficiency, translating to 95% efficiency of removal. The findings proved hierarchy of the performance output of the ML algorithm is consistent with the kinetics results.

Colloidal particle aggregation kinetics and distribution analysis

The optimization and kinetics reports confirmed the outputs of the RF model as most significant for the design specification. The influence of mixing intensity (G) is a key factor in the acknowledgment of the design [36, 37, 45]. The implications of G cut across low sensitivity to operational upsets and eventually its contributions to the disappearance of the primary particles [37]. Investigating the effect of shear and mixing intensity on the aggregation of the colloidal particles that transcend to the removal efficiency $\geq 90\%$ and minima residual ≤ 10 NTU, is largely dependent on the understanding of the hydrodynamics of the floc formation and aggregation process. The values of G (s^{-1}) were evaluated by substituting the values of the slow and high stirring speeds (rpm) used for the coagulation experiment into model Eq. 17.

The results obtained showed that slow mixing (10 rpm) corresponds to $G = 4.16 s^{-1}$ and $167 s^{-1}$ for high stirring (120 rpm). The range of values of G has a direct bearing on the design of the clarifier/flocculation unit [20, 45]. The analysis of the range of values of G confirmed that the degree of mixing effect for the removal of the colloidal particles laid in the regime $4.16 s^{-1} \leq G \leq 167 s^{-1}$. The outcome suggests the velocity gradient falls into the moderate mixing regime [45], indicating that the floc formed was heavy [37, 45], and the flocculation process was flexible enough to keep the floc out of suspension [9, 13]. The optimization output from the RF algorithm indicated the mixing regime was effective in the removal of colloidal particles from the PIE [48]. This outcome suggests that the mixing promoted floc formation, which translates to residual turbidity (≤ 10 NTU) and corresponds to 95% removal efficiency due to a small shear force [22, 24]. The reduced shearing of the floc probably resulted in optimum coagulation [20, 24] yielding 24 mg/L of residual colloids.

However, the time evolution and aggregation distribution [22, 24] of the concentration number of colloidal particles causing turbidity in the PIE were expressed in terms of monomers, dimmers, trimmers, and the cumulative of all colloidal particles per cubic meter [22]. The values of the corresponding concentration number of the colloidal particles causing turbidity in the industrial effluent were derived following the analysis of Eqs. 11, 12, 13, 14, 15. The value of flocculation period (0.83 min) was evaluated from Eq. 6. The concentration number of colloidal aggregates [49] (N_1 , N_2 , N_3 , and ΣNi)

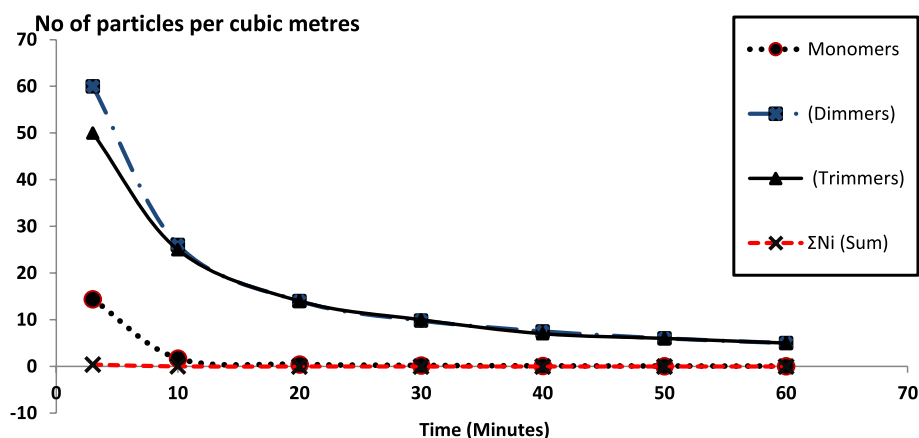


Fig. 6 Time-evolution aggregation and distribution of colloidal particles at predicted optimum

formed per cubic meter [22] was evaluated by substituting the values of $C_0 = 520$ mg/L and $k_f = 3.2 \times 10^{-3}$ L/mg.min obtained from the predicted rate constant of the RF algorithm into Eqs. 12, 13, 14, 15, 16.

The time evolution and distribution of the colloidal particle per cubic meter under the influence of the mixing regime ($4.16 \text{ s}^{-1} \leq G \leq 167 \text{ s}^{-1}$) is shown in Fig. 6. The trend of the distribution curves in Fig. 6 showed that the monomer, dimer, and trimer class colloid particles floc into aggregates rapidly as the flocculation period increased beyond 10 min. The curvatures of the red and black curves indicated the cumulative number of particles (ΣN), and the monomers class colloid particles decreased rapidly at the flocculation period of ≤ 10 min. This outcome indicated the good overall performance of the AS sweep-flocculation kinetics [24]. The AS coagulant under the slow hydrodynamics (4.16 s^{-1}) captured the cumulative and monomer class colloids, reducing turbidity to 9 NTU, with impressive stability ≤ 10 mg/L, and transcends to 95% removal efficiency. The findings are consistent with the results reported in P. S. Nomthandazo et al. [15], Ovuoraye et al. [20], and Nadia et al. [10].

The slope of the curves in Fig. 6 also indicated that the aluminum sulfate sweep flocculation captured considerable quantities of the dimer and trimer classes of colloidal particles present in the medium, resulting in lesser residuals. However, the kinetic was slow, with the flocculation period higher than the optimal (30 min). These results confirmed the destabilization deteriorated under the dominant kinetic switch to higher hydrodynamics [22] resulting from higher stirring ($G = 167 \text{ s}^{-1}$). The findings suggest increasing the residence time beyond 30 min will minimize fine counts present in the system. The kinetic results confirmed the optimization report obtained with the machine learning model.

Machine learning optimized sedimentation tank specifications

The optimal performance of the AS-driven coagulation treatment of the PIE was incorporated into the design of the type 2 sedimentation tank. The sedimentation tank was assumed to be a rectangular structure. The aggregation kinetics proved that the range of the theoretical detention time $t = 30\text{--}60$ min (Fig. 4) is feasible [34], allowing for the colloidal fines count to settle. The high-flow velocity kinetic to the flow deflection of the

sedimentation zone through the density of water can be expressed via the Prandtl number [29, 38].

The dimensions of the tank geometry were deduced following the expression of the mass concentration of the PIE in terms of initial concentration before treatment $C_1 = 520$ mg/L and modeled concentration $C_E = 23$ mg/L. For a mixed liquid (suspension) with varying density, the model equation is expressed in terms of the law of conservation of mass and arbitrary fixed over-flow rate [33]. The inlet flow velocity ($U = 0.048$ cm/s) was set at a volumetric capacity Q (1000 m³/s). The coefficient of viscosity of the effluent at 28 °C is taken to be 0.84 mm²/s and the value for specific gravity (1.002) at $Re < 4$ [29].

The theoretical computation of the tank depth $H_0 = 1349.3 = 1.30$ m was calculated using Eq. 26. The sludge height $H_s = 162$ mm was obtained by substituting the value of H_0 into Eq. 30. The total depth of the sedimentation tank $H = 1151$ mm and corresponding to 1.20 m was evaluated from Eq. 29. The theoretical length of the sedimentation tank $L = 5417$ mm corresponding to 5.40 m was obtained from Eq. 32. The theoretical width of the tank $W = 1.53$ m was evaluated from Eq. 31. The summary of the theoretical sedimentation tank geometry is presented in Table 5. The finding indicates a practical approach to determining the sizing for an economical treatment unit [47].

Conclusions

This study aimed at optimization for the removal of colloidal particles, causing turbidity from PIE to guarantee cleaner effluent discharge via the machine learning algorithm. The performance of the selected ML model (DT, SVM, PM, and RF) was compared using the criterion (≤ 10 NTU) to determine the optimum operating conditions. The train and test datasets were derived from the coagulation-flocculation experiment. The predictive outcomes were compared with the experimentally determine optimum practicable. The results shown at optimum operating conditions (pH 10, dosage of 0.10 g/L, and settling time of 30 min) correspond to residual turbidity ≤ 10 NTU and translate to 23 mg/L concentration. The selected RF regression yielded the most significant outcome with MAE, AAD, and Rmse ≤ 40.10 . The removal efficiency decreased significantly as

Table 5 Machine learning optimized sedimentation tank dimensions

RF model turbidity	Dosage (g/L)	pH	Settling time (Minutes)	Residual concentration (mg/L)	Residual (NTU)
Optimum conditions	0.1	10	30	24	10
Theoretical sedimentation tank geometry					Value
Number of unit					1.00
Tank height (H)					1.20 m
Sludge height (H_0)					1.30 m
Height of settling zone (H_s)					0.16 m
Tank width (W)					1.50 m
Length of tank (L)					5.40 m
Tank surface area (A)					8.10 m ²
Volumetric flow rate (Q)					1000 m ³ /s
Detention time (t)					3600 s

the values of statistical metrics increased. The modeled kinetic rate constant established that the removal of colloidal particles occurred via a sweep flocculation corresponds to 95% removal efficiency. The predictive outputs of the RF model were incorporated to define the type 2 sedimentation tank geometry. It can be inferred from the finding that the predictive capacity of RF is the most reliable. The modeled output confirmed an empirical detention time (60 min) is required to reduce the turbidity in the PIE from 220 to 9 NTU. The finding recommended the ML algorithm as a potential approach to optimizing effluent treatment. The modeled results incorporated into the design equations yielded tank specifications corresponding to a length of 5.40 m, a width of 1.50 m, and a height of 1.20 m will guarantee water recovery with a sludge height of 0.16 m and can be adopted for the design of settling units to guarantee discharged for environmental sustainability.

Abbreviations

PIE	Petroleum industry effluent
RF	Random forest
ML	Machine learning
DT	Decision tree
SVM	Support vector machine
PM	Polynomial regression model
NTU	Nephelometric turbidity unit
AS	Aluminum sulfate
MAE	Mean absolute error
AAD	Average absolute deviation
RMSE	Root-mean-square error
COD	Chemical oxygen demand
BOD	Biological oxygen demand
Std.	Standard deviation

Acknowledgements

The authors wish to thank the Head of Department Chemical Engineering, Dr. V. I. Ugonabo, for his resourcefulness towards the completion of the research work. We also acknowledge the assistance of Engr. Emmanuel Maduakor for his assistance. We sincerely acknowledge the management and staff of Warri Refinery Petrochemical of Company, Nigeria Limited, for allowing access to their facility.

Software permission

The trial versions of the software Python and NCSS PASS version 20 used for this research were obtained prior to permission from the vendors.

Authors' contributions

The manuscript was written via the combined efforts of all authors. PEO created the design matrix for the data, established the kinetics and tank design model and conceptualized the research, and wrote the original manuscript draft. EF managed all software permission, data visualization, and designed the algorithm. AC handled the data training and machine learning optimization algorithm and revised the manuscript. Dr. VIU collected the wastewater samples from the remediation site, examined the floc characteristics, data curation, and supervised the entire research work. The authors read and approved the final manuscript.

Funding

This research was supported by the Nnamdi Azikiwe University Awka, Nigeria. The authors solemnly declare that no external funding was received for this research. The authors were directly involved in the funding of the research.

Availability of data and materials

The authors declare that the data used for this research were obtained from laboratory experimentation with the industrial wastewater. The data stored in a repository will be made available prior to request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 23 September 2022 Accepted: 26 November 2022

Published online: 12 December 2022

References

- World largest refineries, Oil and gas journal (2016) EIA: U.S. Directory of Operable Petroleum Refineries; https://en.m.wikipedia.org/wiki/List_of_oil_refineries
- Refining Crude oil-energy explained, guide to understanding energy, www.tonto.eia.doe.gov
- Adeola OA, Akingboye AS, Ore OT et al (2020) Crude oil exploration in Africa: socio-economic implications, environmental impacts, and mitigation strategies. *J Environ Syst Decis* 42:26–50. <https://doi.org/10.1007/s10669-021-09827-x>
- Zueva S, Corradini V, Ruduka E, Veglio F (2020) Treatment of petroleum refinery wastewater by physiochemical methods* EDP Science E35; Web Conference 161:01042, ICEPP. <https://doi.org/10.1051/e3sconf/202016101042>
- Hammoody Ahmed I, Hassan AA, Sultan HK (2021) Study of electro-fenton oxidation for the removal of oil content in refinery wastewater. *IOP Conf Ser Mat Sci Eng* 1009:012005. <https://doi.org/10.1088/1757-899X/1090/1/012005>
- Varjani S, Joshi R, Srivastava VK, Ngo HH, Guo W (2020) Treatment of wastewater from petroleum industry: current practices and perspectives. *Environ Sci Pollut Res Int*. <https://doi.org/10.1007/s11356019-047205>
- Bant D, Manassia M et al (2020) Combined effects of colloids and SMP on membrane fouling in MBR. *Membrane* 10(6):118
- Mohammad AN, Abdulrazaq KA (2021) Conventional wastewater treatment plant principle and importance factors influencing efficiency. *Des Eng* (8):16009–16027
- Abbas MJ, Al-Sahart RM, Al-Gheethi A, Daud AM (2021) Optimizing FeCl₃ in coagulation-flocculation treatment of dye wastes. *Songklanakarin J Sci Technol* 43(4):1094–1102
- Zaman NK, Rohani R, Yusoff II, Kamsol MA, Basiron SA, Abd Rashid AI (2021) Eco-friendly coagulant versus industrially used coagulants: identification of their coagulation performance, mechanism and optimization in water treatment process. *Int J Environ Res Public Health* 18:9164. <https://doi.org/10.3390/ijerph.18179164>
- Puganeshwary P, Mohd Nordin Adlan HJ, Abdul Aziz H, Murshed MF (2017) Dissolved air flotation for wastewater treatment. *Wastewater treatment in service and utility industries*. <https://doi.org/10.1201/9781315164199>
- Lee CS, Robison J, Chong MF (2014) A review of the application of flocculants in wastewater treatment. *J Process Saf Environ Prot*. <https://doi.org/10.1016/j.esp.2014.04.010>
- Mohammad AN, Abdulrazaq KA (2021) Conventional wastewater treatment plant principle and importance factors influencing efficiency. *Des Eng* (8):16009–16027
- Ovuoraye PE, Ugonabo VI, Nwokocha G.F (2021) Optimization studies on turbidity removal from cosmetics wastewater using aluminum sulfate and blends of fishbone. *SN Appl Sci* 3:488. <https://doi.org/10.1007/s42452-021-04458-y>
- Nomthandazo PS, Rathilal NS, KweinerTetteh E (2021) Coagulation treatment of wastewater kinetics and natural coagulant evaluation. *Molecules* 26:698. <https://doi.org/10.1007/s11696-021-01703-x>
- Zhao W, Xie H, Li J, Zhang L, Zhao Y (2021) Application of alum sludge in wastewater treatment processes Science of Reuse and Reclamation Pathways. *Processes* 9:612. <https://doi.org/10.3390/pr9040612>
- Malik QH (2018) Performance of alum and assorted coagulants in turbidity removal from muddy water. *J Appl Water Sci* 8:40. <https://doi.org/10.1007/s13201-018-0662-5>
- Jaeel AJ, Zaalán SA (2017) Calculation the optimum alum dosages used in several drinking water treatment plants in Wasit Governorate (Iraq) and investigation the effect of pH on alum optimum dosages. *Digital Proceeding of ICOCEE – CAPPADOCIA* (2017), Nevsehir, Turkey
- Benald IO, Oladayo A, Emmanuel O, Chinedu M, Patrick CN, Kelechi NA, Dominic O (2021) Coagulation kinetic study and optimization using response surface methodology for effective removal of turbidity from paint wastewater using natural coagulant* Scientific Africa. Elsevier Publication. <https://doi.org/10.1016/j.sciaf.2021.e00959>
- Ovuoraye PE, Okpala LC, Ugonabo VI, Nwokocha GF (2021) Clarification efficacy of eggshell and aluminium base coagulant for the removal of total suspended solids (TSS) from cosmetics wastewater by coag-flocculation. *Chem Pap* 75(9):4759–4777. <https://doi.org/10.1007/s11696-021-01703-x>
- Ebere Enyoh C, Wang Q, Ovuoraye PE (2022) Response surface methodology for modeling the adsorptive uptake of phenol from aqueous solution using adsorbent polyethylene terephthalate microplastics. *Chem Eng J Adv* (12):100370, ISSN 2666-8211. <https://doi.org/10.1016/j.cjea.2022.100370>
- Ugonabo VI, Emembolu LN, Igwegbe CA, Olaitan SA (2016) Optimal evaluation of coag-flocculation factors for refined petroleum wastewater using plant extract* International Conference Proceedings; Faculty of Engineering, (June 23 - June 24, 2016). Nnamdi Azikiwe University, Awka, Anambra State, Nigeria
- Wang D, Thunell S, Ulrika et al (2021) A machine learning frame work to improve effluent quality control in wastewater treatment plants. *Sci Total Env* 784 (147138). <https://doi.org/10.1016/j.scitotenv.2021.147138>
- Menkiti MC, Igboke PK, Ugodulunwa FXO, Onukwuli OD (2008) Rapid coagulation/flocculation kinetics of coal effluents with high organic content using blended and unblended chitin derived coagulant (CSS). *Res J Appl Sci* 3:317–323
- Smoluchowski M (1999) Versucheiner Mathematischen theorieder koagulation kinetics kolloider Lousugen. *Z Phys Chem* 92:120–168
- Ugonabo VI, Menkiti MC, Onukwuli OD (2016) Micro-kinetics evaluation of coag-flocculation factors for Telefera occidental seed biomass in pharmaceutical effluent system. *J Sci Eng Res*. <https://doi.org/10.1016/j.mex.2019.07.016>
- Hansen SP, Culp GL, Stukenberg JR (1969) Practical application of idealized sedimentation theory in wastewater treatment. *Journal (Water Pollution Control Federation)* 41(8):1421–1444 (<http://www.jstor.org/stable/25039083>)
- Kriš J, Hadi GA (2008) Study the effect of temperature on sedimentation tanks performance. *Water Supply and Water Quality*. In: Proceedings of 20th Jubilee-national, 8th International Scientific and Technical Conference on Water Supply and Water Quality. Poznań, Poland, 15.-18.6.2008, At: Poznań, Poland Polskie Zrzeszenie Inżynierówi Techników Sanitarnych. pp 439–453
- Sadyrbek Djighitekov (2012) Alternative sedimentation theory for rectangular settling tanks design* Conference: Water and Health: Science, Policy and Innovation (October 28 - November 2, (2012). Chapel Hill, North Carolina, USA
- Richardson JF, Zaki WN (1954) Sedimentation and fluidization Part 1. *Trans Inst Chem Eng* 32:35–53

31. Son Jaehyum (2018) "Correlation between sedimentation tank design parameter and sedimentation efficiency" Capstone project, Master degree thesis submitted to KDI School of Public Policy and Management submitted December, 2018 in South Korea Source Wikipedia <https://archives.kdischool.ac.kr/handle/11125/34616>
32. Greenberg LS, Eaton AD. Standard methods for the examination of water and wastewater, 20th ed. APHA, USA
33. A.P.H.A (2015) Standard methods for the examination of water and wastewater 15th edn. American Public Health Association, American water Works Association and Water
34. American Standard Testing and Materials (ASTM) (2015) Water Environ Technol I and II
35. Menkiti MC, Ezemagu IG (2015) Sludge characterization and treatment of produced wastewater using *Typonotonus fuscatus*. *J Pet* 1:51–62
36. Argaman, Kaufman (1968) Design principles for paddle wheel flocculator 1. p 30–120
37. Brathby J (1980) Coagulation and flocculation; with emphasis on water and wastewater treatment. *Filteration and Separation magazine*; Uplands Press Ltd, p 253–1277
38. Djighitekov S. Alternative theory for design of sedimentation tank" Conference paper: Water and Health: Science, Policy and Innovation (October 28 - November 2, 2012). Chapel Hill, North Carolina, USA. <https://researchgate.net/publication/264233253>
39. EPA guidelines for water quality-based decisions (1999) The TMDL process doc. (1991); No. EPA 440/4–91.001
40. WHO guidelines for Drinking-water Quality (2011). The NML classification doc.; 4th Edition, No. WA 675: ISBN 978–9241548151
41. Ohale PE, Onu CE, Ohale NJ, Oba SN (2020) Adsorptive kinetics, isotherm and thermodynamics analysis of fishpond effluent coagulation using chitin derived coagulant from waste *Brachyura* shell. *Chem Eng Journ Adv* 100036. <https://doi.org/10.1016/j.ceja.2020.100036>
42. Zhang Y, Wu Y (2021) Introducing machine learning models to response surface methodologies. <https://doi.org/10.5772/intechopen.9819>
43. Guo H, Kwanho J, Jein L, Jo J, Young MK et al (2015) Prediction of effluent concentration in wastewater treatment plant using machine learning models. *Environ Sci Chinese Academy*:90–101. <https://doi.org/10.1016/j.jes.2015.01.007>
44. Igwegbe CA, Mohammadi L, Ahmadi S, Rahdar A, Khadkhodayi D, Dehghani R, Rahdar S. (2019) Modeling of adsorption of methylene blue dye on Ho-CaWO_4 nanoparticles using response surface methodology (RSM) and artificial neural network (ANN) techniques. *MethodsX*. 19:6:1779–1797. <https://doi.org/10.1016/j.mex.2019.07.0166>
45. Greville AS (1997) How to select a chemical coagulant and flocculants. Alberta waste & wastewater operators' association 22nd annual seminar
46. Elsayed EM, Nour El-Den AA, Elkady MF, Zaatout AA (2020) Comparison of coagulation performance using natural coagulants against traditional ones. *Sep Sci Technol*. <https://doi.org/10.1080/01496395.2020.1795674>
47. Chero E, Torabi M, Hamidreza Z, Anahita G, Bina K (2019) Numerical analysis of the circular settling tank. *J Water Eng Sci Copernicus Pub* 12:39–44. <https://doi.org/10.5194/dwes-1.2-39-2019>
48. Barboza Mariana PJ, La Rovere EL (2020) "Encyclopedia of life support system" Petroleum Engineering- Downstream. Environmental impacts of oil industry
49. Wardzynska R, Smoczyski L, Zaleska-Chrost B (2018) Computer simulation of chemical coagulation and sedimentation of suspended solids. *Ecol Chem Eng S* 25(1):123–131. <https://doi.org/10.1515/eces-2018-0008>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
