Journal of Environmental Management 348 (2023) 119448

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Research article

Application of machine learning algorithms and feature selection methods for better prediction of sludge production in a real advanced biological wastewater treatment plant

Ekin Ekinci [a], Bilge Özbay [b], Sevinç İlhan Omurca [c], Fatma Ece Sayın [b,*], İsmail Özbay [b]

[a] Sakarya University of Applied Sciences, Faculty of Technology, Computer Engineering Department, Sakarya, Turkey
[b] Kocaeli University, Faculty of Engineering, Environmental Engineering Department, Kocaeli, Turkey
[c] Kocaeli University, Faculty of Engineering, Computer Engineering Department, Kocaeli, Turkey

ARTICLE INFO

Handling Editor: Jason Michael Evans

Keywords:
Municipal wastewater
Sludge production
Machine learning models
Prediction
Feature selection

ABSTRACT

Although the management of sewage sludge is an important and challenging task of wastewater treatment, there is a scarcity of studies on the prediction of waste sludge. To overcome this deficiency, the present work aims to develop an appropriate model providing accurate and fast prediction of sewage sludge. With this aim, different machine learning (ML) algorithms were tested by data obtained from a real advanced biological wastewater treatment plant located in Kocaeli, Turkey. In modelling studies, a data set from January 2022 to December 2022 composed of 208 daily measurements was considered. The flow rate of the plant (Q), polyelectrolyte dosage (PD) and removed amounts of total suspended solids (TSS), chemical oxygen demand (COD), biological oxygen demand (BOD), total phosphorous (TP), total nitrogen (TN) were assigned as input parameters to predict sludge production (SP). The precision of the models was evaluated in terms of Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and correlation coefficient (R^2). Among the various tested models Kernel Ridge Regression provided the best accuracy with R^2 value of 0.94 and MAE value of 3.25. Mutual information-based feature selection (MIFS) and correlation-based feature selection (CFS) algorithms were also used in the study in order to enhance the model performance. Thus, higher prediction accuracies were achieved using the selected subset of features. Furthermore, importance contribution of features were calculated and visualized by SHapley Additive exPlanations (SHAP) technique. The overall results of the work indicate the feasibility of ML models for describing the dynamic and complex nature of SP. The process operators may benefit from this modelling approach since it enables accurate and fast estimation of sewage sludge by using fewer and easily measurable parameters.

1. Introduction

Disposal and management of sewage sludge, which is an inevitable by-product of conventional wastewater treatment applications, constitute an important portion of operational costs that approximately account for 50% of the total (Abd El-Wahab et al., 2020). The quantity and characteristics of the sludge are major factors in the budget allocated for sludge disposal and management. SP mechanism is a complex and dynamic process affected by numerous factors such as flow rate and characteristics of the influent, the type of treatment technology, and the experience of the operators (Fytili and Zabaniotou, 2008). An accurate description of this complex mechanism requires powerful models that are capable of considering all affective parameters.

Currently available theoretical formulations based on mass balance may be incompetent in the prediction of SP as they are established on the basis of the accepted biological diversity rate index. The mass balance approach can be described by the following expression:

$$P_{X,TSS} = \text{Heterotrophic Biomass} + \text{Cell Debris} + \text{Nitrifiying biomass} + \text{nbVSS} + \text{Inert solids} \quad (1)$$

As seen from Eq (1) synthesis and debris of cells through metabolism and nitrification are main the processes resulting sewage sludge in the form of total suspended solids ($P_{X,TSS}$). Furthermore, inert solids and

* Corresponding author.
E-mail addresses: ekinekinci@subu.edu.tr (E. Ekinci), bilgealyuz@yahoo.com (B. Özbay), silhan@kocaeli.edu.tr (S.İ. Omurca), ece.sayin@kocaeli.edu.tr (F.E. Sayın), iozbay@kocaeli.edu.tr (İ. Özbay).

https://doi.org/10.1016/j.jenvman.2023.119448
Received 27 June 2023; Received in revised form 20 October 2023; Accepted 20 October 2023
Available online 6 November 2023

non-biodegradable volatile suspended solids (nbVSS) also contribute SP in biological treatment systems. $P_{X,TSS}$ can be calculated by using the mathematical formula given in Eq (2) (Srivastava and Kazmi, 2021):

$$P_{X,TSS} = \frac{QY(S_0 - S)}{1 + k_d(SRT)} * \frac{1}{0,85} + \frac{f_d k_d YQ(S_0 - S)(SRT)}{1 + k_d(SRT)} * \frac{1}{0,85} + \frac{QY_n(NO_x)}{1 + k_{dn}(SRT)} * \frac{1}{0,85} + QX_{0,i} + Q(TSS_0 - VSS_0) \tag{2}$$

Q: Wastewater flow rate (m^3/d)
Y: Biomass yield (g VSS/g COD used)
S_0 and S: Influent and enfluent soluble substrate concentration (BOD or biodegradable soluble COD)(kg/m^3)
k_d: Endogenous decay coefficient (gVSS/g VSS•d)
SRT: Sedimentation retention time (d)
f_d: Fraction of biomass that remains as cell debris (g VSS/g VSS)
Y_n: Cell yield of nitrifiers (g VSS/g NH_4–N)
NO_x: Nitrogen oxidized (kg/m^3)
k_{dn}: Endogenous decay coefficient for nitrifying organisms(gVSS/g VSS•d)
$X_{0,i}$: nbVSS concentration in influent (kg/m^3)
TSS_0: Influent total suspended solids (kg/m^3)
VSS_0: Influent volatile suspended solids (kg/m^3)

Determination of $P_{X\ and\ TSS}$ via mass balance requires calculation and/or acceptance of biological kinetic coefficients, which is a time-consuming and complex task for process operators. Therefore, there is a great demand for novel models for accurate and easy prediction of SP. However, research on this topic is rarely found. In the first known study on this topic, wavelet conjunction models were used to predict the daily SP for a wastewater treatment plant located in Kerman City, Iran (Najafzadeh and Zeinolabedini, 2018). The conjunction models were developed by combination of discrete wavelet transforms (DWT) and single gene expression programming (GEP), model tree (MT) and evolutionary polynomial regression (EPR) models. Results of the study indicated that wavelet coupled-models provided higher accuracy when compared with single ones. In wavelet model tree (WMT) application, RMSE and MAE values decreased 77.4 and 78.01%, respectively and R^2 value improved 92.15% compared to simple MT. Soon after this study, the authors tested efficiency of feed forward back propagation neural network (FFBP-NN) and the radial basis function neural network (RBF-NN) models to predict daily SP for the same plant (Zeinolabedini and Najafzadeh, 2019). Accuracy of the traditional neural network models was improved by employing six mother wavelet (W) functions. In both of the studies, 1-day, 2-day, 3-day and 4-day lag sludge quantities (S(t–1), S(t–2), S(t–3) and S(t–4), respectively) were assigned as input variables to predict the sludge quantity of the present day (S(t)). Although it is possible to make predictions based on daily and/or a few days' SP data, this approach may not provide sufficient interpretation on process dynamics. However, it is a known fact that generation of sewage sludge is a complex process involving numerous factors. Thus, in order to achieve an informative and reliable prediction for SP, more comprehensive data sets of process variables should be considered.

In recent years, ML models have been preferred in many engineering fields due to the superiorities of high speed and accuracy in explaining complex non-linear tasks (Jaffari et al., 2023a; Rajora et al., 2022). A literature search indicates successful applications of ML in the environmental research area (Xia et al., 2022). Fadlil et al. (2022) applied Convolutional Neural Network and Support Vector Machine in order to classify the plastic bottle waste database. In another study, ML models were used to model hydrothermal carbonization of municipal wastes (Zhu et al., 2023). ML algorithms were also efficiently applied to model the adsorption of emerging contaminants onto biochar (Jaffari et al., 2023b; Zhu et al., 2022). Jaffari et al. (2023c) predicted the photocatalytic performance of bismuth ferrite-based materials in the degradation of malachite green by using ML models. Among the twelve ML models, CatBoost provided the best prediction efficiency with the highest R^2 of 0.99 and the lowest MAE of 0.64. El-Rawy et al. (2021) evaluated the performance of wastewater treatment plants by different ML techniques such as traditional feed-forward (TF), deep feed-forward backpropagation (DFB) and deep cascade-forward backpropagation (DCB) networks. Results of the study showed that the DCB network provided the highest accuracy when compared to those of TF and DFB networks. Considering the mentioned literature, ML models may be a promising alternative for explaining complex SP tasks.

Although ML is crucial in the ecological informatics, the high-dimensional datasets and the potential for these datasets to contain a large number of non-informative features provide a serious challenge to learning algorithms. Therefore, a crucial element in reducing complexity, irrelevant features or redundant features is feature selection. It might increase how effectively learning algorithms work. In high-dimensional feature datasets, finding the optimal feature subset is still regarded as an NP-hard task. The search space will expand exponentially as the number of features increases since a dataset with N features consists of 2^{N-1} feature subsets (Atban et al., 2023; Ay et al., 2023). Since classical algorithms are unable to achieve the required outcome in a reasonable amount of time, feature selection algorithms have been employed to select the subset of features. In ecological informatics studies, feature selection studies have great importance (Ba-Alawi et al., 2023; Bagherzadeh et al., 2021; Gadegaonkar et al., 2023; Kushwaha et al., 2023; Lap et al., 2023).

In the scope of this work, different regression algorithms, namely Multiple Linear Regression (MLR), Elastic-Net (EN) Regression, Categorical Boosting (CatBoost) Regression, Kernel Ridge Regression (KRR), Extreme Gradient Boosting (XGBoost) Regression, and Light Gradient Boosting Machine Regression (LightGBMR) were used to predict SP. These algorithms provide a number of benefits over conventional regression algorithms. The effects of different regression approaches have been comprehensively explored. For this aim, to apply widely used regression algorithms MLR and EN Regression; to combine the power of the kernel trick with the standard least-squares regression KRR algorithm; to create a more accurate model by correcting the mistakes of the previous weak learners, the Gradient boosted regression algorithms CatBoost, XGBoost, LightGBM have been considered.

In the first step of the study, preselected regression algorithms were run with seven features (flow rate of the plant, PD and ΔTSS, ΔCOD, ΔBOD, ΔTP and ΔTN). To identify the most informative input features, MIFS and correlation-based feature selection CFS methods were applied. These methods evaluate subsets of features using a measure independent of the inductive learning algorithm because they presume total independence between the learning machine and the data (Ebiaredoh-Mienye et al., 2022).

The main contributions of this paper can be summarized as follows:

- to develop an accurate model for the prediction of SP by using a real-world dataset obtained from an advanced biological treatment plant in Kocaeli, Turkey.
- to test the performance of different ML algorithms for SP modelling
- to reduce the dimensionality of the feature space, resulting in a more accurate regression.

While the most successful regression was achieved with KRR, both feature selection algorithms successfully selected fewer features, greatly reduced the number of features, and improved the KRR's regression accuracy.

The prevailing novelty of this study is the accurate prediction of SP using easily measurable operational parameters. There is no similar study, according to our best knowledge. Furthermore, it is thought that the findings of the study will contribute to the determination of sludge minimization strategies in activated sludge processes.

2. Materials and methods

2.1. Wastewater treatment description

This study was carried out using data from the advanced biological treatment plant in the eastern region of Kocaeli, Turkey. The construction area lies on latitude 40°45′04″ North and longitude 29°58′ 05″ East and covers 29.13 ha area, of which 0.24 ha is closed. The plant was designed to serve the equivalent population of 447,000 with 140,000 m^3/day treatment capacity. Advanced biological treatment is implemented in the plant. Long aerated activated sludge process is utilized in which nitrogen and phosphorus removal along with carbon oxidation takes place. The flow chart of the plant is presented in Fig. 1.

As can be seen from Fig. 1, the treatment plant consists of a succession of anaerobic, anoxic and aerobic tanks, a settling tank, and a sludge return connection transporting the settling sludge to the anaerobic tank inlet. Treatment sludge with high phosphorus content obtained as a result of treatment processes is disposed of in licensed facilities after being subjected to thickening, aerobic sludge digestion and dewatering processes. In dewatering, 0.2 % cationic polyelectrolyte solution is dosaged to the entry of the decanter during sludge feeding. 3 kg polyelectrolyte is consumed per ton of dewatered sewage sludge to provide 20 % TSS content. There is also an Ultraviolet (UV) disinfection unit with a capacity of 17.000 m^3/day. The treated wastewater, which is subjected to advanced treatment, is used as process water in industrial establishments and for irrigation of green areas.

The efficiency of the treatment plant is controlled by monitoring conventional water quality parameters (BOD, COD, TSS, TN, TP). BOD is analyzed according to the respirometric method using OxiTop Control 6 WTW instrumentation, whereas COD is measured according to the titrimetric method (ASTM D1252-A). ISO 29441 and SM 4500 P– I standard methods are followed for TN and TP analyses. TSS were measured according to the SM 2540 D gravimetric method.

2.2. Modelling studies

In the study, the used data set was composed of 208 daily average measurements of flow rate, PD and ΔTSS, ΔCOD, ΔBOD, ΔTP, ΔTN (input variables) and SP (output). Data collection was performed in the period from January to December 2022. The lines, including even one missing variable, were extracted from the raw data set.

2.2.1. Experimental setup and evaluation metrics

The dataset split for each algorithm and model parameters should be originally determined individually. The dataset is divided into two parts as train set/test set split, with different ratios of 70:30, 75:25, and 80:20. For 70:30 ratios, there are 145 samples in the train set and 63 samples in the test set. For 75:25 ratios, there are 156 samples in the train set and 52 samples in the test set. For 80:20 ratios, there are 166 samples in the train set and 42 samples in the test set.

Between the lasso penalty (alpha = 1) and the ridge-regression penalty (alpha = 0), the EN penalty offers a helpful middle ground (Cho et al., 2009). The EN Regression works similarly to the lasso with alpha = 1 - epsilon for some small epsilon >0, but it is more resistant to severe correlations among predictor variables. Thus, the EN Regression was implemented for this study using an alpha value of 1. Regularization enhances the problem's conditioning and lowers the estimates' variance. Stronger regularization is indicated by larger values. For KRR, the regularization parameter alpha was used as 1. The number of trees is a critical parameter that plays an important role in forecasting the final value to realize the XGBoost regression model (Peng et al., 2023). The number of trees was selected as 100 for the forecasting. The loss function used in training was RMSE for the CatBoost regressor. The reason why we use RMSE is that larger errors are punished more severely by RMSE than smaller ones (Hancock and Khoshgoftaar, 2020a). The number of trees used for prediction was 100 and passed as a parameter for the LightGBM Regressor. As a weighting component, the learning rate was used to slow down the learning process and was set to 0.1. and boosting type was selected as the traditional Gradient Boosting Decision Tree (GBDT). The decision tree serves as the weak learner in the GBDT approach, which integrates them into a robust model (Jing et al., 2023).

The EN Regression and KRR models were implemented for this study using an alpha value of 1. The XGBoost regression model accepted 100 trees as a parameter for the forecasting. The loss function used in training was RMSE for the CatBoost regressor. The number of trees used for prediction was 100 and passed as a parameter for the LightGBM Regressor. The learning rate was set to 0.1, and the boosting type was selected as the traditional Gradient Boosting Decision Tree (GBDT).

To see the effect of feature selection algorithms, the performances of forecasting algorithms are measured by using several evaluation metrics such as MAE, MSE, MAPE, RMSE and R^2. MAE is the expected value of the sum of absolute errors of the actual and predicted daily generated SP. MSE is the expected value of the square of the difference between the actual and predicted daily generated SP. MAPE is the average of absolute

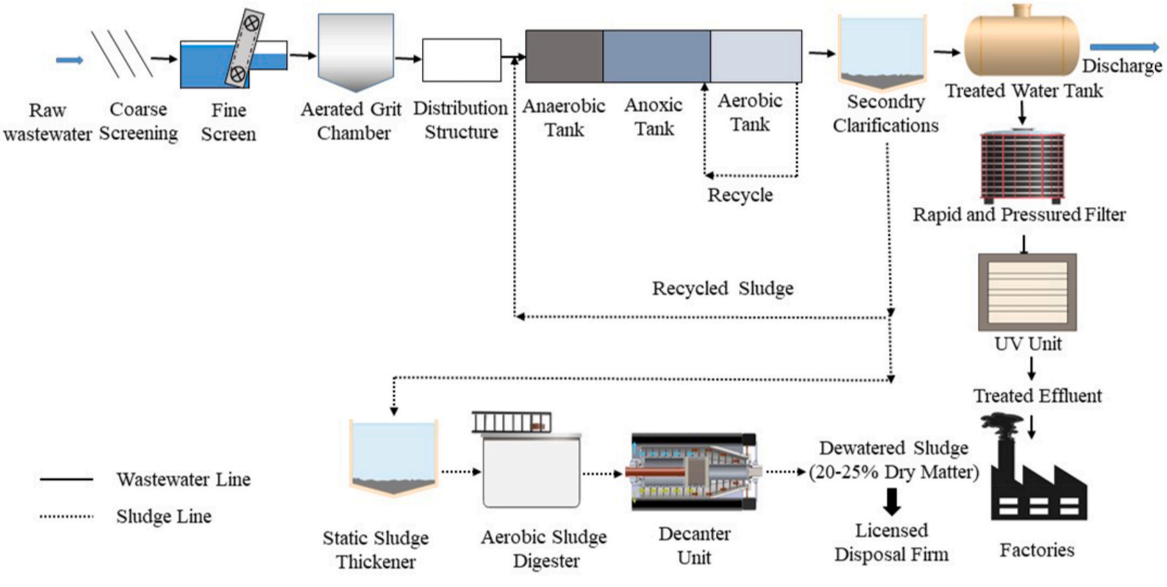

Fig. 1. Flow chart of the investigated treatment plant.

percentage errors of the actual and predicted daily generated SP. RMSE is the square root of the MSE. R^2 quantifies how closely the data is to the fitted regression line by calculating the proportion of the daily generated SP (dependent variable) variance that can be explained by a linear model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{3}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{4}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}(|y_i - \widehat{y}_i| / y_i) * 100 \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{6}$$

$$R^2 = 1 - \sum_{i=1}^{n}\frac{(y_i - \widehat{y}_i)^2}{(y_i - \overline{y})^2} \tag{7}$$

In the equations above, y_i is the accurate value, $\widehat{y}_i$ is the predicted value, and $\overline{y}$ is the mean of the daily generated SP; n is the number of observations in the test set. While MAE, MSE, MAPE, and RMSE are expected to be small, R^2 is expected to be high. According to our data, the most accurate way to assess average error magnitude is using the MAE.

2.2.2. *Machine learning models*

In this section, the regression and the feature selection methods used in modelling municipal wastewater sludge are summarized.

2.2.2.1. Multiple Linear Regression. In ML, regression models are used to evaluate the relationship between dependent and independent variables. The MLR model is represented by Eq (8).

$$\widehat{y}_i = (\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n) + \varepsilon_i \tag{8}$$

in the MLR model, $\widehat{y}_i$ is the dependent variable, $x_1 \ldots x_n$ are independent variables, $\theta_0 \ldots \theta_n$ are the regression parameters, and ε_i is the noise parameter. The objective is to estimate the parameters of the model by an error minimization algorithm such as least squares. The sum square error of the model is defined as in Eq (9):

$$SSE = \sum_{i}^{m}(y_i - \widehat{y}_i)^2 \tag{9}$$

To estimate the model parameters, an iterative error minimization algorithm such as stochastic Gradient descend can be used, where α is learning rate.

Stochastic_Gradient_Descend (training examples, α)

Each training example is a pair of the form $< \overrightarrow{x}, y >$; $\overrightarrow{x}$ is the vector of input features; y is the target output value, α is the learning rate.

```
● Set initial value of each $\theta_i$ to a modest random number
● Until it convergence, Do
    o Initialize each $\Delta\theta_i$ to zero
    o For each $< \overrightarrow{x}, y >$ in training examples, Do
          ▪ Compute the output $\widehat{y}$.
          ▪ For each $\theta_i$, Do.
$\Delta\theta_i \leftarrow \Delta\theta_i + \alpha(y - \widehat{y})x_i$

  o Update each parameter $\theta_i$, Do
$\theta_i \leftarrow \theta_i + \Delta\theta_i$
```

2.2.2.2. Elastic-Net regression. In linear regression (LR) models, the prediction performance is highly affected by the bias–variance trade-off. Therefore, an accurate penalization method is critical to achieve higher prediction rates. EN approach considers lasso and ridge parameters to reduce the bias (Minaravesh and Aydin, 2023).

While Lasso and Ridge regression models perform L1 and L2 regularizations, respectively, the EN model combines these two regularization terms and balances the proportion of them by using a weight parameter. The loss function is defined as in Eq (10):

$$L(\theta) = \frac{1}{2m}\sum_{i}^{m}\left(y^{(i)} - \theta^T x^{(i)}\right)^2 + \alpha\left(\lambda\sum_{j}^{n}|\theta_j| + \frac{(1-\lambda)}{2}\sum_{j}^{n}\theta_j^2\right) \tag{10}$$

where m represents training data size, $x^{(i)}$ is the data point, and $y^{(i)}$ is its label. Where $\lambda \in [0,1]$ controls the strength of L1 and L2 penalties, and $\alpha > 0$ are the regularization parameters of EN.

The EN Regression is an LR model that can be accurately applied when the features are strongly correlated (Zou and Hastie, 2005). In our experiments, we applied the EN Regression model, which can more accurately reveal the linear relations of the features with each other.

2.2.2.3. Kernel Ridge Regression. Kernel methods have been successfully incorporated into various ML models, such as support vector machines (Varma et al., 2019). KRR (Saunders et al., 1998) is widely used for regression problems as a computationally efficient model. It also has a high generalization capability. KRR is based on the non-linear transformation of raw data. The sample data $X \in R^{n\times m}$ in low dimensional space is first transferred to a high dimensional feature space using a non-linear mapping function ($\varnothing$). The non-linear regression problem is then solved in a high dimensional space using the linear regression approach (Fan et al., 2021). (Detailed principles are presented in Supplementary Materials).

2.2.2.4. Gradient Boosting. Ensemble Learning (EL) is one of the learning approaches in ML that aims to obtain better predictive performance by combining the predictions from weak learners (base learners). In EL, rather than creating a strong prediction hypothesis, it is important to unify the results of multiple weak hypotheses for the final prediction. Bagging and boosting are the two main classes of EL. The key idea in bagging is to train more than one model in parallel on different samples of the original dataset and combine the predictions. Moreover, boosting has a pipeline of sequential classifiers in which each classifier corrects the predictions made by priors, and this learning process is repeated many times. Finally, it outputs a weighted average of the predictions.

Gradient Boosting is a powerful boosting algorithm that can handle problems with noisy data and complex dependencies, as in heterogeneous features and can also return good results without massive data. It works by training decision trees as weak learners on gradient-based functions and incorporating them into the EL model as "boosted" participants. In our study, we applied the XGBoost, LightGBM, and CatBoost, which are state-of-the-art Gradient Boosting algorithms.

2.2.2.5. Extreme Gradient Boosting. XGBoost (Friedman, 2001) is based on decision trees as the other Gradient-boosting algorithms. In XGBoost, the combined tree structures are as in the classification and regression trees (CART) model. Because it is a regression model, real-valued scores are associated with leaf nodes instead of class labels in the CART model. Thus, to compute a final output, each individual tree's prediction scores are summed up as in Eq (11).

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{11}$$

in the Equation above, while K is the number of trees, f_k is a function of a tree in F. F is the space of regression trees. f_k contains the tree structure and leaf node scores. The objective function is given in Eq (12).

$$obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^k\right) + \sum_{k=1}^{K} w(f_k) \tag{12}$$

In XGBoost, the regularization term $w(f_k)$ is the complexity of the f_k. (Detailed principles are presented in Supplementary Materials)

2.2.2.6. Categorical Boosting regression. CatBoost (Prokhorenkova et al., 2018) is an ensemble-based ML method that is implemented based on the Gradient boosting algorithm on symmetric decision trees. Yandex proposed CatBoost in 2017 as one of the Boosting family algorithms. Catboost can both preprocess categorical data efficiently and implement ordered boosting to avoid overfitting for small datasets. Compared to other boosting algorithms, it has a better generalization capacity and higher prediction accuracy rates. Besides, CatBoost can provide faster convergence as it reduces the need for hyperparameter optimization (Zeng et al., 2023). It has been indicated that the performance of CatBoost is relative to hyperparameter tuning, such as the learning rate and the number of trees (Hancock and Khoshgoftaar, 2020b).

2.2.2.7. Light gradient machine boosting. LightGBM (Ke et al., 2017) is a boosting algorithm proposed as a part of the Microsoft Distributed Machine Learning Toolkit (DMTK) project. While other GBDT-based algorithms, such as XGBoost or CatBoost, grow trees horizontally, LightGBM grows the tree vertically. As a result, the algorithm is guaranteed to have benefits, including fast processing of massive amounts of data and features, high prediction rates, parallel learning, and less resource (RAM) utilization. The experiments conducted by Ke et al. (2017) showed that LightGBM provided almost the same accuracy as other GBDT-based algorithms while speeding up training 20 times. Since computational complexity in GDBT algorithms depends on the number of features, such algorithms in big data are quite time-consuming. Unlike other algorithms, two techniques are designed to solve this problem: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). The intuition behind GOSS is the instances with the larger gradients will contribute more to the information gain. Therefore, it is aimed to keep the data instances, which have large gradients and drop the instances, which have small gradients randomly. By doing this, a more accurate gain estimation has been achieved than uniformly random sampling. In the case of sparse feature space (such as, one-hot encoding), EFB aims to reduce the number of features with re-categorizing mutually exclusive features into bundles.

Another important aspect in LightGBM is using a histogram-based algorithm to identify the optimal split points. Continuous feature values are discretized into bins by a histogram-based technique, and during training, feature histograms are constructed from the feature bins (Ke et al., 2017). To summarize, LightGBM has many improvements and produces successful results in experiments involving both classification and regression.

2.2.3. Feature selection

Feature selection is a very crucial task, which has been extensively studied by the ML community. The goal of feature selection is to select the best feature subset with k features that yields the minimum generalization error. In traditional ML approaches, feature selection methods are applied at the data preprocessing stage for an accurate classification or regression. In feature selection algorithms, the features are ranked according to their relevance to the dependent variable in the supervised learning.

Feature selection algorithms look up the whole feature space and evaluate possible subsets by the relevance measures calculated by using various techniques. Two of the important relevance measures are Mutual Information (MI) (Battiti, 1994) and correlation coefficient. According to an evaluation function based on correlation, CFS ranks feature subsets. The feature subset evaluation function of CFS is defined as follows:

$$P_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{13}$$

where S is a subset of k features, P_S is the potential measurement of a feature subset S which contains k features. While $\overline{r_{cf}}$ represents the mean feature class correlation $(f \epsilon S)$, $\overline{r_{ff}}$ represents the average feature-feature inter-correlation. These correlations are computer as:

$$\overline{r_{xy}} = \frac{E\left[(x-\mu_x)\left(y-\mu_y\right)\right]}{\sigma_x\sigma_y} \tag{14}$$

where μ_x , μ_y are the mean values and σ_x, σ_y are the standard deviation values of x and y, respectively.

eqs (13) and (14) form the core of CFS for ranking the feature subsets.

Information measures defined in information theory, such as MI, are also used to compute the relevance or redundancy of features (Battiti, 1994).

The MI, on the other hand, is able to capture non-linear correlations between features, whereas the correlation is unable to do so. On the other hand, for computations based on correlation, all features must be numerical features, however, MI can deal with both numerical and categorical features (Vinh et al., 2012). The representation of proposed method is given with Fig. 2.

3. Results and discussion

3.1. Evaluation of the used data

Statistical evaluation of the used data has been presented in Table 1. According to the classification of municipal wastewater considering the strength type of wastewater, strong wastewaters have COD, BOD and TSS concentrations of approximately 1000 mg/L, 300 mg/L and 400 mg/L, respectively (Metcalf et al., 2014). The treated wastewater can be classified as medium strength considering average TN and TP concentrations, whereas relatively higher COD, BOD, and TSS values refer to strong wastewater characteristics. On the other hand, wastewater classification based on BOD/COD ratio put forward three groups as slowly bio-degradable (0.2–0.4), average bio-degradable (0.4–0.5) and readily bio-degradable (0.5–0.8). In this study, BOD/COD ratio is calculated as 0.29 ± 0.09, indicating hardly the bio-degradable nature of the influent. As seen from Table 1, the national discharge limits are ensured considering pollution concentrations in the effluent. A daily average 60 tons of dewatered sewage sludge is generated in the process.

In the study, interactions between the parameters were also examined by bivariate correlation analyses. Determined Pearson's coefficients are summarized in Table 2. When investigating the impacts of model inputs on SP, the highest positive correlation was determined for polyelectrolyte dosage with R^2 value of 0.95. This is an expected result as polyelectrolyte is the fundamental chemical utilized in the dewatering process proportional to incoming sludge volume. Flow rate has a moderate positive correlation with SP ($R^2 = 0.60$), whereas all pollution parameters exhibited weak positive interactions. Since SP relies on the synthesis of new cells via biodegradation of organic materials, ΔBOD and ΔCOD have relatively higher R^2 values (0.43 and 0.42, respectively). These results are all in good accordance with the results of feature selection applications as polyelectrolyte dosage, flow rate and organic matter removal were the common selected features of the applications.

When investigated the relationships between the input variables, ΔBOD has significant positive correlations with ΔCOD and ΔTSS ($R^2 >$ 0.80), as expected. On the other hand, weaker positive correlations were obtained for individual bivariate relations between ΔTP and ΔBOD, ΔCOD, ΔTSS explained by R^2 of 0.45 approximately. A significant

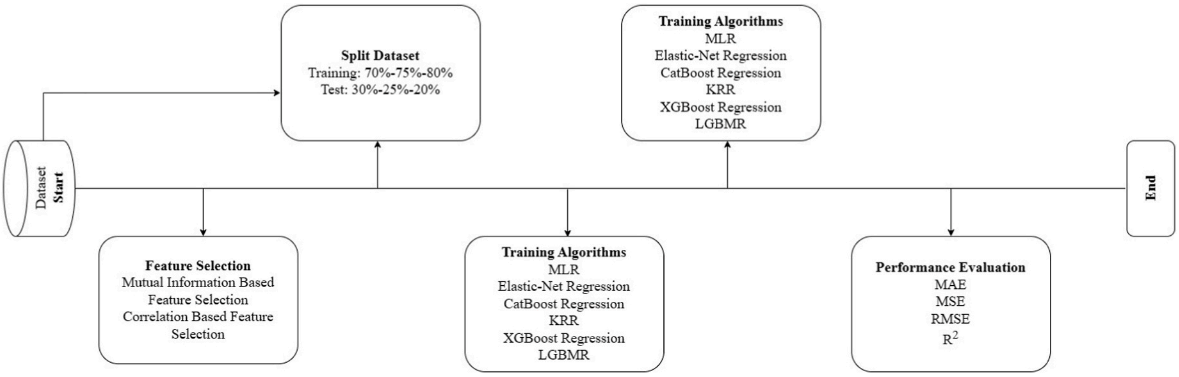

Fig. 2. Representation of proposed method.

Table 1
Average values of the model parameters.

Parameter	Unit	Average Value		
Input		Influent	Effluent	Discharge Standard (UWRT, 2006)
BOD	mg/L	305 ± 211	11 ± 6	25
COD	mg/L	1117 ± 857	40 ± 28	125
TSS	mg/L	719 ± 517	12 ± 11	60
TN	mg/L	34.60 ± 12.03	6.71 ± 2.16	15
TP	mg/L	4.99 ± 2.51	0.50 ± 0.27	2
Q	m^3/day	97,543 ± 14,355		
PD	kg/day	154 ± 66		
Outputrowhead				
SP	ton/day	60 ± 25		

interaction was observed between polyelectrolyte dosage and flow rate ($R^2 = 0.60$) which is also in good accordance with SP.

3.2. Results of SHAP analysis

SHAP is a useful method used to describe the supervised learning models which provides an importance value for each of the input parameters individually (Mangalathua et al., 2020). Model sensitivity to variables can be explored by the calculated SHAP values for the attributes.

In the study, SHAP values were determined for the KRR model in order to examine and validate the analysis results from another point of view.

As clearly seen from the SHAP plot, PD is the attribute with the greatest influence on SP and it is followed by Q (Fig. 3). These results are consistent with those obtained from MIFS and CFS methods. However, significance of ΔTP has increased while significance of both ΔCOD and ΔBOD have reduced in the SHAP analysis. These results can be explained by the fact that SHAP values emphasize the contribution of each feature within each sample, whereas the model's feature contribution reflects the feature's weight.

3.3. Results of the modelling studies

The main motivation behind this study is that the feature selection algorithms could determine the best representative features for regression algorithms. To see if this motivation is satisfied, various regression models were constructed, and numerous feature selection algorithms were applied to the original feature set.

When MIFS and CFS algorithms were applied to the dataset consisting of seven features in total, the number of features decreased to three for both algorithms. While the MIFS algorithm selected the flow rate of the plant, ΔBOD and daily PD, the CFS algorithm selected the daily flow rate of the plant, ΔCOD and daily PD.

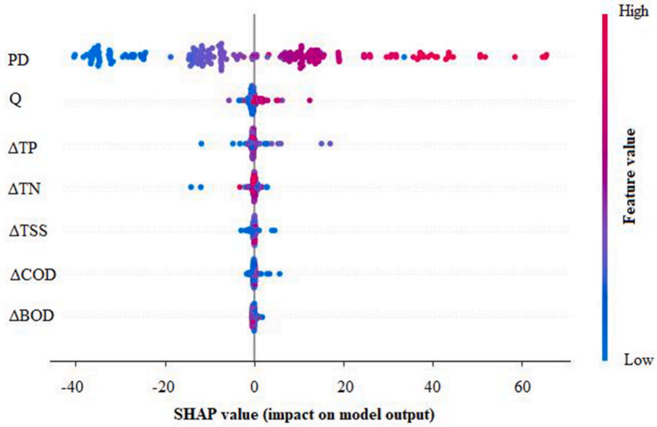

Fig. 3. Impact of input features on SP for KRR model.

Table 2
Pearson's correlation coefficients between the model parameters.

	Q (m^3/d)	ΔTSS (mg/L)	ΔCOD (mg/L)	ΔBOD (mg/L)	ΔTP (mg/L)	ΔTN (mg/L)	PD (kg/d)	SP (tons/d)
Q	1	0.32[a]	0.30[a]	0.39[a]	0.04	−0.06	0.60[a]	0.60[a]
ΔTSS		1	0.91[a]	0.92[a]	0.43[a]	0.12	0.38[a]	0.37[a]
ΔCOD			1	0.84[a]	0.47[a]	0.06	0.42[a]	0.42[a]
ΔBOD				1	0.46[a]	0.19	0.43[a]	0.43[a]
ΔTP					1	0.17	0.06	0.06
ΔTN						1	0.05	0.05
PD							1	0.95
SP								1

[a] Correlation is significant at the 0.01 level (2-tailed).

The selected features are found to be meaningful, considering the well-known literature on the activated sludge process. Although the biological oxidation of ammonium to nitrate yields a certain amount of cell mass, this quantity is significantly lower than that of coming from organic materials, and it is generally negligible in the determination of SP (Reynolds and Richards, 1996). Substrate removal is the prominent factor affecting SP, and substrate load is defined by BOD and COD parameters. Results of feature selection applications are in good accordance with the literature as COD and BOD parameters representing organic load are selected. Additionally, the selection of PD is also sensible since polymer was added into the entry of the decanter proportional to the flow rate of the sludge.

When investigated the data distribution of the selected features (Fig. 4), it is clearly seen that presented dataset is normally distributed. There are no missing data in the dataset for any of the attributes considered. This demonstrates that the feature-selected datasets, which include three attributes, are useful and should be taken into account for the development and evaluation of various ML algorithms. Obtained feature-selected datasets were split into training and test sets and ML models were applied.

The performance evaluation metrics (MAE, MSE, MAPE, RMSE and R^2) are presented in Table 3. This table provides comparison of MLR, EN, CatBoost, KRR, XGBoost and LightGBMR regression algorithms based on the original dataset and feature selected dataset under the same conditions.

As a result of the comparison, it was seen that the results obtained with the feature selection were more successful than the results obtained from the original data set. According to average MAE the best results

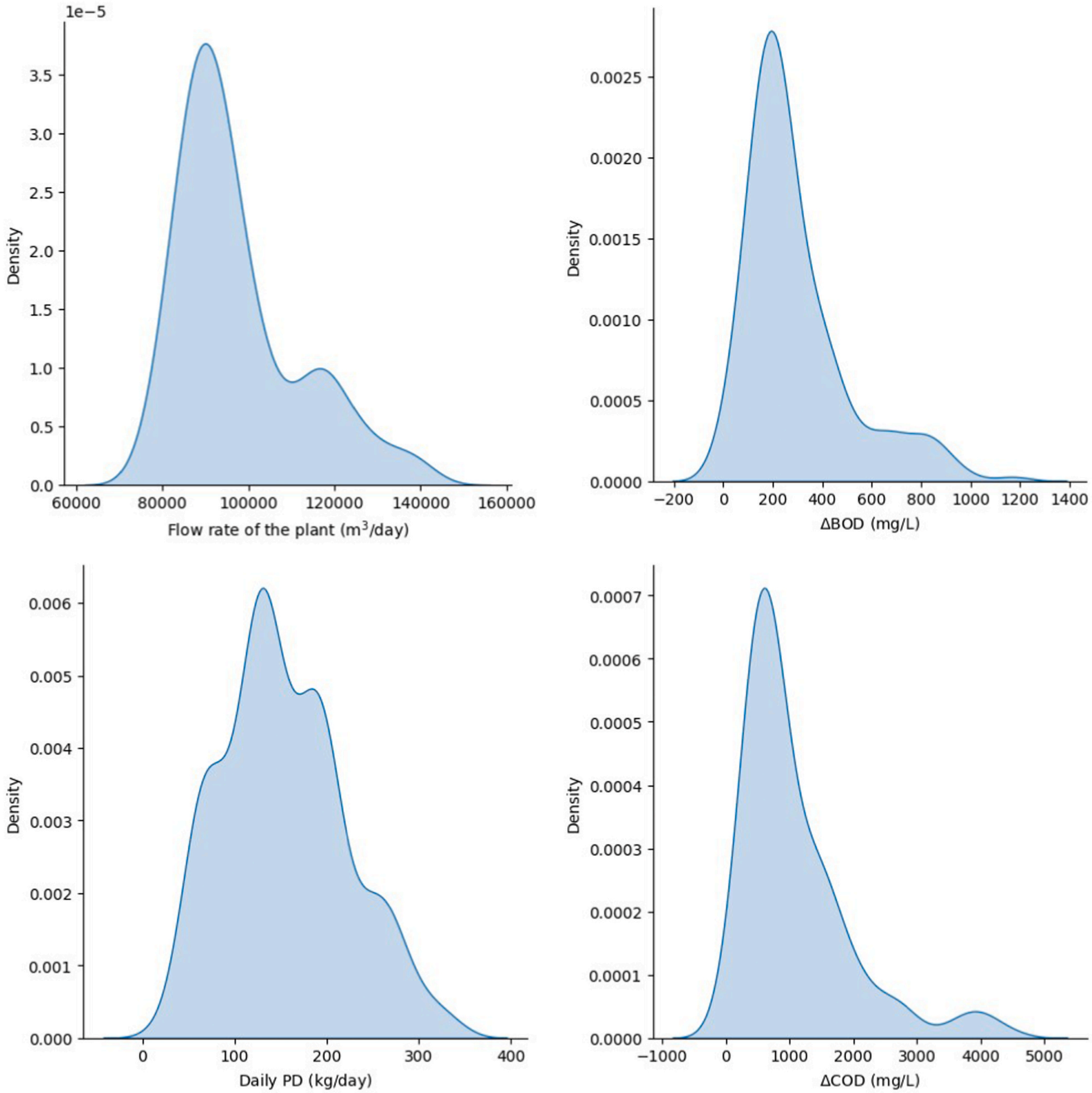

Fig. 4. Data distribution of selected features (a) Flow rate of the plant, (b) ΔBOD, (c) Daily PD, (d) ΔCOD.

Table 3
Performance evaluation of different ML models for SP prediction.

Dataset	Algorithm	70:30					75:25					80:20				
		MAE	MSE	MAPE	RMSE	R^2	MAE	MSE	MAPE	RMSE	R^2	MAE	MSE	MAPE	RMSE	R^2
Original	MLR	3.42	39.55	6.10	6.29	0.94	3.27	43.88	5.95	6.62	0.92	3.41	51.93	5.71	7.21	0.89
	EN	3.35	39.55	6.03	6.29	0.94	3.24	43.93	5.94	6.63	0.92	3.34	51.78	5.62	7.20	0.89
	CatBoost	4.28	68.27	5.90	8.26	0.89	3.43	53.49	5.28	7.31	0.9	3.63	62.93	5.20	7.93	0.87
	KRR	3.25	38.85	5.63	6.23	0.94	3.19	43.61	5.61	6.60	0.92	3.22	51.23	5.11	7.16	0.89
	XGBoost	4.61	109.11	6.76	10.45	0.83	3.51	61.2	5.53	7.82	0.89	4.0	75.87	5.86	8.71	0.84
	LightGBMR	5.13	74.32	7.66	8.62	0.89	4.63	56.05	7.35	7.49	0.90	4.69	60.22	7.37	7.76	0.88
MIFS	MLR	2.94	35.20	5.36	5.93	0.95	2.77	38.68	5.10	6.22	0.93	2.81	46.12	4.70	6.79	0.91
	EN	2.95	35.25	5.38	5.94	0.95	2.78	38.71	5.11	6.22	0.93	2.82	46.15	4.71	6.79	0.91
	CatBoost	3.98	59.85	6.18	7.74	0.91	3.45	67.14	5.57	8.19	0.88	3.61	76.30	5.25	8.74	0.84
	KRR	1.93	33.23	2.79	5.76	0.95	1.97	38.78	2.77	6.23	0.93	2.11	47.19	2.72	6.87	0.90
	XGBoost	4.01	63.95	6.49	8.00	0.90	3.44	52.15	5.87	7.22	0.90	3.69	62.21	5.58	7.89	0.87
	LightGBMR	4.73	63.18	7.55	7.95	0.90	4.35	45.20	7.47	6.72	0.92	4.07	47.47	6.76	6.89	0.90
CFS	MLR	3.01	35.92	5.45	5.99	0.94	2.74	38.51	5.04	6.21	0.93	2.83	46.22	4.75	6.80	0.90
	EN	3.02	35.95	5.46	6.00	0.94	2.75	38.54	5.05	6.21	0.93	2.83	46.24	4.76	6.80	0.90
	CatBoost	3.88	61.24	5.30	7.83	0.91	3.37	61.17	5.15	7.82	0.89	3.00	51.84	4.41	7.20	0.89
	KRR	1.98	33.53	2.80	5.79	0.95	1.90	38.44	2.61	6.20	0.93	2.10	47.20	2.73	6.87	0.90
	XGBoost	3.27	45.76	5.04	6.76	0.93	2.95	46.63	4.54	6.83	0.91	3.23	57.15	4.86	7.56	0.88
	LightGBMR	5.15	70.97	7.90	8.42	0.89	4.17	49.30	7.16	7.02	0.91	4.05	57.14	6.47	7.56	0.88

were obtained with data split of 75:25. When the data split of 75:25 is examined by comparing the original dataset with MIFS and CFS, respectively, the results obtained are summarized as follows:

In MLR model, MIFS and CFS provided decrease of 15.09% and 16.04% in MAE values whereas there was no significant difference noted for R^2. When evaluated results of EN algorithm, the most remarkable decrease was determined for MAE values by 14.2% and 15.13%, respectively for MIFS and CFS. Similar to MLR algorithm, there was only 1% increase in the R^2 values. When compared with the others, CatBoost algorithm was not efficient considering 2.5% and 1.4% decreases observed in the R^2 value. Increase of 25.52% and 14.37% recorded in MSE values also indicates the inadequacy of the algorithm for this study. With the KRR algorithm, MAE values decreased by 38.23% and 40.67%; MSE values decreased by 11.08% and 11.86%; MAPE values decreased by 50.62% and 53.48% and finally RMSE values decreased by 5.71% and 6.12%. Additionally, an increase of 0.8% and 0.9% were obtained in R^2 values. When error values of XGBoost were evaluated, decreases of 2.11% and 16.09% in MAE, decreases of 14.78% and 23.81% in MSE, an increase of 6.15% and a decrease of 17.9% in MAPE and decreases of 7.7% and 12.72% in RMSE were determined. R^2 values have slightly increased by 1.7% and 2.7%. In LightGBMR algorithm, MAE values decreased by 5.92% and 9.98%, MSE values decreased by 19.35% and 12.03%. An increase of 1.63% and a decrease of 2.59% were obtained for MAPE, and decreases of 10.19% and 6.2% were observed for RMSE

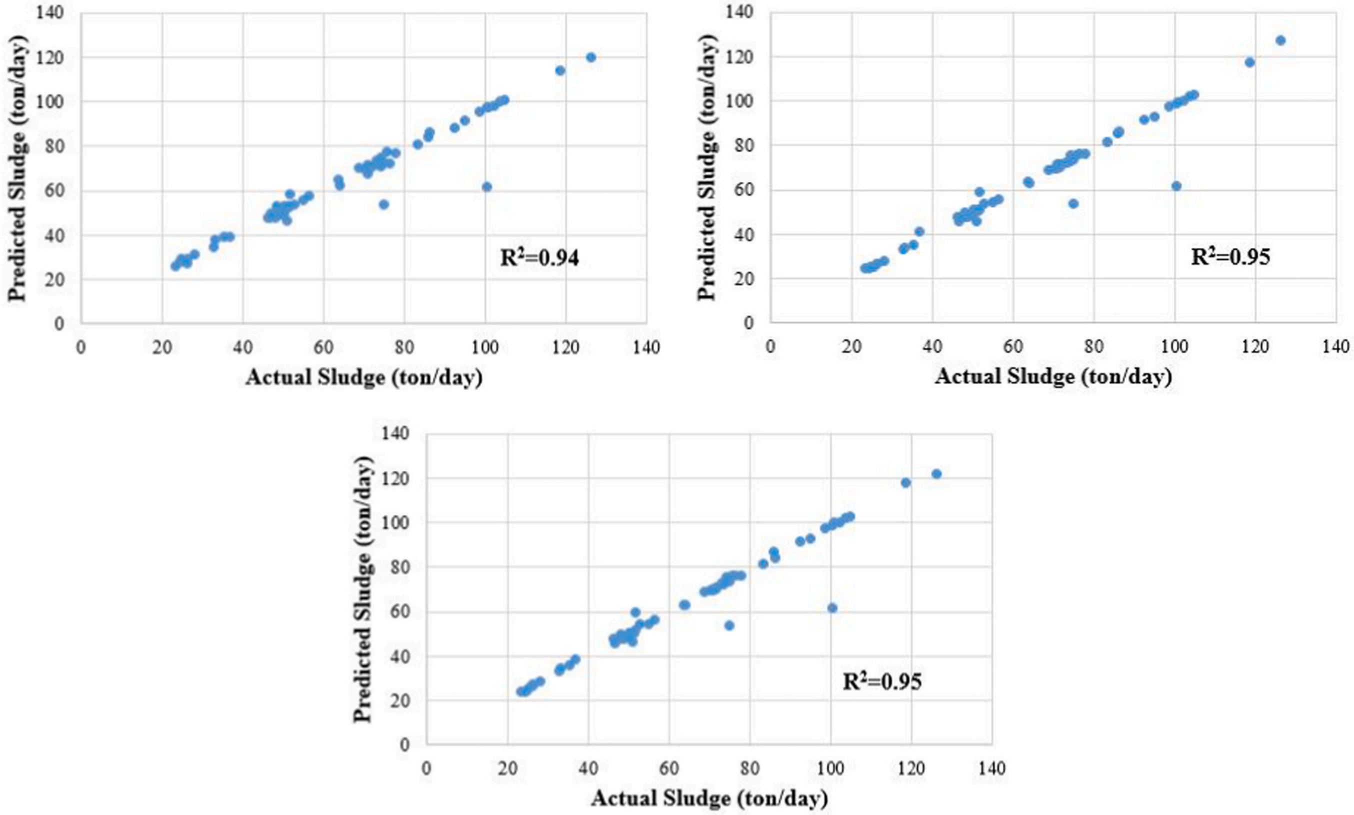

Fig. 5. Regression plots of KRR for test set of (a) original observations (b) MIFS and (c) CFS.

values.

As a result, for the original dataset and feature-selected datasets, The KRR algorithm outperforms the MLR, EN, CatBoost, XGBoost and LightGBMR. Furthermore, CatBoost performed quite poorly for the dataset used relative to MLR, EN, KRR, XGBoost and LightGBMR.

KRR provided MAE between 1.90 and 2.10 for feature-selected datasets whereas for original dataset, MAE values were in the range of 3.19 and 3.25. Both feature selection algorithms significantly decreased the number of features effectively. Selected fewer features enhanced the KRR's predictive capability. The regression plots of the best models are given in Fig. 5.

When Fig. 6 is evaluated, it has been seen that there are only two fluctuations, as shown in the black boxes. As a result, prediction results are nearly the same as actual values indicating promising efficiency of KRR algorithm for SP prediction.

The KRR model works well because it can deal with nonlinearity issues that can arise in datasets. The expansion of the linear ridge regression, known as KRR, allows for the acceptance of non-linear data. The kernel technique is an extremely effective weapon. The algorithm behaves nonlinearly while using the kernel. The kernel uses linear regression algorithms on non-linear data.

4. Conclusions

The major aim of this study was to predict municipal wastewater SP for a real advanced biological treatment plant. With this aim, the ML modelling approach, which is well-known for its superiority in explaining complex tasks, was applied. Among the tested algorithms, KRR provided the highest modelling efficiency with lower error values (3.25, 38.85 and 6.23, respectively, for MAE, MSE and RMSE) and a higher R^2 value of 0.94. MIFS and CFS algorithms were also applied to improve prediction accuracy using fewer selected features. Both of the algorithms decreased the original data set from seven features to three, providing similar meaningful results. Flow rate of the plant and polyelectrolyte dosage were the common features of the new data sets, whereas removal of organic matter was explained in terms of BOD and COD removal in MIFS and CFS algorithms, respectively. Selected features have improved the accuracy of the model at a ratio of 40.7 %, considering lower MAE values. Additionally, results of correlation analysis and SHAP applications have also demonstrated that PD and Q are the parameters with the greatest influence on SP.

Overall, the results of the study provide promising future perspectives on the usability of ML algorithms combined with feature selection approaches for modelling SP in advanced biological treatment systems. The limitation of the study is the construction of the model using a limited data set obtained from the process of a certain configuration. Prediction of sludge generation for wastewater plants with different capacities and treatment processes may require different ML algorithms. Future suggestions include the application of ML models using larger datasets from different treatment plants to obtain more representative and generalized predictions.

Credit author statement

Ekin Ekinci: Methodology, Software, Visualization. **Bilge Özbay**: Project administration, Conceptualization, Investigation, Writing - Review & Editing. **Sevinç İlhan Omurca**: Methodology, Software, Visualization, Conceptualization. **Fatma Ece Sayın**: Writing - Review & Editing, Visualization. **İsmail Özbay**: Conceptualization, Investigation, Visualization.

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.

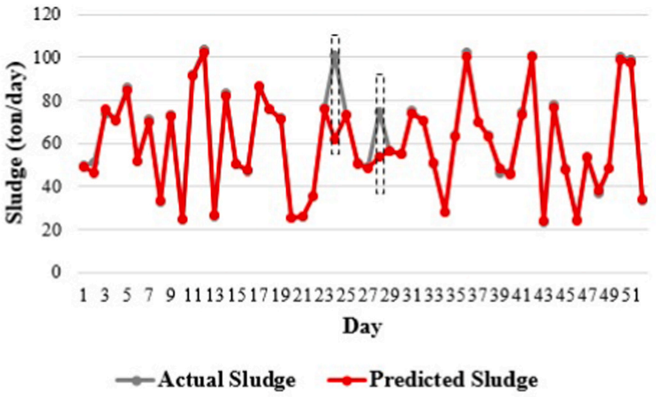

Fig. 6. Actual and predicted SP for the test set of KRR with CFS.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2023.119448.

References

Abd El-Wahab, H., Meligi, G.A., Hassaan, M.G., Lin, L., 2020. New water-based flexographic ink based on new ter-polymer nano-particles as eco-friendly binders – Part II. Pigment Resin Technol. 49, 473–482. https://doi.org/10.1108/PRT-12-2019-0111.

Atban, F., Ekinci, E., Garip, Z., 2023. Traditional machine learning algorithms for breast cancer image classification with optimized deep features. Biomed. Signal Process Control 81, 104534. https://doi.org/10.1016/j.bspc.2022.104534.

Ay, Ş., Ekinci, E., Garip, Z., 2023. A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases. J. Supercomput. 79, 11797–11826. https://doi.org/10.1007/s11227-023-05132-3.

Ba-Alawi, A.H., Heo, S.K., Aamer, H., Chang, R., Woo, T.Y., Kim, M.H., Yoo, C.K., 2023. Development of transparent high-frequency soft sensor of total nitrogen and total phosphorus concentrations in rivers using stacked convolutional auto-encoder and explainable AI. J. Water Process Eng. 53, 103661 https://doi.org/10.1016/j.jwpe.2023.103661.

Bagherzadeh, F., Mehrani, M.J., Basirifard, M., Roostaei, J., 2021. Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. J. Water Process Eng. 41, 102033 https://doi.org/10.1016/j.jwpe.2021.102033.

Battiti, R., 1994. Using mutual ınformation for selecting features in supervised neural net learning. IEEE Trans. Neural Network. 5, 537–550. https://doi.org/10.1109/72.298224.

Cho, S., Kim, H., Oh, S., Kim, K., Park, T., 2009. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. BMC Proc. 3, 3–8. https://doi.org/10.1186/1753-6561-3-s7-s25.

Ebiaredoh-Mienye, S.A., Swart, T.G., Esenogho, E., Mienye, I.D., 2022. A machine learning method with filter-based feature selection for ımproved prediction of chronic kidney disease. Bioengineering 9. https://doi.org/10.3390/bioengineering9080350.

El-Rawy, M., Abd-Ellah, M.K., Fathi, H., Ahmed, A.K.A., 2021. Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. J. Water Process Eng. 44, 102380 https://doi.org/10.1016/j.jwpe.2021.102380.

Fadlil, A., Umar, R., Sunardi Nugroho, A.S., 2022. Comparison of machine learning approach for waste bottle classification. Emerg. Sci. J. 6, 1075–1085. https://doi.org/10.28991/ESJ-2022-06-05-011.

Fan, P., Deng, R., Qiu, J., Zhao, Z., Wu, S., 2021. Well logging curve reconstruction based on kernel ridge regression. Arabian J. Geosci. 14 https://doi.org/10.1007/s12517-021-07792-y.

Friedman, J., 2001. Greedy function approximation : a gradient boosting machine. Ann. Stat. 29, 1189–1232.

Fytili, D., Zabaniotou, A., 2008. Utilization of sewage sludge in EU application of old and new methods-A review. Renew. Sustain. Energy Rev. 12, 116–140. https://doi.org/10.1016/j.rser.2006.05.014.

Gadegaonkar, S.S., Mander, Ü., Espenberg, M., 2023. A state-of-the-art review and guidelines for enhancing nitrate removal in bio-electrochemical systems (BES). J. Water Process Eng. 53, 103788 https://doi.org/10.1016/j.jwpe.2023.103788.

Hancock, J., Khoshgoftaar, T.M., 2020a. Performance of CatBoost and XGBoost in Medicare Fraud detection. Proc. - 19th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 572–579. https://doi.org/10.1109/ICMLA51294.2020.00095, 2020.

Hancock, J.T., Khoshgoftaar, T.M., 2020b. CatBoost for big data: an interdisciplinary review. J. Big Data 7. https://doi.org/10.1186/s40537-020-00369-8.

Jaffari, Z.H., Abbas, A., Lam, S.M., Park, S., Chon, K., Kim, E.S., Cho, K.H., 2023a. Machine learning approaches to predict the photocatalytic performance of bismuth ferrite-based materials in the removal of malachite green. J. Hazard Mater. 442, 130031 https://doi.org/10.1016/j.jhazmat.2022.130031.

Jaffari, Z.H., Jeong, H., Shin, J., Kwak, J., Son, C., Lee, Y.G., Kim, S., Chon, K., Hwa Cho, K., 2023b. Machine-learning-based prediction and optimization of emerging contaminants' adsorption capacity on biochar materials. Chem. Eng. J. 466, 143073 https://doi.org/10.1016/j.cej.2023.143073.

Jaffari, Z.H., Abbas, A., Umer, M., Kim, E.S., Cho, K.H., 2023c. Crystal graph convolution neural networks for fast and accurate prediction of adsorption ability of Nb2CTx towards Pb(ii) and Cd(ii) ions. J. Mater. Chem. A 11, 9009–9018. https://doi.org/10.1039/d3ta00019b.

Jing, H., Chen, Z., Wang, Xinghao, Wang, Xueqing, Ge, L., Fang, G., Xiao, D., 2023. Gradient boosting decision tree for rotor temperature estimation in permanent magnet synchronous motors. IEEE Trans. Power Electron. 38, 10617–10622. https://doi.org/10.1109/TPEL.2023.3291464.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30, 30. NIPS 2017.

Kushwaha, N.L., Rajput, J., Suna, T., Sena, D.R., Singh, D.K., Mishra, A.K., Sharma, P.K., Mani, I., 2023. Metaheuristic approaches for prediction of water quality indices with relief algorithm-based feature selection. Ecol. Inf. 75, 102122 https://doi.org/10.1016/j.ecoinf.2023.102122.

Lap, B.Q., Phan, T.T.H., Nguyen, H., Du Quang, L.X., Hang, P.T., Phi, N.Q., Hoang, V.T., Linh, P.G., Hang, B.T.T., 2023. Predicting Water Quality Index (WQI) by feature selection and machine learning: a case study of an Kim Hai irrigation system. Ecol. Inf. 74, 101991 https://doi.org/10.1016/j.ecoinf.2023.101991.

Mangalathua, S., Hwang, S.H., Jeon, J.S., 2020. Failure mode and effects analysis of RC members based on machinelearning-based SHapley Additive exPlanations (SHAP) approach. Eng. Struct. 219, 110927 https://doi.org/10.1016/j.engstruct.2020.110927.

Metcalf, L., Eddy, H.P., Tchobanoglous, G., Stensel, H.D., Tsuchihashi, R., Burton, F., 2014. Wastewater Engineering: Treatment, Disposal, and Reuse. McGraw Hill, New York, USA.

Minaravesh, B., Aydin, O., 2023. Environmental and demographic factors affecting childhood academic performance in Los Angeles county: a generalized linear elastic net regression model. Remote Sens. Appl. Soc. Environ. 30, 100942 https://doi.org/10.1016/j.rsase.2023.100942.

Najafzadeh, M., Zeinolabedini, M., 2018. Derivation of optimal equations for prediction of sewage sludge quantity using wavelet conjunction models: an environmental assessment. Environ. Sci. Pollut. Res. 25, 22931–22943. https://doi.org/10.1007/s11356-018-1975-5.

Peng, J., Sun, W., Xu, J., Zhou, G., Xie, L., Han, H., Xiao, Y., Chen, J., Li, Q., 2023. Analyzing process parameters for industrial grinding circuit based on machine learning method. Adv. Powder Technol. 34, 104113 https://doi.org/10.1016/j.apt.2023.104113.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 31, 1–23. NeurIPS 2018).

Rajora, G.L., Sanz-Bobi, M.A., Domingo, C.M., 2022. Application of machine learning methods for asset management on power distribution networks. Emerg. Sci. J. 6, 905–920. https://doi.org/10.28991/ESJ-2022-06-04-017.

Reynolds, T.D., Richards, P.A., 1996. Unit Operations and Processes in Environmental Engineering, second ed. PWS Publishing Company, Boston, USA.

Saunders, C., Alexander, G., Vovk, V., 1998. Ridge regression learning algorithm in dual variables. In: Appears in Proceedings of the 15th International Conference on Machine Learning.

Srivastava, G., Kazmi, A.A., 2021. Estimation of biomass composition and biological sludge production in SBR plants of India. Int. Journal of Sci. and Res. ISSN 2319–7064.

Urban Wastewater Treatment Regulation (UWTR), 2006. Official Gazette No.26047. Republic of Turkiye Ministry of Environment, Urbanization and Climate Change. https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=9844&MevzuatTur=7&MevzuatTertip=5. (Accessed 27 September 2023).

Varma, A.K., Thakur, L.S., Shankar, R., Mondal, P., 2019. Pyrolysis of wood sawdust: effects of process parameters on products yield and characterization of products. Waste Manag. 89, 224–235. https://doi.org/10.1016/J.WASMAN.2019.04.016.

Vinh, L.T., Lee, S., Park, Y.T., D'Auriol, B.J., 2012. A novel feature selection method based on normalized mutual information. Appl. Intell. 37, 100–120. https://doi.org/10.1007/s10489-011-0315-y.

Xia, W., Jiang, Y., Chen, X., Zhao, R., 2022. Application of machine learning algorithms in municipal solid waste management: a mini review. Waste Manag. Res. 40, 609–624. https://doi.org/10.1177/0734242X211033716.

Zeinolabedini, M., Najafzadeh, M., 2019. Comparative study of different wavelet-based neural network models to predict sewage sludge quantity in wastewater treatment plant. Environ. Monit. Assess. 191 https://doi.org/10.1007/s10661-019-7196-7.

Zeng, H., Shao, B., Dai, H., Yan, Y., Tian, N., 2023. Prediction of fluctuation loads based on GARCH family-CatBoost-CNNLSTM. Energy 263, 126125. https://doi.org/10.1016/j.energy.2022.126125.

Zhu, Xinzhe, He, M., Sun, Y., Xu, Z., Wan, Z., Hou, D., Alessi, D.S., Tsang, D.C.W., 2022. Insights into the adsorption of pharmaceuticals and personal care products (PPCPs) on biochar and activated carbon with the aid of machine learning. J. Hazard Mater. 423, 127060 https://doi.org/10.1016/j.jhazmat.2021.127060.

Zhu, Xinzhe, Liu, B., Sun, L., Li, R., Deng, H., Zhu, Xiefei, Tsang, D.C.W., 2023. Machine learning-assisted exploration for carbon neutrality potential of municipal sludge recycling via hydrothermal carbonization. Bioresour. Technol. 369, 128454 https://doi.org/10.1016/j.biortech.2022.128454.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.