University of Birjand

Water Harvesting Research Vol. 7, No. 1, Spring & Summer 2024, Original paper p. 140-150

10.22077/JWHR.2024.8007.1148

Comparative Analysis of Machine Learning Algorithms for Forecasting Effluent Chemical Oxygen Demand in Wastewater Treatment Plants

Samira Gerami[a&*], Abolfazl Akbarpour[b]

[a]MSc, Department of Civil Engineering, University of Birjand, Birjand, Iran
[b]Professor, Department of Civil Engineering, University of Birjand, Birjand, Iran

[*]Corresponding Author, E-mail address: samiragerami9397@yahoo.com

Received: 06 August 2024/ **Revised**: 09 September 2024/ **Accepted**: 15 September 2024

Abstract

Accurate prediction of wastewater effluent parameters is crucial for evaluating the performance of wastewater treatment plants, as it significantly contributes to reducing time, energy, and costs. This study employed three machine learning algorithms such as Artificial Neural Network (ANN), Support Vector Machine (SVM), and Gaussian Process Regression (GPR) in order to forecast the output COD values of Wastewater Treatment Plant No. 1 in Parkand Abad, Mashhad, Iran. The input data for the models included BOD_5, COD, TSS, Temprature, and pH of influent sewage, recorded daily from March 2018 to June 2019. The findings indicated that the SVM model surpassed the ANN and GPR models in predicting effluent COD parameters across all three phases, with GPR also performing better compared to ANN throughout the training, validation, and testing stages. The SVM model achieved values of r = 0.82, R^2 = 0.67, RMSE = 19.02, MAPE = 0.069, and MAE = 13.26 during the training phase, and the model exhibits values of r= 0.74, R^2= 0.45, RMSE=28.02, MAPE=0.080, and MAE=18.46 in the testing phase.

Keywords: Chemical Oxygen Demand, Coefficient of Determination, Machine Learning Algorithms, Wastewater Treatment Plant

1. Introduction

The rapid growth of urban development in both residential and urban areas imposes significant stress on the environment, often underestimated when compared to economic and industrial progress, particularly in developing countries (Yel and Yalpir, 2011). This escalating concern for environmental issues has led professionals to shift their focus toward the effective operation and management of Wastewater Treatment Plants (WWTPs). The inadequate performance of a WWTP can lead to severe environmental and public health issues. The discharge of effluent from these plants into receiving water bodies can contribute to the spread of various illnesses among humans (Hamed et al., 2004; Karri et al., 2021; Mjalli et al., 2007). Critical factors for assessing treatment system performance in a WWTP include parameters such as Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Suspended Solids (SS), Total Nitrogen (TN), and Total Phosphorus (TP). Furthermore, the design and operation of treatment systems play a crucial role, along with the scrutiny of wastewater discharge limits to ensure their compatibility with the receiving environment (Bekkari and Zeddouri, 2019; Türkmenler and Murat, 2017). Organic pollutants are among the major contaminants of wastewater, and COD is the most common test to estimate the concentration of organic matter in wastewater samples (Abouzari et al., 2021).

With the rise in the number and significance of treatment plants, there is an increasing demand for innovative methods to forecast and analyze pollution parameters (Türkmenler and Murat, 2017). Additionally, enhancing the safety and management of a WWTP can be accomplished through the creation of a modeling tool (Matheri et al., 2022). This tool is designed to predict plant performance by analyzing historical data related to specific key parameters linked to product quality (Hamed et al., 2004; Mjalli et al., 2007). Traditional testing methods used for assessing the operational parameters of the plant can be expensive and time-consuming, presenting obstacles to achieving efficient and effective process control (Tufaner and Demirci, 2020; Türkmenler and Murat, 2017). However, it's important to note that WWTPs encompass a range of complex physical, biological, and chemical processes. Many of these processes exhibit nonlinear behaviors that pose challenges in characterizing them using linear mathematical models (Hamed et al., 2004; Mjalli et al., 2007).

The focus of artificial intelligence (AI) techniques is primarily on forecasting diverse phenomena, whether artificial or natural, across different domains. Machine learning (ML), a subset of artificial intelligence, involves identifying unique patterns within given data to enable predictions or classifications (Bagherzadeh et al., 2021). In recent years, there has been a rapid increase in the use of artificial intelligence techniques for modeling and predicting environmental phenomena. This growth is fueled by their superior accuracy compared to mechanical models (Ye et al., 2020). These algorithms can efficiently learn complex relationships with greater effectiveness than traditional statistical methods (Khatri et al., 2020; Mohammad et al., 2020).

The ease and accuracy of predictions have driven the adoption of machine learning, particularly Artificial Neural Networks (ANNs), as a promising alternative for modeling the wastewater treatment process. This trend is further supported by advancements in computational capabilities (Bekkari and Zeddouri, 2019). ANNs utilize a series of nonlinear equations to identify complex patterns and relationships between input and output variables. As a result, they emerge as powerful and efficient tools for prediction, estimation, simulation, and classification.

Many studies have focused on modeling influent or effluent wastewater parameters. For instance, ANN models have been used to predict methane production from the digester of a biogas plant, achieving an R^2 value of 0.87 (Qdais et al., 2010). Hamed et al. (2004) employed two ANN models to forecast BOD and SS effluent concentrations for a primary WWTP in Cairo over ten months. These neural network models were trained and tested using daily datasets of BOD and SS measurements, providing accurate estimates for the datasets. Hamoda et al. (1999) evaluated the efficiency of a municipal WWTP in Ardiya using an artificial neural network backpropagation model. The findings demonstrated that ANNs serve as a versatile tool for modeling WWTPs, offering an alternative approach to predicting their performance. Rene and Saidutta (2008) also developed a Back Error Propagation (BEP) neural network to forecast the BOD5 and COD concentrations of refinery wastewater. In a separate study, Vyas et al. (2011) implemented two ANN-based models for predicting BOD at both the inlet and outlet of the Govindpura wastewater treatment plant in Bhopal. The study constructed a three-layered feedforward ANN with a backpropagation learning algorithm to make predictions for these parameters.

Oliveira-Esquerre et al. (2002) introduced a method to predict the BOD of effluent from the WWTP at RIPASA S/A Celulose e Papel in Brazil. Their work demonstrated the highest predictive accuracy by preprocessing the data with Principal Component Analysis (PCA) before inputting it into a backpropagation neural network. Additionally, Kardam et al. (2010) utilized a two-layer ANN model to forecast the efficiency of Shelled Moringa Oleifera (SMOS) in removing Cd (II) ions. This ANN model integrated Back Propagation (BP) with Principal Component Analysis to predict the sorption efficiency of SMOS for the specified metal ion. A sigmoid axon function was used for both the input and output layers, along with the Lunberg-Marquardt Algorithm (LMA). This approach resulted in a minimal Mean Squared Error (MSE) during training and cross-validation, accurate up to the ninth decimal place.

Türkmenler and Murat (2017) developed an ANN to forecast the BOD effluent of a wastewater treatment plant in Turkey. Their research demonstrated the successful application of the ANN model in accurately predicting the daily BOD levels in the effluent discharged from biological wastewater treatment plants. Manu and Thalla (2017) employed the support vector machine (SVM) and adaptive neuro-fuzzy inference system (ANFIS) models to evaluate the performance of Kjeldahl nitrogen removal in a large-scale aerobic biological WWTP in India. The findings indicated that the SVM method effectively models the biological processes in the WWTP.

Tahraoui et al. (2023) utilized Gaussian Process Regression (GPR) in conjunction with the dragonfly optimization algorithm (GPR-DA) to forecast the reduction rates of Dissolved Organic Carbon (DOC), absorbance at 254 nm (UV254), and turbidity. The study included experimental validation to compare the efficiency of the GPR-DA model with the Response Surface Methodology (RSM) model. The outcomes highlighted the superior performance of GPR-DA over RSM. Kerem and Yuce (2022) employed various regression techniques linear regression (LR), extreme gradient boosting (XGB), GPR, ridge regression (RR), Lasso regression (LASReg), and Bayesian ridge regression (BR)—to forecast the potential for recovering electrical energy from sewage sludge at the Kahramanmaraş Advanced Biological Wastewater Treatment Plant in Turkey. The study revealed that the XGB method was the most successful model for this purpose.

Nafsin and Li (2022) utilized different ML models to predict BOD_5 in the Buriganga river of Bangladesh. The results demonstrated that the best prediction model was RF-SVM with R^2 value of 0.91%. Hejabi et al. (2021) evaluated the performance of SVM and ANN models in predicting the effluent quality of Tabriz WWTP.

In this research, the operational dynamics of Wastewater Treatment Plant No. 1 in Parkand Abad, located in Mashhad, were modeled employing Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Gaussian Process Regression (GPR). These modeling techniques empower plant operators to predict the expected effluent quality by taking into account the characteristics of the input waste stream at specific locations (Hamed et al., 2004). The main objective of the study was to model and optimize the wastewater treatment process with a focus on COD output, considering variations in data across different time periods and seasonal temperature changes. The study aimed to develop machine learning models using algorithms such as ANN, SVM, and GPR to determine the most effective COD forecasting model. Daily operational data spanning from March 2018 to June 2019 under varying operational conditions were utilized. The COD_{eff} output parameter was predicted based on five input factors: pH_{inf}, TSS_{inf}, T_{inf}, BOD_{5inf}, and COD_{inf}.

2. Materials and Methods

2.1. Case study of WWTP and data description

Wastewater Treatment Plant No. 1 of Parkand Abad is situated in the northwest of Mashhad City, Iran. It serves a community of approximately 100,000 residents, managing an average daily influent flow of 17,000 cubic meter. The plant utilizes an aeration lagoon with complete mixing. It is equipped with two aeration lagoons, two sedimentation tanks, a clarifier tank, and a disinfection unit. For a visual overview of the WWTP process, refer to Fig. 1. Table 1 presents a summary of essential statistical characteristics, encompassing the minimum, mean, maximum, and standard deviation.

In this research, a database was utilized to construct predictive models for COD_{eff} values, using daily influent data of COD, BOD_5, TSS, Temperature (T), and pH as input variables, with daily COD of effluent as the target variable. The daily records detailing the qualities of influent and effluent were thoroughly studied and analyzed from March 2018 to June 2019, covering all seasonal variations for the variables under investigation. T and pH determinations were performed in the field immediately after sample collection using online sensors, while the remaining influent characteristics were recorded through sampling and analysis following standard wastewater analysis methods (Baird et al., 2017).

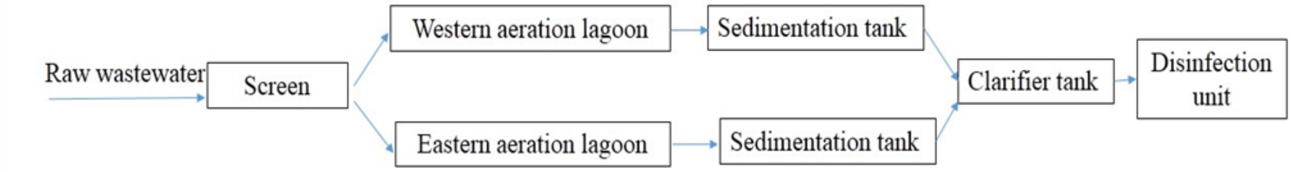

Fig. 1. Schematic of the WWTP process

Table 1. Data set statistical properties

Parameters	Unit	Min	Max	Mean	SD
Inputs model parameters					
pH_{inf}	-	7.4	8.9	7.84	0.18
T_{inf}	(^{0}C)	0	35.7	22.51	3.52
TSS_{inf}	(mg/L)	209	1192	492.12	165.4
COD_{inf}	(mg/L)	300	988	821.13	96.65
BOD_5	(mg/L)	211	560	365	61.42
Output model parameter					
COD_{eff}	(mg/L)	101	331	200.76	34.73

In this study, 3 methods ANN, SVM, and GPR were applied using Python 3 software to estimate the output COD values. The dataset was divided, allocating 75% for training, 15% for testing, and 10% for validation purposes. To enhance modeling precision, the dataset was normalized using the min-max method to scale the values within the range of 0 and 1, as shown in (Eq. 1). (Aldaghi and Javanmard, 2021; Bagherzadeh et al., 2021).

$$x_i = \frac{x_u - x_{(\min)}}{x_{(\max)} - x_{(\min)}} \quad (1)$$

where, x_i represents the standardized data value, x_u denotes the observed data, $x_{(\min)}$ is the minimum, and $x_{(\max)}$ is the maximum value within the measured data set.

2.2. Modeling approaches

2.2.1. Artificial Neural Networks

The ANN used in this study is a multilayer perceptron (MLP) that is fully interconnected, consisting of three layers: input, hidden, and output. The number of hidden layers in the network depends on the complexity of the dataset (Bagherzadeh et al., 2021). A crucial aspect of setting up a neural network model is determining both the number of hidden layers and the number of neurons within each hidden layer. The backpropagation algorithm, commonly employed for learning in multilayered feedforward networks, is used in this study. In backpropagation networks, data is processed sequentially from the input layer through the hidden layer and finally to the output layer. The goal is to optimize the weights to achieve output values that closely match the desired target values (Tosun et al., 2016; Wang et al., 2023).

The model includes 5 input neurons, corresponding to the number of inputs. It also features five hidden layers with 12, 8, 8, 8, and 20 neurons, respectively, designed to capture the complexity of the data. The number of neurons in each hidden layer was determined through a trial-and-error process to achieve the lowest error value for the model. The Selu activation function was used for the hidden layers to establish precise connections. The output layer, responsible for predicting the target variable (COD_{eff}), consists of a single neuron and uses the Relu activation function. The optimization process employed Adam's algorithm with MSE as the loss function, running for 50 epochs.

2.2.2. Support Vector Regression

The Support Vector Machine (SVM) is a cutting-edge method for classification and regression, specifically designed to address complex regression challenges. It is based on the concept of structural risk minimization, strategically mitigating overfitting by balancing the model's complexity. To handle nonlinear problems, SVM employs kernel functions, transforming them into linear counterparts within a multidimensional feature space (Bagga et al., 2023; Manu and Thalla, 2017; Wang et al., 2023) . In this study, an RBF (Radial Basis Function) kernel was used to characterize the effluent quality from Parkand Abad's Wastewater Treatment Plant No. 1 in relation to its COD_{eff} value.

2.2.3. Gaussian Process Regression

The GPR is a nonparametric Bayesian regression approach renowned for its ability to measure forecast uncertainty and effectively handle limited datasets. These models, based on nonparametric kernels, are considered probabilistic models (Kerem and Yuce, 2022). GPR comprehensively evaluates all admissible functions to fit experimental data (Ng et al., 2020). The underlying premise of GPR is that the output measurements, y are produced as Eq. 2 (Park et al., 2017).

$$y = f\big(x(k)\big) + \mathcal{E} \qquad (2)$$

where x represents the input variable measurements, f signifies the unknown functional equation, and ε denotes Gaussian noise characterized by a zero mean and variance σ_n^2. GPR employs a Gaussian Process (GP) as a prior to model the distribution on the target function (x). Within GPR, the function values $f^{1:n} = (f^1, \dots, f^n)$ corresponding to the inputs $x^{1:n} = (x^1, \dots, x^n)$ are considered stochastic variables, where $f^i = f(x^i)$. A GP is characterized as a collection of random variables forming a stochastic process, assuming that any finite set of these variables follows a joint Gaussian distribution. A GP can accurately depict the distribution of an unknown function f(x) using its mean function $m(x) = E[f(x)]$ and a kernel function $k(x, x')$ that estimates the covariance $E\big[\big(f(x) - m(x)\big)\,\big(f(x') - m(x')\big)\big]$. The kernel (covariance) function measures the geometric distance, operating on the premise that inputs closer to each other exhibit higher correlation in their functional values. The representation of the prior on the function values is depicted by (Eq. (3)) (Park et al., 2017):

$$(f^{1:n}) = GP(m(0), k(0{,}0)) \qquad (3)$$

The mean function, denoted as m(0), captures the general pattern in the target function's values, while the kernel function k(0,0) is used to estimate covariance.

Within GPR, the kernel (covariance) function reveals the underlying structure of the target function. Consequently, the type of kernel function, denoted as k(x,x'), and its parameters play a crucial role in the overall representability of the GPR model and significantly influence its predictive accuracy. A variety of kernel functions can be utilized for this purpose(Tahraoui et al., 2023). In this study, four kernel functions were employed: the squared exponential kernel, exponential kernel, Matérn 5/2, and rational quadratic kernel, as described by (Eqs. (4), (5), (6) and (7)), respectively. The function that performs best based on statistical evaluation criteria is ultimately selected.

- Squared Exponential Kernel

$$k\big(x_i, x_j \big| \theta\big) = \sigma_f^2 exp\left[\frac{1}{2}\,\frac{(x_{i-}x_{j)})^{T}\,(x_{i-}x_j)}{{\sigma_l}^2}\right] \qquad (4)$$

where σ_l represents the characteristic length scale, and σ_f is the signal standard deviation.

- Exponential Kernel

$$k\big(x_i, x_j \big| \theta\big) = \sigma_f^2 exp\left[\frac{-r}{\sigma_l}\right] \qquad (5)$$

where $r = \sqrt{(x_{i-}x_j)^T\,(x_i - x_j)}$.

- Matérn 5/2

$$k\big(x_i, x_j \big| \theta\big) = \sigma_f^2\left(1 + \frac{\sqrt{5}}{\sigma_l} r + \frac{5r^2}{3\sigma_f^2}\right) \exp\left(\frac{-\sqrt{5}}{\sigma_l} r\right) \qquad (6)$$

- Rational Quadratic Kernel

$$k\big(x_i, x_j \big| \theta\big) = \sigma_f^2 (1 + \frac{r^2}{2\alpha\sigma_l^2})^{-\alpha} \qquad (7)$$

where α represents a positive scale mixture parameter.

2.3. Model evaluation

In this study, the performance of the proposed predictive models was evaluated using several metrics: R^2, r, $RMSE$, $MAPE$ and MAE, as shown in (Eqs. (8), (9), (10), (11) and (12)), respectively (Aalami et al., 2021; An et al., 2023; Bhagat et al., 2021).

$$R^2 = 1 - \frac{\sum_i^n (\alpha_{i-}\rho_i)^2}{\sum_i^n (\alpha_{i-}\mu_\alpha)^2} \qquad (8)$$

$$r = \frac{\sum_i^n (\rho_{i-}\omega_\rho)\,(\alpha_{i-}\mu_\alpha)}{\sqrt{\sum_i^n (\alpha_{i-}\mu_\alpha)^2\, \sum_i^n (\rho_{i-}\omega_\rho)^2}} \qquad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_i^n (\alpha_{i-}\rho_i)^2} \qquad (10)$$

$$MAE = \frac{1}{n}\sum_i^n |\alpha_{i-}\rho_i| \qquad (11)$$

$$MAPE = \frac{1}{n}\left[\sum_{i}^{n}\left|\frac{\rho_i - \alpha_i}{\rho_i}\right|\right] \qquad (12)$$

The index i ranges from $1\, to\, n$, representing each observation in the dataset, where n is the total number of records. Here, α_i denotes the predicted model output, ρ_i represents the actual (real) values, ω_ρ is the mean value of the ρ values, and μ_α is the mean value of the α values.

3. Results and Discussion

3.1. Prediction results

The objective of this study is to assess and compare various machine learning models to identify the most effective one for predicting COD output. A model is considered excellent when it achieves a low RMSE and a high r value (Güçlü and Dursun, 2010). Table 2, Table 3 and Table 4 provide a summary of performance metrics for the ANN, SVM, and GPR models during the training, validation, and testing stages, respectively. These tables present the model performance criteria based on the dataset under consideration.

As shown in Table 2, the optimal ANN model exhibits the following metrics during the training stage: r = 0.7, R^2 = 0.48, RMSE = 23.82, MAPE = 0.093, and MAE = 18.40. Similarly, during the validation phase, the metrics are r = 0.67, R^2 = 0.43, RMSE = 27.64, MAPE = 0.099, and MAE = 21.14. When it comes to model testing for predicting the COD_{eff} variable, the metrics are r = 0.63, R^2 = 0.3, RMSE = 31.58, MAPE = 0.090, and MAE = 20.99.

The performance metrics for the SVM model are presented in Table3. During the training phase, the model achieved values of r = 0.82, R2 = 0.67, RMSE = 19.02, MAPE = 0.069, and MAE = 13.26. In the validation stage, the metrics were computed as r = 0.79, R2 = 0.6, RMSE = 23.07, MAPE = 0.084, and MAE = 17.70. For the testing phase, the SVM model yielded r = 0.74, R2 = 0.45, RMSE = 28.02, MAPE = 0.080, and MAE = 18.46.

Table 4 presents the coefficients (r and R^2) and performance metrics (RMSE, MAPE, and MAE) for the training, validation, and testing phases using four different kernel functions. According to the table, the Matérn 5/2 kernel function exhibited superior statistical coefficients and lower errors compared to the other three kernel functions. For the Matérn 5/2 kernel function in the training phase, the values were r = 0.79, R^2 = 0.62, RMSE = 20.34, MAPE = 0.079, and MAE = 15.55. Similarly, in the validation stage, the metrics were r = 0.71, R^2 = 0.47, RMSE = 26.78, MAPE = 0.096, and MAE = 20.52. In the testing phase, the values were r = 0.7, R^2 = 0.38, RMSE = 29.73, MAPE = 0.084, and MAE = 19.57.

Upon closer examination of Table 2, Table 3 and Table 4, it becomes evident that the SVM model exhibits lower RMSE, MAPE, and MAE values across the training, validation, and testing phases compared to the ANN and GPR models. Additionally, the SVM model demonstrates higher coefficients (r and R^2) compared to the ANN and GPR models. These findings indicate that the SVM algorithm outperformed the ANN and GPR models, offering superior accuracy and performance throughout the training, validation, and testing phases for predicting COD output. Furthermore, the performance analysis of the remaining two models reveals that the GPR model achieved higher accuracy in predictions across all three phases in terms of r, R^2, RMSE, MAPE, and MAE compared to the ANN model. The comparative performance evaluation of these models based on the metrics r and RMSE is depicted in Fig. 2 and Fig. 3 respectively.

Prediction of important parameters such as COD output of the WWTP is a topic that a large number of researchers make effort to provide different methods to increase its predicting accuracy. Antwi et al. (2018) contributed to the field of predicting COD output in WWTPs by evaluating the efficiency of COD removal in an upflow anaerobic sludge blanket (UASB) reactor. They employed a feedforward Backpropagation Artificial Neural Network (BPANN) and utilized the PCA method to select input variables. The activation functions chosen for the hidden layer and output layer were tansig and purelin, respectively. Through a comprehensive evaluation of eleven training algorithms, the Levenberg-Marquardt algorithm (LMA) was identified as the most optimal. The BPANN model demonstrated impressive performance with an R^2 value of 87%, indicating its potential for controlling and optimizing the anaerobic digestion process.

Abouzari et al. (2021) undertook a study that involved twelve regression models, including both linear and nonlinear approaches. Their objective was to identify the most efficient method for estimating COD levels in the effluent of the clarifier unit within a petrochemical WWTP. Among these models, the piece-wise linear regression with breakpoint method emerged as the most viable option. It achieved an MSE value of 0.041, an r value of 0.835, and an R^2 value of 0.694 for predicting the COD_{eff} parameter. This study underscores the precision and cost-effectiveness of mathematical and intelligent modeling as an alternative to laborious and costly laboratory tests for forecasting chemical oxygen demand levels.

Bekkari and Zeddouri (2019) employed the BPANN approach to predict the ten-month performance of the Touggourt WWTP concerning COD_{eff}. The outcomes of their study demonstrated the efficacy of the ANN model, attaining correlation coefficients of 0.89, 0.96, and 0.87 for the learning, validation, and testing stages, respectively.

Sharafati et al. (2020) utilized three ML models: Ada Boost Regression (ABR), Gradient Boost Regression (GBR), and Random Forest Regression (RFR) to forecast essential effluent characteristics such as COD. The study results showed that GBR exhibited superior performance compared to RFR and ABR, achieving an R^2 of 0.75 and an RMSE value of 9.6 mg/L. Similarly, Bagheri et al. (2016) developed MLPANN-GA and RBFANN-GA models to predict effluent COD values in a submerged membrane bioreactor. The study demonstrated remarkable accuracy in the predictions, as indicated by significantly low RMSE values and high R^2 values close to one for both models when comparing the predicted and measured COD values.

Nourani et al. (2018) employed different AI models, including FFNN, ANFIS, SVM, and MLR, to predict effluent variables such as COD_{eff}. The results indicated that the ANFIS model outperformed other models, achieving evaluation criteria for COD_{eff} modeling around 0.9 for R^2 and 0.005 for RMSE. Following ANFIS, the FFNN model performed second best, followed by SVM, with the MLR model ranking last in terms of predictive accuracy.

Granata et al. (2017) utilized SVM and regression trees to forecast effluent concentrations of BOD, COD, TSS, and TDS in a WWTP. Both models exhibited resilience, reliability, and strong generalizability. However, SVM demonstrated slightly superior performance compared to the regression tree. The analysis of existing literature indicates a lack of a universally superior model applicable across all scenarios. The effectiveness of various models seems to depend on the specific conditions of each WWTP. Therefore, there is an urgent need to develop more robust and effective models that leverage available information to better accommodate the diverse requirements of WWTPs (Bagherzadeh et al., 2021; Nourani et al., 2018).

Table 2. ANN model performance statistics

	Effluent COD		
	Training	**validation**	**Testing**
r	0.70	0.67	0.63
R^2	0.48	0.43	0.30
RMSE	23.82	27.64	31.58
MAPE	0.093	0.099	0.090
MAE	18.40	21.14	20.99

Table 3. SVM model performance statistics

	Effluent COD		
	Training	**validation**	**Testing**
r	0.82	0.79	0.74
R^2	0.67	0.60	0.45
RMSE	19.02	23.07	28.02
MAPE	0.069	0.084	0.080
MAE	13.23	17.70	18.46

Table 4. Performances of the different GPR models tested

Effluent COD															
	Training					validation					Testing				
Kernel Function	r	R^2	RMSE	MAPE	MAE	r	R^2	RMSE	MAPE	MAE	r	R^2	RMSE	MAPE	MAE
Squared Exponential Kernel	0.69	0.48	23.84	0.092	18.37	0.67	0.42	27.92	0.099	21.16	0.64	0.31	31.49	0.091	21.12
Exponential Kernel	0.79	0.60	21.07	0.079	15.80	0.69	0.42	27.90	0.099	21.32	0.68	0.33	30.98	0.090	20.95
Matern 5/2	0.79	0.62	20.34	0.079	15.55	0.71	0.47	26.78	0.096	20.52	0.70	0.38	29.73	0.084	19.57
Rational Quadratic Kernel	0.77	0.56	22.02	0.082	16.47	0.70	0.42	28.03	0.099	21.45	0.68	0.32	31.26	0.092	21.21

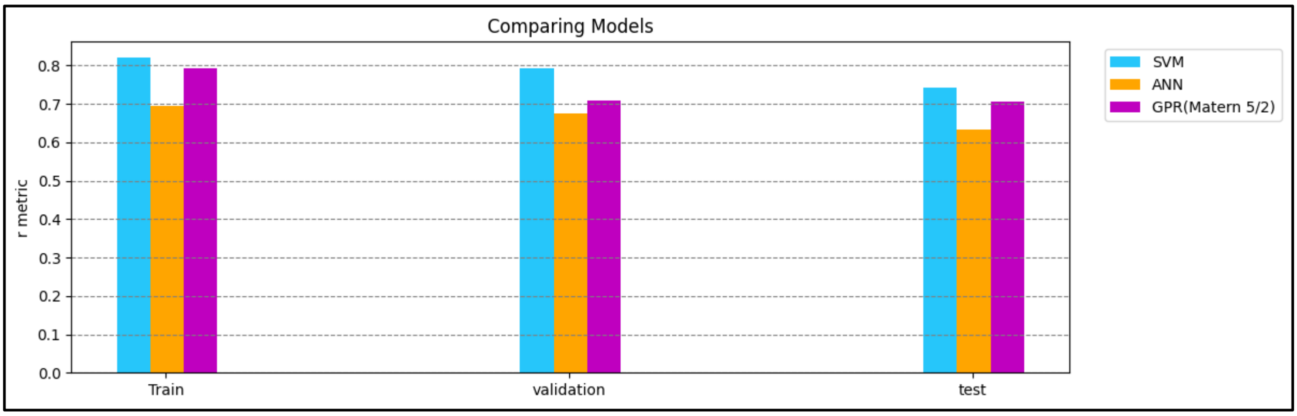

Fig. 2. Comparison of the models performance in terms of r

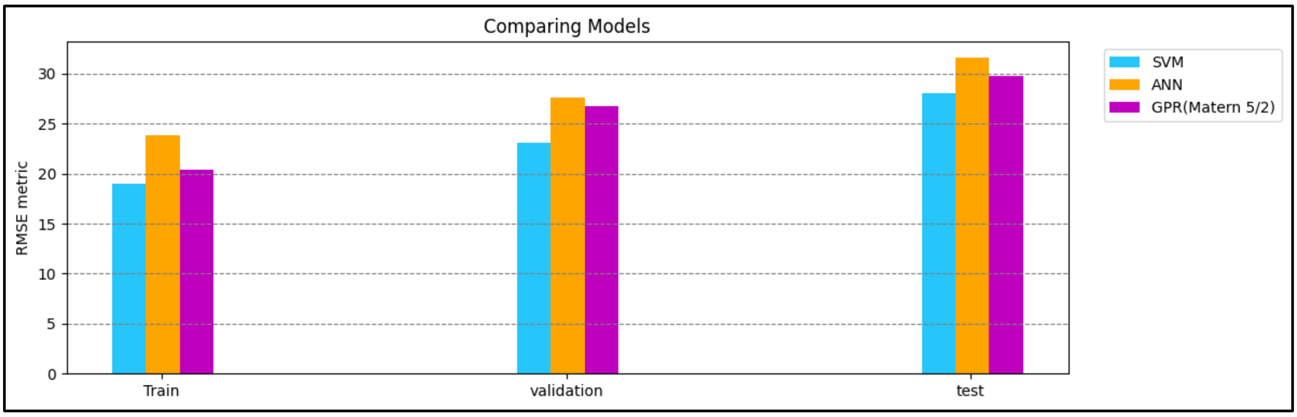

Fig. 3. Comparison of the models performance in terms of RMSE

4. Conclusion

In this study, the effectiveness of ANN, SVM, and GPR was evaluated in predicting the effluent COD parameter of WWTP No. 1 in Parkand Abad. Daily operational data from March 2018 to June 2019 were utilized for this analysis. Various metrics including r, R^2, RMSE, MAPE, and MAE were calculated to gauge the predictive performance of these machine learning algorithms. The findings indicated that SVM outperformed the other models, demonstrating superior predictive accuracy based on the evaluation criteria. Specifically, the SVM model showed higher accuracy compared to the ANN and GPR models, with GPR also performing better than ANN across all phases. Therefore, the Support Vector Machine method emerged as a robust approach for predicting COD_{eff} concentrations at any WWTP.

The models proposed in this study, particularly the SVM method, demonstrated satisfactory predictive performance compared to recent research findings. However, upon reviewing the various models presented in Tables 2, 3 and 4, it becomes apparent that their ability to predict COD values did not reach the level of accuracy seen in laboratory observations. This discrepancy may stem from the relatively limited sample size used in the study. Nevertheless, integrating a larger dataset into the systems enhances the likelihood of achieving a closer match between the modeled and observed values of COD_{eff}, thereby improving the models' suitability in terms of model evaluation criteria. Consequently, these models not only exhibit reliability in predicting various variables but also offer advantages for the downstream operations of the WWTP process (Abouzari et al., 2021).

5. Acknowledgments

The authors express their gratitude to the management of Wastewater Treatment Plant No. 1 in Parkand Abad, Mashhad, Iran for their collaboration in providing and allowing the use of the data.

6. Disclosure Statement

No potential conflict of interest was reported by the authors

7. References

Aalami, M. T., Hejabi, N., Nourani, V., & SAGHEBIAN, S. (2021). Investigation of artificial intelligence approaches capability in predicting the wastewater treatment plant performance (Case study: Tabriz wastewater treatment plant). *Amirkabir Journal of Civil Engineering, 53*(3), 1033-1048.

Abouzari, M., Pahlavani, P., Izaditame, F., & Bigdeli, B. (2021). Estimating the chemical oxygen demand of petrochemical wastewater treatment plants using linear and nonlinear statistical models–A case study. *Chemosphere, 270*, 129465.

Aldaghi, T., & Javanmard, S. (2021). The evaluation of wastewater treatment plant performance: A data mining approach. *Journal of Engineering, Design and Technology*.

An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. *Sensors, 23*(9), 4178.

Antwi, P., Li, J., Meng, J., Deng, K., Quashie, F. K., Li, J., & Boadi, P. O. (2018). Feedforward neural network model estimating pollutant removal process within mesophilic upflow anaerobic sludge blanket bioreactor treating industrial starch processing wastewater. *Bioresource technology, 257*, 102-112.

Bagga, P. J., Patel, K. M., Makhesana, M. A., Şirin, Ş., Khanna, N., Krolczyk, G. M., Pala, A. D., & Chauhan, K. C. (2023). Machine vision-based gradient-boosted tree and support vector regression for tool life prediction in turning. *The International Journal of Advanced Manufacturing Technology, 126*(1), 471-485.

Bagheri, M., Mirbagheri, S. A., Kamarkhani, A. M., & Bagheri, Z. (2016). Modeling of effluent quality parameters in a submerged membrane bioreactor with simultaneous upward and downward aeration treating municipal wastewater using hybrid models. *Desalination and Water Treatment, 57*(18), 8068-8089.

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. *Journal of Water Process Engineering, 41*, 102033.

Baird, R., Rice, E., & Eaton, A. (2017). Standard methods for the examination of water and wastewaters. *Water Environment Federation, Chair Eugene W. Rice, American Public Health Association Andrew D. Eaton, American Water Works Association*.

Bekkari, N., & Zeddouri, A. (2019). Using artificial neural network for predicting and controlling the effluent chemical oxygen demand in wastewater treatment plant. *Management of Environmental Quality: An International Journal, 30*(3), 593-608.

Bhagat, S. K., Tiyasha, T., Awadh, S. M., Tung, T. M., Jawad, A. H., & Yaseen, Z. M. (2021). Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models. *Environmental Pollution, 268*, 115663.

Granata, F., Papirio, S., Esposito, G., Gargano, R., & De Marinis, G. (2017). Machine learning algorithms for the forecasting of wastewater quality indicators. *Water, 9*(2), 105.

Güçlü, D., & Dursun, Ş. (2010). Artificial neural network modelling of a large-scale wastewater treatment plant operation. *Bioprocess and biosystems engineering, 33*, 1051-1058.

Hamed, M. M., Khalafallah, M. G., & Hassanien, E. A. (2004). Prediction of wastewater treatment plant performance using artificial neural networks. *Environmental Modelling & Software, 19*(10), 919-928.

Hamoda, M. F., Al-Ghusain, I. A., & Hassan, A. H. (1999). Integrated wastewater treatment plant performance evaluation using artificial neural networks. *Water Science and Technology, 40*(7), 55-65.

Hejabi, N., Saghebian, S. M., Aalami, M. T., & Nourani, V. (2021). Evaluation of the effluent quality parameters of wastewater treatment plant based on uncertainty analysis and post-processing approaches (case study). *Water Science and Technology, 83*(7), 1633-1648.

Kardam, A., Raj, K. R., Arora, J. K., Srivastava, M. M., & Srivastava, S. (2010). Artificial neural network modeling for sorption of cadmium from aqueous system by shelled Moringa oleifera seed powder as an agricultural waste. *Journal of Water Resource and Protection, 2*(4), 339.

Karri, R. R., Ravindran, G., & Dehghani, M. H. (2021). Wastewater—sources, toxicity, and their consequences to human health. In *Soft computing techniques in solid waste and wastewater management* (pp. 3-33). Elsevier.

Kerem, A., & Yuce, E. (2022). Electrical energy recovery from wastewater: prediction with machine learning algorithms. *Environmental Science and Pollution Research*, 1-14.

Khatri, N., Khatri, K. K., & Sharma, A. (2020). Artificial neural network modelling of faecal coliform removal in an intermittent cycle extended aeration system-sequential batch reactor based wastewater treatment plant. *Journal of Water Process Engineering, 37*, 101477.

Manu, D., & Thalla, A. K. (2017). Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the removal of Kjeldahl Nitrogen from wastewater. *Applied Water Science, 7*, 3783-3791.

Matheri, A. N., Mohamed, B., Ntuli, F., Nabadda, E., & Ngila, J. C. (2022). Sustainable circularity and intelligent data-driven operations and control of the wastewater treatment plant. *Physics and Chemistry of the Earth, Parts A/B/C, 126*, 103152.

Mjalli, F. S., Al-Asheh, S., & Alfadala, H. (2007). Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. *Journal of Environmental Management, 83*(3), 329-338.

Mohammad, A. T., Al-Obaidi, M. A., Hameed, E. M., Basheer, B. N., & Mujtaba, I. M. (2020). Modelling the chlorophenol removal from wastewater via reverse osmosis process using a multilayer artificial neural network with genetic algorithm. *Journal of Water Process Engineering, 33*, 100993.

Nafsin, N., & Li, J. (2022). Prediction of 5-day biochemical oxygen demand in the Buriganga River of Bangladesh using novel hybrid machine learning algorithms. *Water Environment Research, 94*(5), e10718.

Ng, K. H., Gan, Y., Cheng, C. K., Liu, K.-H., & Liong, S.-T. (2020). Integration of machine learning-based prediction for enhanced Model's generalization: Application in photocatalytic polishing of palm oil mill effluent (POME). *Environmental Pollution, 267*, 115500.

Nourani, V., Elkiran, G., & Abba, S. (2018). Wastewater treatment plant performance analysis using artificial intelligence–an ensemble approach. *Water Science and Technology, 78*(10), 2064-2076.

Oliveira-Esquerre, K., Mori, M., & Bruns, R. E. (2002). Simulation of an industrial wastewater treatment plant using artificial neural networks and principal components analysis. *Brazilian Journal of Chemical Engineering, 19*, 365-370.

Park, J., Lechevalier, D., Ak, R., Ferguson, M., Law, K. H., Lee, Y.-T., & Rachuri, S. (2017). Gaussian process regression (GPR) representation in predictive model markup language (PMML). *Smart and sustainable manufacturing systems, 1*(1), 121.

Qdais, H. A., Hani, K. B., & Shatnawi, N. (2010). Modeling and optimization of biogas production from a waste digester using artificial neural network and genetic algorithm. *Resources, Conservation and Recycling, 54*(6), 359-363.

Rene, E. R., & Saidutta, M. (2008). Prediction of BOD and COD of a refinery wastewater using multilayer artificial neural networks. *Journal of Urban and Environmental Engineering, 2*(1), 1-7.

Sharafati, A., Asadollah, S. B. H. S., & Hosseinzadeh, M. (2020). The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. *Process Safety and Environmental Protection, 140*, 68-78.

Tahraoui, H., Belhadj, A.-E., Triki, Z., Boudellal, N. R., Seder, S., Amrane, A., Zhang, J., Moula, N., Tifoura, A., & Ferhat, R. (2023). Mixed coagulant-flocculant optimization for pharmaceutical effluent pretreatment using response surface methodology and Gaussian process regression. *Process Safety and Environmental Protection, 169*, 909-927.

Tosun, E., Aydin, K., & Bilgili, M. (2016). Comparison of linear regression and artificial neural network model of a diesel engine fueled with biodiesel-alcohol mixtures. *Alexandria Engineering Journal, 55*(4), 3081-3089.

Tufaner, F., & Demirci, Y. (2020). Prediction of biogas production rate from anaerobic hybrid reactor by artificial neural network and nonlinear regressions models. *Clean Technologies and Environmental Policy, 22*, 713-724.

Türkmenler, H., & Murat, P. (2017). Performance assessment of advanced biological wastewater treatment plants using artificial neural networks. *International Journal of Engineering Technologies IJET, 3*(3), 151-156.

Vyas, M., Modhera, B., Vyas, V., & Sharma, A. (2011). Performance forecasting of common effluent treatment plant parameters by artificial neural network. *ARPN Journal of Engineering and Applied Sciences*, *6*(1), 38-42.

Wang, Y., Cheng, Y., Liu, H., Guo, Q., Dai, C., Zhao, M., & Liu, D. (2023). A review on applications of artificial intelligence in wastewater treatment. *Sustainability*, *15*(18), 13557.

Ye, Z., Yang, J., Zhong, N., Tu, X., Jia, J., & Wang, J. (2020). Tackling environmental challenges in pollution controls using artificial intelligence: A review. *Science of the Total Environment*, *699*, 134279.

Yel, E., & Yalpir, S. (2011). Prediction of primary treatment effluent parameters by Fuzzy Inference System (FIS) approach. *procedia computer science, 3,* 659-665.