International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 14, No.1 (Oct-23)

http://dx.doi.org/10.12785/ijcds/1401113

A Comparative Study to Forecast the Total Nitrogen Effluent Concentration in a Wastewater Treatment Plant Using Machine Learning Techniques

Fuad Ahmad Musleh

Department of Civil Engineering, College of Engineering, University of Bahrain, Isa Town, Bahrain

Received 9 Sep. 2023, Revised 19 Oct. 2023, Accepted 27 Oct. 2023, Published 31 Oct. 2023

Abstract: With the global population increasing and water scarcity becoming a pressing issue worldwide, wastewater treatment has emerged as a crucial solution to meet growing water demands. Wastewater treatment plants (WWTPs) play a vital role in this regard, and the integration of new technologies, such as Machine Learning (ML), holds immense potential for their optimization. This study focuses on evaluating and comparing the performance of four ML regressors - Light Gradient Boosting regressor (LGBM), Random Forest regressor (RF), Support Vector Regressor (SVR), and Ridge Regression - in predicting Total Nitrogen (TN) concentration in a WWTP. The results indicate that the Random Forest regressor outperformed the other algorithms, demonstrating superior performance in correlation coefficient (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). These findings highlight the efficacy of the Random Forest regressor as a valuable tool for accurate TN concentration prediction in WWTPs. By leveraging ML techniques, WWTPs can enhance operational efficiency and contribute to sustainable water management, addressing the global challenge of water scarcity.

Keywords: *Artificial Intelligence, Wastewater Treatment Plant, Machine Learning Regressors, Total Nitrogen*

1. INTRODUCTION

Due to global water scarcity, the utilization of wastewater in various sectors has become essential. However, to ensure the preservation of the water environment, it is crucial to maintain high-quality standards for recycled water. This emphasizes the importance of optimizing WWTPs. WWTPs are intricate systems involving a complex interplay of biological, physical, and chemical reactions and processes. Their primary objective is to reduce pollutants like Chemical Oxygen Demand (COD), total suspended solids (TSS), Biological Oxygen Demand (BOD), and pathogenic organisms, as well as remove nutrients like Ammonia (NH3) and phosphorus[1]. Despite the high variability in influent characteristics and discharge quantities experienced in WWTPs on different time scales, it is imperative to uphold consistent performance in effluent discharge[2]. One significant pollutant in wastewater is nitrogen, and its concentration must be reduced to meet specific standards before wastewater is discharged into the environment. Accordingly, predicting the concentration of Total Nitrogen (TN), which comprises nitrate, ammonia, nitrite, and organically bonded nitrogen, in influent wastewater treatment plants is of utmost importance[3].

ML, a sector of Artificial Intelligence (AI), has been distinguished as a powerful tool for this prediction task, offering high accuracy while reducing the time, sampling requirements, energy, and costs compared to mechanistic models[4]. ML algorithms are capable of efficiently handling complex relationships and procedures, surpassing traditional statistical methods. Several ML algorithms, such as fuzzy logic (FL), model trees, Artificial Neural Network (ANN), and hybrid intelligent techniques, have contributed to the field of wastewater treatment[5].

This study aims to measure and compare the predictive capabilities of four regression algorithms—Ridge, SVR, LGBM, and RF in predicting TN concentration in WWTPs. The analysis will utilize the Full-Scale Wastewater Treatment Plant Dataset obtained from Kaggle and follow the six-stage Cross-Industry Standard Protocol for Data Mining (CRISP-DM) methodology. The subsequent sections will cover a literature review, methodology, dataset description and preparation,

E-mail:fuadakm@hotmail.com

prediction algorithms, findings, and a discussion and conclusion.

2. Literature Review

Wastewater treatment plants were established to precisely conduct the water treatment process, the success of the procedure is based on effluent discharge standards and the influent's water quality. There is much research that has studied the WWTPs using artificial intelligence techniques.

In this research performed by Aghdam, et al. [6] the prediction of BOD5 and COD levels in wastewater involved training various models, including Gene Expression Programming (GEP), multilayer perception neural networks, multi-linear regression, k-nearest neighbors, gradient boosting, and regression trees. Monthly data from the inflow of seven wastewater treatment plants in Hong Kong over a three-year period formed the basis of the training dataset. Remarkably, the GEP model demonstrated superior accuracy, yielding R^2 values of 0.784 for BOD5 and 0.861 for COD. Sensitivity analysis conducted through Monte Carlo simulation unveiled TSS (Total Suspended Solids) as the primary influencing factor for both BOD5 and COD levels, with a 10% increase in TSS leading to a 7.94% increase in BOD5 and a 7.92% increase in COD. The modeling results obtained from GEP align well with the underlying chemistry of wastewater quality and offer potential for broader application to other sewage sources, such as industrial wastewater and leachate.

To start, novel methods from feature selection methods namely stepwise selection, genetic algorithms, and forward selection were performed by Tomperi, et al. [7] . This was followed by establishing a k-fold model for the forecasting of TN concentration. A R^2=0.69 was achieved by the model. Additionally, in this study performed by Xu, et al. [8] , Deep Learning (DL) and ML models were applied to forecast effluent phosphorus levels using data spanning nine years from a small-scale WWTP. Pearson correlation analysis has been employed to identify 42 variables, revealing internal correlations among them. Initially, five ML relevant input features from a pool of regression models were employed, the support vector machine model achieved the highest coefficient of determination (R^2) of 0.637 for predicting effluent phosphorus load. Subsequently, long short-term memory a DL model successfully predicted phosphorus load a day in advance, yielding an R^2 reading equal to 0.496. Lastly, an anomaly alarm system was proposed, based on historical data to diminish the number of permit violations, achieving a maximum accuracy of 79.7% by comparing seven ML classification models to predict phosphorus concentration.

A novel integrated model with a rolling decomposition method for predicting influent ammonia nitrogen (NH3-N) in wastewater treatment was proposed in this study by Yan, et al. [9] . The model surpasses the performance of individual GRU models, exhibiting substantial reductions of 16.69% in RMSE, 13.02% in MAE, and 11.90% in MAPE. When compared to an integrated model trained with information leakage, the proposed model demonstrates significant improvements, achieving reductions of 42.34% in RMSE, 41.06% in MAE, and 39.34% in MAPE. These findings underscore the enhanced accuracy and reliability of the integrated model, providing valuable insights for intelligent wastewater treatment and the development of sustainable urban environments.

The study conducted by Sadri Moghaddam and Mesghali [10] introduces novel hybrid modeling and optimization tools, where to forecast TN in treated wastewater from the Southern Tehran Wastewater Treatment Plant (STWWTP) an integration between KNN, SVR, DT, and RF algorithms with Bayesian optimization algorithm (BOP), Ensemble models, such as voting average and stacked regression, were employed to enhance predictions. Achieving a superior performance with an R^2 of 0.640, RMSE of 2.378, and MAE of 1.838 on the test data the hybrid ensemble model using KNN-BOP and SVR-BOP proved to be the most optimal. This accurate prediction model can provide early warnings about eutrophication-related water pollution caused by total nitrogen concentration.

A ML algorithms based intelligent tool for optimizing sewage sludge (SS) disposal was introduced in this study by Adibimanesh, et al. [11] Three ML models (Artificial Neural Network (ANN), Parallel, and Chained) were implemented using the SciKit-Learn library in Python. To predict optimal changing variables for system outputs, simulation data from ASPEN PLUS software was utilized by the optimizer. Validation using data from a WWTP in Gdynia, Poland, demonstrated the enhanced forecasting ability of the ML models, with R^2 values of 0.85, 0.94, and 0.91 for models A, B, and C, respectively. The optimized approach achieved approximately 6% savings in energy consumption for SS incineration, contributing to addressing the energy crisis and reducing costs.

The study carried out by Manzo, et al. [12] examined the performance of a WWTP in Esquel, Patagonia, over a two-year period. The impact of climatic conditions on nutrient dissipation, suspended solids, and dissolved oxygen was examined at six sampling points. The results revealed that climatic variables, such as rainfall patterns and air temperature, influenced the WWTP's functioning and efficiency in mitigating nutrient loads and sediment retention. Nitrate loads were extremely higher in 2018, advocating operational failures, while ammonia levels remained consistently high throughout both years. The WWTP showed moderate success (36%) in reducing.

suspended solids in 2018 but was inefficient in 2019. Effluent nutrient levels exceeded regulatory limits, particularly during summer, threatening the ecological integrity of the receiving stream.

An exogenous input, was utilized along with a dynamic nonlinear autoregressive network by Yang, et al. [13] along with two models which were used for predicting the effluent quality, these are the static ANN and input PCA-NARX hybrid model. PCA-NARX achieved impressive results of (RMSETN = 0.8 mg/L and RMSECOD = 2.9 mg/L) when predicting total hydrogen and the effluent COD. Moving forward, in this study, the presence and removal efficiencies (REs) of 17 psychoactive drug residues were investigated in Slovene municipal wastewater treatment plants by Abushammala, et al. [14]. In influent, effluent, and receiving rivers drug residues were detected. REs varied among different drugs and treatment technologies, with THC-COOH, nicotine, amphetamine and cocaine residues showing the highest REs (>90%) and methadone residues the lowest (<30%). The moving biofilm bed reactor (MBBR) had lower removal rates for certain drugs. While seasonal changes in drug residue levels and Res were observed, no consistent pattern emerged. In silico predictions indicated potential effects on aquatic plants, and environmental risk assessment identified risks associated with nicotine, methadone, EDDP, morphine, and MDMA. Regular monitoring and regulatory actions are recommended to protect aquatic organisms.

Bagherzadeh, et al. [15] proposed a novel approach to improve the prediction of the concentration of TN. Therefore, wrapper filter, and embedded which are different FS methods were utilized; moreover, three ML methods ANN, GBM, and RF were used, and their performance was differentiated in terms of RMSE, R^2, and MAR. The findings showed the superiority of Mutual Information in FS and a more efficient performance by GBM and RF when compared to ANN.

A new methodology was proposed by Lotfi, et al. [16] in this research. This methodology is based on combining nonlinear outlier robust extreme learning Machine technique (ORELM) and linear stochastic model (ARIMA) to forecast the concentration of both BOD and COD in the influents and effluents of WWTPs. The hybrid models improved the prediction capabilities by attaining an effluent R^2 of 99%.

A two-layered stacked (LSTM) network was utilized for the prediction of TN by Yaqub, et al. [17]. A dataset consisting of 1876 testing and 6000 training attributes was used; a low value was attained for the average model error as MSE equaled 0.015.Zhou, et al. [18] suggested a novel approach utilizing hybrid simulation for the precise prediction of TN. In this approach a parallel-serial hybrid model consisting of both mechanistic and ML models was utilized on a dataset with 400 attributes for train and another 250 for test. The results attained indicate that the previously mentioned combination results in high accuracy as a R^2 of 0.81 was achieved; additionally, comparing this combination to previous recent studies with standalone ML models proves the superiority of this combination for predicting the concentration of TN.

For the purpose of forecasting the concentration of effluent COD, Liu, et al. [19] suggested a model comparing the LSTM NN model based on an attention mechanism (AM) and adaptive hybrid mutation particle swarm optimization (AHMPSO) with an optimized model composed of LSTM, LSTM-AM, and PSO-LSTM-AM. Results revealed a decrease of 8.993%-25.996%, 7.803%-19.499%, 9.669%-27.551%, and 3.313%-11.229% in the MAPE, RMSE, MAE, and the R^2 respectively.

Matheri, et al. [20] used a forecasting model dependent on AI and ANN utilizing MATLAB as a platform to study the relationship between trace metals and COD. The results achieved showed that ANN was more efficient with mean squared error (MSE) of *2.7059e*-14 to 2.3175e-15, R^2 of 0.98–0.99, RMSE of 0.0049–0.8673 and sum of square error (SSE) of 0.00029–0.1598.

The efficiency of removing the BOD, suspended solids (TSS), sulfide, and COD at El-Berka WWTP in Egypt is forecasted in this study performed by El-Rawy, et al. [21] using two different methods. In the first method multiple models were used, these are deep cascade-forward backpropagation (DCB), deep feed-forward backpropagation (DFB), and the traditional feed-forward (TF). These models were found to be effective in the prediction process; moreover, DCB was found to be the most accurate.

The aim of the study performed by Wijaya and Oh [22] was to enhance the understanding of keystone taxa in membrane bioreactor (MBR) wastewater treatment systems. Based on microbiome data a ML model was evolved to forecast for a full-scale MBR systems its operational characteristics, achieving an average accuracy of >91.6%. Ferruginibacter was identified as a key organism in the MBR system, responsible for metabolizing complex organic polymers. Through ML regression modeling, significant temporal patterns of Ferruginibacter in response to water temperature were discovered. This ML approach provides valuable insights into the complex ecological interactions of crucial taxa, enabling the implementation of sustainable and predictive management strategies for MBR systems.

In this study, Toffanin, et al. [23] suggest the implementation of LSTM network to design an oxygen concentration controller for the activated sludge process in wastewater treatment. The LSTM model is developed using data collected from a real plant over nearly a month. A comparative analysis is conducted between the performance of the LSTM model and a standard AutoRegressive model with eXogenous input (ARX). Both models effectively capture the oscillation frequencies and overall behavior, but the LSTM model exhibits enhanced accuracy specifically in terms of amplitude with a fitting index of 60.56% compared to 41.20% for ARX. The LSTM model shows satisfactory performance for oxygen concentration control.

3. Research Methodology and Approach

A. Back Ground of the Research Study

The study utilized the Google Colab platform as the foundation for conducting the research. The programming phase involved employing Python libraries such as LuciferML and Scikit-learn, both known for their capabilities in ML. Four distinct ML techniques, namely Ridge, SVR, LGBM regressor, and RF regressor, were applied to the dataset for analysis purposes. Additionally, the study adhered to the six-phase methodology known (CRISP-DM), ensuring a systematic and structured approach to the research process [24].

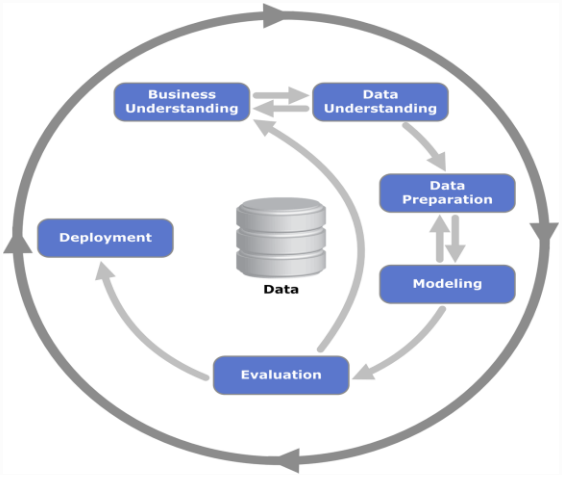

Figure 1. Phases of the CRISP-DM Methodology.

B. Data Set Description

The study made use of the Full-Scale Wastewater Treatment Plant Dataset sourced from Kaggle [25]. This dataset is a result of merging open access data from the Melbourne Airport weather station and Melbourne water. The merging process was performed based on the record date column, combining power consumption, hydraulic, biological, and climate data. The dataset comprises 1383 instances, featuring 19 variables and a single target variable.

A detailed summary of the variables can be found in Table 1. Furthermore, Figure 2 visually represents the annually gathered data spanning from 2014 to 2019. Through this dataset, the study gained valuable insights for analysis and exploration within the wastewater treatment domain.

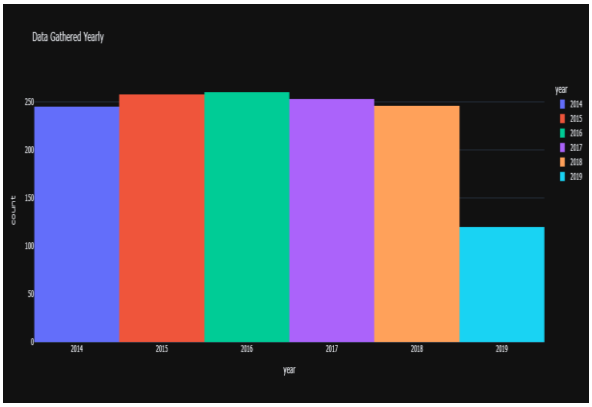

Figure 2. Data Gathered Yearly from 2014 till 2019.

TABLE I. DATASET DESCRIPTION

Attributes	Data Type
Avg_outflow (AO)	Float64
Avg_inflow (AI)	Float64
Energy Consumption (EC)	Int64
Ammonia (NH3)	Float64
Biological Oxygen Demand (BOD)	Float64
Average Temperature (AT)	Float64
Maximum Temperature (MaxT)	Float64
Minimum Temperature (MinT)	Float64
Atmospheric Pressure (AP)	Float64
Average Humidity (AH)	Int64
Total Rainfall (TR)	Float64
Average Visibility (AV)	Float64
Average Wind Speed (AWS)	Float64
Maximum Wind Speed (MWS)	Float64
Chemical Oxygen Demand (COD)	Float64
Total Nitrogen (TN)	Float64

C. Correlation Matrix

The Heatmap Correlation matrix is a statistical technique employed to visualize the interrelationships and dependencies among features [4]. In Figure 3, it is evident that the dataset's features exhibit a negative correlation, indicating an inverse association between them. Consequently, all the features were included in this study to comprehensively investigate their mutual influences and effects. By considering the entirety of the features, a thorough examination of their interplay and impact can be conducted, leading to more comprehensive insights and findings.

.

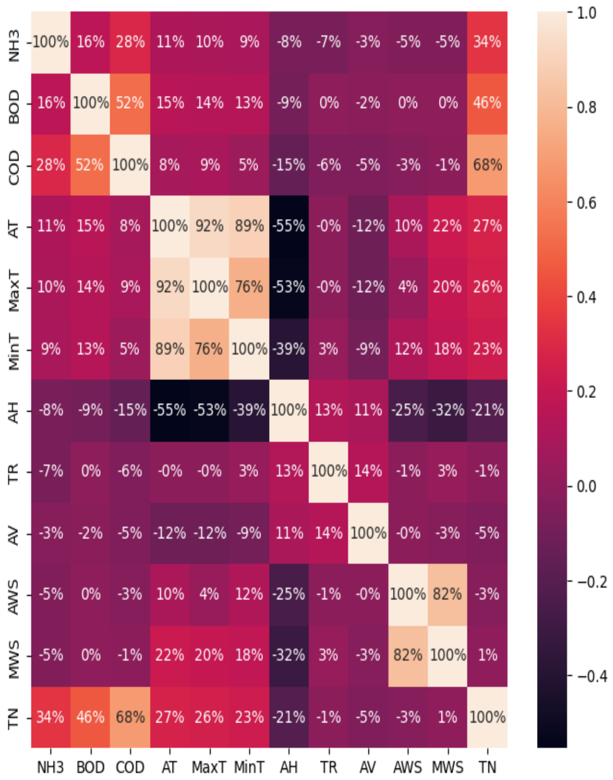

Figure 3. Heatmap Correlation Matrix.

A positive correlation was distinguished between TN which is the target and the features COD, BOD, NH3, MaxT, MinT, AT with correlation percentages of 68%, 46%, 34%, 27%, 26%, and 23%, respectively.

D. Data Preparation.

Following the data exploration stage, the data preparation phase commenced, involving several crucial steps. These steps encompassed addressing missing data, performing data scalarization, conducting feature selection, and finally, splitting the dataset.

- Missing Data

The dataset was clean without any missing value.

- Feature Selection (FS)

FS is a process were redundant and irrelevant attributes are discarded; lowering the number of attributes will enhance the predictive capabilities of the model [26]. The suggested models manipulate 1383 instances of WWTP dataset with 20 features among which only 11 features were selected including EC, NH3, BOD, TR, MaxT, MinT, AH, AV, AWS, MWS, COD, and TN concentration using the SelectPercentile method.

- Data Scalarization

The MinmaxScaler function constricts the data into a specified range, most commonly laying inclusively between 0 and 1, without altering the underlying distribution since the differences are also scaled. It ensures that the values are scaled to a specific range while preserving the original shape of the data [27].

- Splitting Data

The researchers initially divided the data into two group categories creating a 4 to 1 split ratio, allocating 80% of the data for training and reserving the remaining 20% for the testing procedure.

E. Modelling

The four ML algorithms Ridge, SVR, LGBM regressor, and RF regressor are implemented to Total Nitrogen Prediction.

RF regressor: RF regressor, put forward by Breiman in the year 2001, is an enhanced version of the bagging algorithm. It is a homogeneous, bagging-type ensemble method that utilizes decision trees. RF employs bootstrapped samples of data and arbitrarily chosen subsets of features to construct multiple decision trees, which are then combined through a voting mechanism. The main objective of RF is to improve predictive accuracy while mitigating overfitting through the use of averaging. In contrast to simple bagging, RF introduces a key modification: at each split in a tree, a random subset of variables is considered instead of all input variables. This random feature selection strategy promotes greater diversity and reduces the risk of relying too heavily on specific subsets of features.

RF algorithm commences by creating bootstrap samples of the dataset, from which individual decision trees are built. Notably, RF applies the bootstrap sampling technique to both cases and features (input variables), further enhancing the ensemble's diversity and robustness. The construction of RF models involves tuning parameters such as the number of cases and variables considered at each split and the number of trees to build. By carefully adjusting these parameters, RF models can yield highly accurate predictions, often surpassing the performance of simple bagging and boosting methods. In summary, the RF regressor is a powerful ensemble method that leverages decision trees, bootstrapped samples, and random feature selection to improve predictive accuracy while controlling overfitting, making it a valuable tool in ML and data analysis. (i.e., AdaBoost) [28].

SVR is a regression algorithm capable of handling both linear and non-linear regression problems. It operates based on the principles of Support Vector Machines (SVM), which are a family of generalized linear models. SVR makes regression decisions by evaluating the value of a linear combination of input features. Support vector machines have gained popularity in the field of machine learning due to their strong predictive capabilities and solid theoretical foundation. SVMs are a type of supervised learning technique that constructs input-output functions using labeled training data. These functions can serve as classifiers, assigning cases into predefined classes, or regressors, estimating continuous numerical values for desired outputs [29].

Ridge Regression: This approach for parameter estimation is widely recognized and employed. It analyses any data with collinearity problem, where the independent variables are highly correlated, usually arising in multiple linear regression with unbiased least-squares and large variances which result in predicted values far away from the actual values. This technique can be characterized as a form of regularization used to address regression models. In this method, the loss function corresponds to the linear least square's function, while the regularization component is determined by the l2-norm. Ridge regression has been used in many fields such as econometrics, chemistry, and engineering [30].

LGBM is a highly popular algorithmic framework for gradient boosting that utilizes tree-based learning techniques. It operates in a manner like AdaBoost, where predictors are added sequentially to an ensemble, each correcting the mistakes made by its predecessor. Unlike AdaBoost, which adjusts instance weights, Gradient Boosting in LGBM focuses on fitting the new predictor to the residual errors of the previous predictor. The fundamental concept of boosting involves the gradual introduction of new models to the ensemble. During each iteration, a new weak base learner model is trained by considering the errors of the entire ensemble. In the case of gradient boosting, the learning process progressively fits new models to provide a more precise estimation of the class variable. To minimize the loss function, each new model is aligned with the negative gradient of the system, employing the gradient descent method [31].

The performance precision of four ML algorithms was carried out based on RMSE, MAE, and R^2. The models' performance is measured using three evaluation metrics: RMSE, MAE, and R^2. RMSE and MAE both gauge the disparity between two vectors: the vector of predictions and the vector of target values. The selection of the norm index determines the emphasis on larger or smaller values. Thus, RMSE is more responsive to outliers than MAE. However, in situations where outliers are exceptionally infrequent, like in a bell-shaped curve, RMSE performs admirably and is typically the preferred choice.

Root Mean Square Error: It is a commonly employed performance measure in regression models. It offers valuable information about the typical number of errors made by the system in its predictions, with a greater emphasis placed on larger errors. RMSE provides a measure of the sample standard deviation, illustrating the variations between the predicted and observed values. By accounting for both the magnitude and direction of errors, RMSE provides a comprehensive assessment of the model's predictive accuracy [32]. This is illustrated in Equation 1.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - P_i)^2} \tag{1}$$

***Mean Absolute Error*:** It is a widely utilized metric in regression analyses. It quantifies the average absolute difference between the predicted value and the actual value, normalized by the total number of data points. MAE provides a measure of the average magnitude of errors, disregarding their direction. By assessing the average deviation between predictions and actual values, MAE offers valuable insights into the overall effectiveness of the model [33]. This is illustrated illustrated in Equation 2.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - P_i| \tag{2}$$

Correlation Coefficient: It is a statistical metric that provides insight into the degree to which the relationship between an independent variable and a dependent variable accounts for the variability observed in the data [34]. This is illustrated in Equation 3.

$$R^2 = 1 - \frac{\sum(a_i - P_i)^2}{\left(a_i - \mu_a\right)^2} \tag{3}$$

Where n is the overall readings count and $i = 1,2,..n$ is the number of current readings, and n is the total number of records. Considering μ_a as mean value a_i for output, and p_i as real value.

4. Results and Discussion

In this study, the RF regressor had the maximum accuracy with a RMSE=3.48, R^2 =0.64, and MAE=0.92 for the test dataset followed by LGBM, SVR and Ridge. as shown in Table III.

TABLE II. PERFORMANCE COMPARISON BETWEEN FOUR REGRESSORS

Models	R^2	MAE	RMSE
Ridge Regression	0.55	1.24	4.35
SVR	0.53	1.19	4.54
LGBM regressor	0.60	1.16	3.80
RF Regressor	0.64	0.92	3.48

The R^2 readings obtained for the different regression algorithms were as follows: Ridge Regression (0.55), SVR (0.53), LGBM Regressor (0.60), and RF Regressor (0.64) as shown in Figure.4. These results indicate the goodness of fit or the proportion of variance explained by each respective model. The higher R^2 values for LGBM Regressor and RF Regressor suggest that these models have a better ability to capture and predict the outcome variable compared to Ridge Regression and SVR. Therefore, LGBM Regressor and RF Regressor exhibit stronger performance in explaining the correlation between the attributes and the outcome parameter in the dataset.

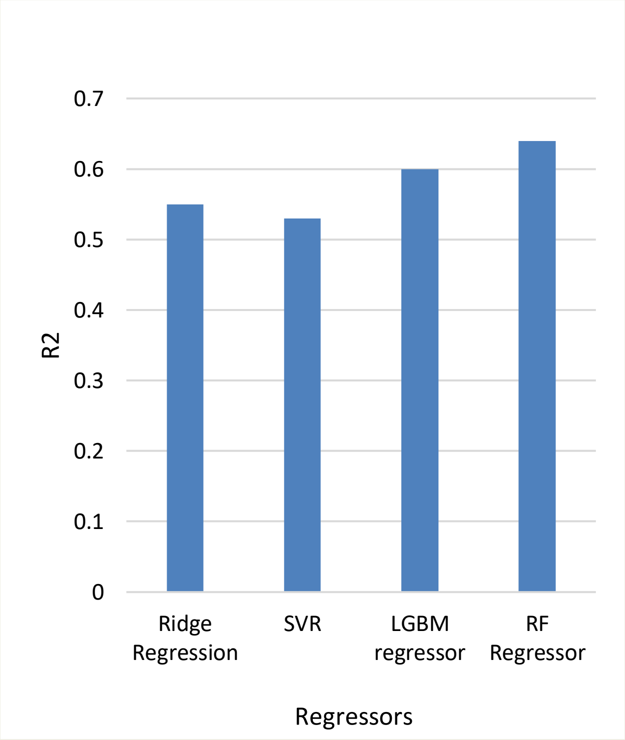

Figure 4. R^2 plot for the proposed Algorithms

The RMSE values for the different regression models were as follows: Ridge Regression (4.35), SVR (4.54), LGBM Regressor (3.80), and RF Regressor (3.48) as shown in Figure.5. These results suggest that the RF and LGBM regressors Regressor models achieved lower RMSE values, indicating better accuracy in predicting the outcome variable compared to Ridge Regression and SVR. Thus, RF Regressor and LGBM Regressor models exhibit stronger performance in terms of minimizing prediction errors.

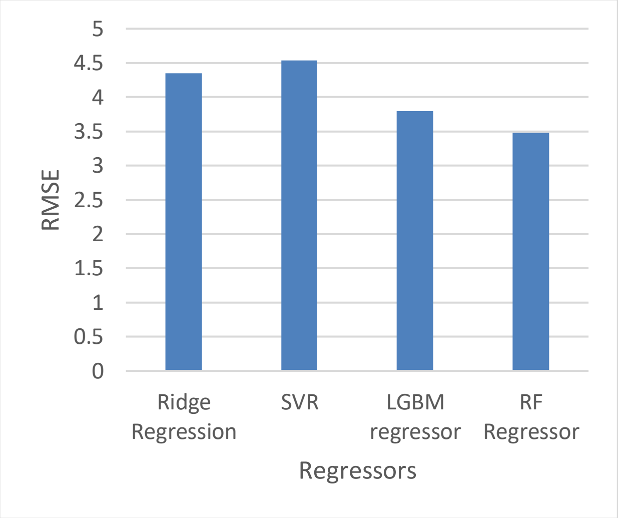

Figure 5. RMSE plot for the proposed Algorithms

The MAE values were as follows: Ridge Regression (1.24), SVR (1.19), LGBM Regressor (1.16), and RF Regressor (0.92) as shown in Figure.6. RF Regressor achieved the lowest MAE, indicating superior accuracy in predicting the outcome variable compared to other models, followed by LGBM Regressor.

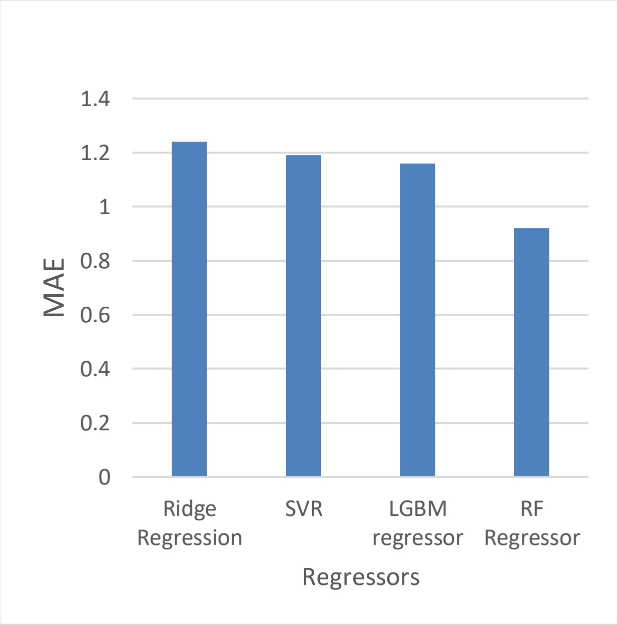

Figure 6. MAE plot for the proposed Algorithms

Many attributes affect the prediction process of influent TN concentration in WWTP. Figure.5 displays the significance level for each attribute in the prediction process. It can be observed that COD is by far the most significant followed by NH3 and EC and ending with AH and AV which have the least significance.

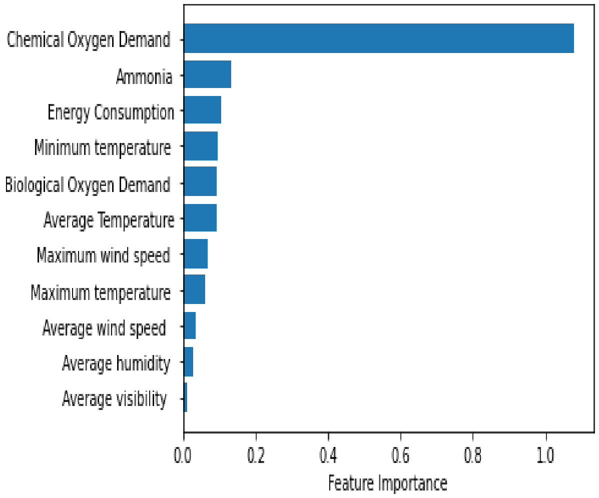

Figure 7. Performance plot of proposed Algorithms

The results obtained in this study are aligned with the results of the study conducted by Moghaddam and Mesghai [10]; which also predicted total nitogen, since in both studies the maximum obtained R^2 was 0.64; however, in this study it was attained through RF while in theirs it was achieved through KNN-BOP and SVR-BOP. Additionally, in a study carried out by Xu, et al [8] to forecast effluent phosphorous level, SVC attained an R^2 of 0.63. Results obtained in any conducted study depend on the chosen model, parameters in it, quality and size of dataset; in addition to, the preprocessing procedure conducted. In this study the performance of four algorithms was compared under the same conditions, making the comparison precise.

5. Conclusion and Future Direction

The principal ambition of the comparative research analysis was to perform an evaluation and performance comparison of the predictive power of the RF regressor, LGBM, SVR, and Ridge Regression in estimating the influent TN concentration in WWTPs. The models are evaluated based on metrics like RMSE, MAE, and R^2. The findings unequivocally demonstrate the RF regressor's superior performance compared to the other regressors.

The RF regressor excellence is due to the possession of a notable advantage, being its unique working approach, which involves introducing additional randomness during the tree growth process. Unlike traditional methods that prioritize the most significant feature for node splitting, the RF regressor selects the best feature from a randomly chosen subset of features. This strategy fosters a diverse range of trees, resulting in a more robust and accurate model overall.

Furthermore, the study identifies Chemical Oxygen Demand (COD) as the most significant attribute in predicting influent TN concentration. The effect of COD surpasses that of other attributes, as it serves as a reliable indicator of organic matter content in wastewater, exerting a direct influence on nitrogen levels. Ensuring accurate measurement and consideration of COD values is crucial for effective TN concentration prediction and management in WWTPs, thereby facilitating efficient wastewater treatment and environmental protection.

Addressing the challenges faced by ML in this field, this study contributes to improving the accuracy and reliability of TN concentration prediction in WWTPs. The findings have significant implications for environmental concerns, particularly in identifying and managing critical effluents. Enhanced prediction of TN concentration can optimize the recovery process of nitrogen during wastewater treatment and water purification, which plays a vital role in energy and economic development.

In conclusion, this study underscores the importance of COD in predicting TN concentration and highlights the efficacy of ML algorithms, specifically the RF regressor, in enhancing WWTP performance. The potential of these algorithms, including the RF regressor, lies in their ability to accurately predict TN concentrations, facilitating resource allocation, process optimization, and compliance with environmental regulations. These advancements contribute to sustainable water management, addressing the global challenge of water scarcity. The valuable insights gained from this research benefit researchers, engineers, and policymakers involved in wastewater treatment, fostering the development of efficient and sustainable solutions for water resource management.

This study makes a significant contribution to TN concentration prediction in WWTPs using four regression algorithms: Ridge, SVR, LGBM, and RF. The findings enhance predictive modeling in this field and provide novel methodologies for accurate TN estimation.

Several recommendations for future studies in this field can be proposed:

- Subsequent research should focus on developing the most accurate model to assist specialists in predicting TN concentrations in WWTPs. This can involve exploring novel algorithmic approaches and incorporating advanced techniques to further enhance the efficacy of the prediction process.
- Future works should consider expanding the range of ML algorithms employed in the comparative analysis. In addition to the existing models, the inclusion of deep learning approaches and hybrid models can

provide valuable insights and potentially improve the accuracy of TN concentration predictions.

- Attention should be given to the exploration of additional datasets in future studies. Incorporating diverse and comprehensive datasets can yield interesting outcomes, enhance the generalizability of the models, and provide an all-inclusive predictive relationship between attributes and TN concentrations.
- Future research endeavors can be directed towards predicting other crucial wastewater effluent parameters, such as BOD, NH3, and COD, within the context of WWTPs. Expanding the scope of prediction to encompass these parameters will provide a more holistic approach to wastewater management and facilitate comprehensive decision-making processes.

By addressing these recommendations, future studies can contribute to advancing the field of TN concentration prediction in WWTPs, enabling more accurate and effective wastewater treatment strategies.

6. References

[1] M. J. Song *et al.*, "Identification of primary effecters of N2O emissions from full-scale biological nitrogen removal systems using random forest approach," *Water Research*vol. 184, p. 116144, 2020.

[2] X. Chen, A. T. Mielczarek, K. Habicht, M. H. Andersen, D. Thornberg, and G. Sin, "Assessment of full-scale N2O emission characteristics and testing of control concepts in an activated sludge wastewater treatment plant with alternating aerobic and anoxic phases," *Environmental science technology*. vol. 53, no. 21, pp. 12485-12494, 2019.

[3] V. Vasilaki, E. Volcke, A. Nandi, M. Van Loosdrecht, and E. Katsou, "Relating N2O emissions during biological nitrogen removal with operating conditions using multivariate statistical techniques," *Water research*,vol. 140, pp. 387-402, 2018.

[4] P. M. Ching, R. H. So, and T. Morck, "Advances in soft sensors for wastewater treatment plants: A systematic review," *Journal of Water Process Engineering*,vol. 44, p. 102367, 2021.

[5] V. Vasilaki, S. Danishvar, A. Mousavi, and E. Katsou, "Data-driven versus conventional N2O EF quantification methods in wastewater; how can we quantify reliable annual EFs?," *Computers Chemical Engineering*,vol. 141, p. 106997, 2020.

[6] E. Aghdam, S. R. Mohandes, P. Manu, C. Cheung, A. Yunusa-Kaltungo, and T. Zayed, "Predicting quality parameters of wastewater treatment plants using artificial intelligence techniques," *Journal of Cleaner Production,* vol. 405, p. 137019, 2023.

[7] J. Tomperi, E. Koivuranta, and K. Leiviskä, "Predicting the effluent quality of an industrial wastewater treatment plant by way of optical monitoring," *Journal of water process engineering,* vol. 16, pp. 283-289, 2017.

[8] Y. Xu, Z. Wang, S. Nairat, J. Zhou, and Z. He, "Artificial Intelligence-Assisted Prediction of Effluent Phosphorus in a Full-Scale Wastewater Treatment Plant with Missing Phosphorus Input and Removal Data," *ACS ES T Water*,2023.

[9] K. Yan, C. Li, R. Zhao, Y. Zhang, H. Duan, and W. Wang, "Predicting the ammonia nitrogen of wastewater treatment plant influent via integrated model based on rolling decomposition method and deep learning algorithm," *Sustainable Cities Society*,vol. 94, p. 104541, 2023.

[10] S. Sadri Moghaddam and H. Mesghali, "A new hybrid ensemble approach for the prediction of effluent total nitrogen from a full-scale wastewater treatment plant using a combined trickling filter-activated sludge system," *Environmental Science Pollution Research*,vol. 30, no. 1, pp. 1622-1639, 2023.

[11] B. Adibimanesh, S. Polesek-Karczewska, F. Bagherzadeh, P. Szczuko, and T. Shafighfard, "Energy consumption optimization in wastewater treatment plants: machine learning for monitoring incineration of sewage sludge," *Sustainable Energy Technologies Assessments*,vol. 56, p. 103040, 2023.

[12] L. M. Manzo, L. B. Epele, C. N. Horak, Y. A. Assef, and M. L. Miserendino, "Variability in nutrient dissipation in a wastewater treatment plant in Patagonia: a two-year overview," *Environmental Management*,vol. 71, no. 4, pp. 773-784, 2023.

[13] Y. Yang *et al.*, "Prediction of effluent quality in a wastewater treatment plant by dynamic neural network modeling," *Process Safety Environmental Protection*,vol. 158, pp. 515-524, 2022.

[14] M. F. Abushammala, N. E. A. Basri, R. Elfithri, M. K. Younes, and D. Irwan, "Modeling of methane oxidation in landfill cover soil using an artificial neural network," *Journal of the Air Waste Management Association*,vol. 64, no. 2, pp. 150-159, 2014.

[15] F. Bagherzadeh, M.-J. Mehrani, M. Basirifard, and J. Roostaei, "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,* vol. 41, p. 102033, 2021.

[16] K. Lotfi *et al.*, "Predicting wastewater treatment plant quality parameters using a novel hybrid linear-nonlinear methodology," *Journal of environmental management,* vol. 240, pp. 463-474, 2019.

[17] M. Yaqub, H. Asif, S. Kim, and W. Lee, "Modeling of a full-scale sewage treatment plant to predict the nutrient removal efficiency using a long short-term memory (LSTM) neural network," *Journal of Water Process Engineering,* vol. 37, p. 101388, 2020.

[18] P. Zhou, Z. Li, S. Snowling, B. W. Baetz, D. Na, and G. Boyd, "A random forest model for inflow prediction at wastewater treatment plants," *Stochastic Environmental Research Risk Assessment*,vol. 33, no. 10, pp. 1781-1792, 2019.

[19] X. Liu, Q. Shi, Z. Liu, and J. Yuan, "Using LSTM Neural Network Based on Improved PSO and Attention Mechanism for Predicting the Effluent COD in a Wastewater Treatment Plant," *IEEE Access,* vol. 9, pp. 146082-146096, 2021.

[20] A. N. Matheri, F. Ntuli, J. C. Ngila, T. Seodigeng, and C. Zvinowanda, "Performance prediction of trace metals and cod in wastewater treatment using artificial neural network," *Computers Chemical Engineering*,vol. 149, p. 107308, 2021.

[21] M. El-Rawy, M. K. Abd-Ellah, H. Fathi, and A. K. A. Ahmed, "Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques," *Journal of Water Process Engineering,* vol. 44, p. 102380, 2021.

[22] J. Wijaya and S. Oh, "Machine learning reveals the complex ecological interplay of microbiome in a full-scale membrane bioreactor wastewater treatment plant," *Environmental Research*,vol. 222, p. 115366, 2023.

[23] C. Toffanin, F. Di Palma, F. Iacono, and L. Magni, "LSTM Network for the Oxygen Concentration Modeling of a Wastewater Treatment Plant," *Applied Sciences,* vol. 13, no. 13, p. 7461, 2023.

[24] V. Krishnaswamy, N. Singh, M. Sharma, N. Verma, A. J. J. o. E. P. Verma, and Management, "Application of CRISP-DM methodology for managing human-wildlife conflicts: an empirical case study in India," *Journal of Environmental Planning Management*,pp. 1-27, 2022.

[25] Bagehrzadeh and Faramarz, “Full Scale Wastewater Treatment Plant Data”, Mendeley Data, 2021

[26] A. Mavrommatis and G. Christodoulou, "Comparative Experimental Study of Flow through Various Types of Simulated Vegetation," *Environmental Processes,* vol. 9, no. 2, pp. 1-15, 2022.

[27] A. Amarilla, "Scalarization methods for many-objective virtual machine placement of elastic infrastructures in overbooked cloud computing data centers under uncertainty," *arXiv preprint arXiv:.04245*, 2018.

[28] M. S. I. Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," *Journal of King Saud University-Computer Information Sciences*.2021.

[29] H.-L. Chen, B. Yang, G. Wang, S.-J. Wang, J. Liu, and D.-Y. Liu, "Support vector machine based diagnostic system for breast cancer using swarm intelligence," *Journal of medical systems,* vol. 36, no. 4, pp. 2505-2519, 2021.

[30] S. Mangalathu, K. Karthikeyan, D.-C. Feng, and J.-S. Jeon, "Machine-learning interpretability techniques for seismic performance assessment of infrastructure systems," *Engineering Structures,* vol. 250, p. 112883, 2022.

[31] F. S. El Mustapha Azzirgue, T. A. T. El Khalil Cherif, and N. Mejjad, "Using Machine Learning Approaches to Predict Water Quality of Ibn Battuta Dam (Tangier, Morocco)," 2022.

[32] A. Elawwad, M. Matta, M. Abo-Zaid, and H. Abdel-Halim, "Plant-wide modeling and optimization of a large-scale WWTP using BioWin's ASDM model," *Journal of Water Process Engineering,* vol. 31, p. 100819, 2019.

[33] M. Salgot, M. Folch, and Health, "Wastewater treatment and water reuse," *Current Opinion in Environmental Science,* vol. 2, pp. 64-74, 2018.

[34] N. Khatri, K. K. Khatri, and A. Sharma, "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,* vol. 37, p. 101477, 2020.

Fuad Ahmad Musleh, Assistant Professor at the University of Bahrain. Ph.D. and master's degree from the University of Alabama in Huntsville, B.Sc. from Jordan University of Science and Technology in Jordan. Interested in research related to flow through vegetation, water and environmental conservation.