

1 **Comparative study on total nitrogen prediction in wastewater treatment** 2 **plant and effect of various feature selection methods on machine learning** 3 **algorithms performance**

4 Faramarz Bagherzadeh^{1*}, Mohamad-Javad Mehrani², Milad Basirifard³, Javad Roostaei⁴

5 1- Faculty of Mechanical Engineering, Gdansk University of Technology, Narutowicza Street 11/12, 80-233 Gdansk, Poland

6 2- Faculty of Civil and Environmental Engineering, Gdansk University of Technology, Narutowicza Street 11/12, 80-233 Gdansk, Poland

7 3- Department of Environmental Engineering, Faculty of Engineering, University of Tehran, Enghelab Sq., Tehran, Iran

8 4- Department of Environmental Science and Engineering, University of North Carolina, Chapel Hill, 27599, USA

9 *Corresponding Author: Faramarz Bagherzadeh (Email: s179532@student.pg.edu.pl)

10 **Link:** <https://doi.org/10.1016/j.jwpe.2021.102033>

11 **Abstract**

12 Wastewater characteristics prediction in wastewater treatment plants (WWTPs) is valuable and
13 can reduce the number of sampling, energy, and cost. Feature Selection (FS) methods are used
14 in the pre-processing section for enhancing the model performance. This study aims to evaluate
15 the effect of seven different FS methods (filter, wrapper, and embedded methods) on enhancing
16 the prediction accuracy for total nitrogen (TN) in the WWTP influent flow. Four scenarios
17 based on FS suggestions were defined and compared by three supervised Machine Learning
18 (ML) algorithms, i.e. Artificial Neural Network (ANN), Random Forest (RF), and especially
19 Gradient Boosting Machine (GBM). Input parameters, as daily time-series including pH, DO,
20 COD, BOD, MLSS, MLVSS, NH₄-N, and TN concentration, were used. Data set divided into
21 train and unseen test data-sets, and performance precision of all models was carried out based
22 on Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and correlation coefficient
23 (R²). Results reveal that scenario IV which was suggested by Mutual Information, including
24 NH₄-N, COD, BOD, and DO had the best result rather than other FS methods. Furthermore,
25 decision tree algorithms (RF and GBM) revealed better performance results in comparison to
26 neural network algorithm (ANN). GBM generalized the dataset patterns very well and

27 produced the best performance on unseen data-set, which shows the effectiveness of this state-
28 of-the-art ML algorithm for wastewater components prediction.

29 **Keywords:** *ANN; RF; GBM; Feature selection; total nitrogen*

30 **1 Introduction**

31 Nitrogen is one of the major wastewater pollutions, which should be reduced to the standard
32 level before discharging wastewater to the environment [1, 2]. Total nitrogen (TN) is primarily
33 presented in wastewater as ammonia, nitrite, nitrate, and organically bonded nitrogen [3].
34 Monitoring the TN from the influent of wastewater treatment plants (WWTPs) plays a
35 significant role in the performance of nutrient removal systems, controlling sludge production,
36 and operation of different parts of wastewater treatment processes [4].

37 Wastewater parameters especially nutrient compounds are really important for engineers to
38 understand and calculate at the beginning and end of treatment [5]. To obtain the necessary
39 information, the operator should determine the characteristics of the raw wastewater by
40 receiving data from sensors or collecting samples and analyzing the influent/effluent flow of
41 the plant. Insufficiently treated wastewater which is one of the nutrient sources can cause many
42 health problems by entering into the water bodies like groundwater systems [6]. However,
43 many facilities have been upgraded by WWTPs to progress in the removal of nutrient pollutants
44 which resulting in a drastic decrease in the discharged nutrients from WWTPs [7, 8].

45 Artificial Intelligent (AI) technics, mostly used to predict natural or artificial processes in
46 various disciplines. Machine learning (ML) as a subset of AI, is a process of recognizing a
47 special pattern based on the given data for prediction or classification purposes [9]. Recently,
48 modeling and prediction of environmental phenomena using AI technics are rapidly increased
49 due to their high accuracy rather than mechanistic models [10]. These algorithms can learn
50 sophisticated relations more efficiently than statistical methods [11-13].

51 A fully connected neural network known as ANN model acts as a universal function estimator,
52 and each neuron in the network contains learnable parameters (weight and bias). A feed-
53 forward ANN can be used for WWTPs influent and or effluent quality prediction [14, 15].

54 Many studies have been addressed to model the influent or effluent wastewater parameters. For
55 instance, ANN models are employed to estimate the methane production in a biogas

56 optimization scenario while having ($R^2=0.87$) [16]. Also, another similar modeling was
57 conducted to follow the correlation of supplements membrane bioreactor of WWTP [17].
58 Ansari et al. integrated a hybrid genetic algorithm with fuzzy logic (GA-FIS) model to increase
59 the prediction of missing value in the wastewater parameters record like COD, BOD, and NH_4 -
60 N and compared it with fuzzy logic (ANFIS) model. Results presented that integrated GA-FIS
61 had lower errors in contrast to ANFIS prediction [18]. In another study, Abba et al. studied an
62 extreme learning machine (ELM) model combined with kernel principal component analysis
63 (KPCA) for prediction of pH, turbidity, total dissolved solids, and hardness, which had the
64 highest accuracy for almost all predicted components ($R^2 > 0.95$) [19]. Random forest (RF) and
65 Gradient Boosting Machine (GBM) are other state-of-the-art and powerful ML methods [20-
66 22]. An RF prediction model was found a useful and powerful method for the evaluation of
67 reliability prediction of small WWTPs in the UK [23].

68 On the other hand, the feature selection (FS) process is utilized in the pre-processing section
69 for increasing the speed of training and enhancing the prediction precision as well as
70 simplifying the models [24]. Although there have been many different FS methods, most
71 forecasting studies just use correlation models, like the Pearson correlation method. Hence, a
72 comparative evaluation of the FS effect on enhancing the accuracy of simulation for WWTPs
73 components is still required. Also, prediction of WWTP components with RF and GBM is less
74 used in comparison to other ML techniques i.e. ANN, SVM, etc. [23, 25, 26].

75 This study aims to investigate the effect of various feature selection methods for enhancing the
76 prediction performance of TN in the WWTPs. The specific objectives of this paper are: i)
77 Defining scenarios according to the different FS suggestions and compare together, ii) Create
78 ML models by using algorithms such as ANN, RF, and GBM for comparing scenarios and find
79 the best TN forecasting model, and iii) Evaluate the potential of using a state-of-the-art GBM
80 algorithm as a new ML model for TN prediction and compared with the conventional methods
81 (RF and ANN).

82 **2 Methodology**

83 **2.1 Case study of WWTP and data description**

84 In this research, a data set from North Torbat WWTP for nutrients removal which is located in
85 the north of Iran was investigated. This WWTP is designed for a population of 350,000 PE

86 with a mean-daily influent flow of 71,500 m³/d. The WWTP consists of a primary
87 sedimentation tank, anaerobic/aerobic reactors, and a clarifier. The pH and DO are monitored
88 using online sensors, and the rest of the influent characteristics are recorded using sampling
89 and analysis based on the standard for wastewater analysis method [27].

90 A data set consisting of 800 records (almost 2.5 years between 2015-2017) daily recording of
91 total nitrogen (TN), Ammonia nitrogen (NH₄-N), biological oxygen demand (BOD), chemical
92 oxygen demand (COD), mixed liquor suspended solids (MLSS), Mixed liquor volatile
93 suspended solid (MLVSS), pH, dissolved oxygen (DO) for the training of models. Also, 30
94 days from the last data set was selected for the test of models (unseen data).

95 Furthermore, for obtaining an accurate model, the data set should be normalized, and
96 unnecessary (redundant) features should be eliminated (feature selection) to avoid overfitting
97 issues [28, 29]. One of the main points of this study is to compare different applicable feature
98 selection methods and their effects on model precision. TN was selected as a target of
99 prediction in this study due to the level of importance in the WWTPs as a critical influent
100 quality index and the rest of parameters were selected as input data based on feature selection
101 ranking.

102

103 **2.2 Feature selection (FS)**

104 The main goal of feature selection (FS) is to obtain the most relevant input data from a dataset.
105 Considering a dataset with M features, then 2^M subsets of features are available, and the FS
106 methods are responsible for introducing the best subset. In each method, related to the criterion
107 and application of the model several functions are responsible to optimize and evaluate the
108 subset. The FS methods are divided into three major categories: filters, wrappers, and
109 embedded methods [30- 32].

110 Filter methods emphasize on characteristics of each feature and they evaluate the features based
111 on the properties without employing any clustering algorithm to guide the search [30]. Wrapper
112 methods use clustering algorithms. If the introduced subset increased the accuracy of the
113 clustering algorithms, then the subset earns a higher score [30]. Embedded techniques combine
114 all the advantages of wrappers and filters. They construct an ML algorithm, and it performs
115 feature selection while training the model [33].

116 In this study, variance threshold [34], analysis of variance (ANOVA) [35], mutual information
117 (MI) [36,37], Pearson correlation (PC) [38], backward elimination (BE) [39], random forest
118 (RF) [40], and Least Absolute Shrinkage and Selection Operator (LASSO) were used [41,42].
119 The details of each method can be found in supplementary information.

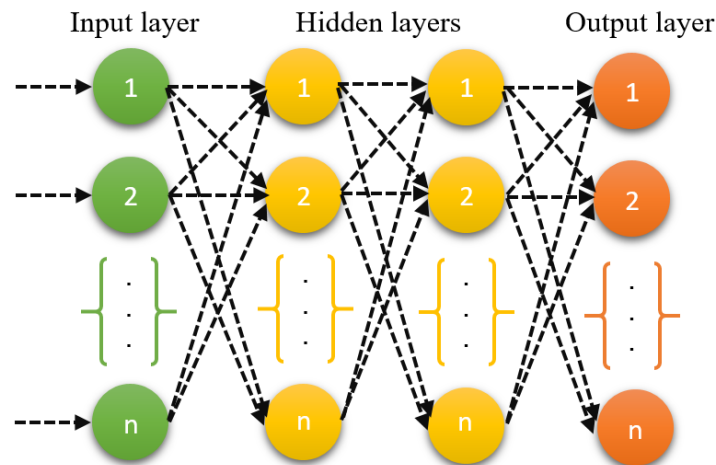
120

121 2.3 Modeling approaches

122 2.3.1 Artificial neural networks (ANNs)

123 An artificial neural network (ANN) is a fully connected multilayer perceptron (MLP) with
124 three layers: input, hidden, and output (Fig. 1). The network may have several hidden layers
125 concerning the level of complexity of the data set [43, 44].

126 In this study, the number of input neurons of the model is equal to the number of input features
127 which depends on the scenario (considered subset). Also, two hidden layers with 15 and 10
128 neurons are designed to capture the complexity of the model. For having a smooth and accurate
129 connection between layers, we used the ReLU activation function for the hidden layers. Finally,
130 there is a single neuron in the output layer to predict the target variable (TN). The optimization
131 process was performed by Adam's algorithm concerning the mean squared error (MSE) as a
132 loss function with 100 epochs.



133

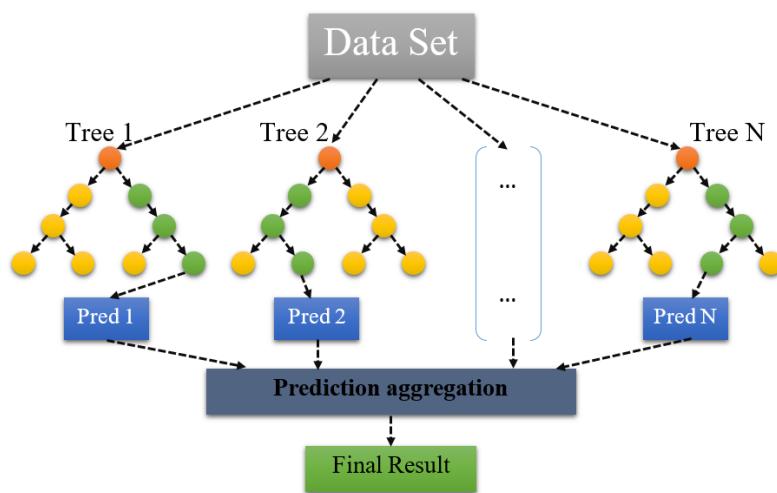
134

135

Fig. 1. Fully connected artificial neural network (ANN)

136 **2.3.2 Random forest (RF)**

137 Ensemble learning is a technique that combines the prediction results of multiple algorithms to
138 obtain a better final result. Random forest (RF) is an ensemble method that uses bootstrap
139 aggregation to generate decision trees. The final output of the model is an aggregation of the
140 prediction based on the decision trees (Fig. 2). This method helps to consider all potential
141 features fairly and prevents trees to become highly correlated [45]. In this study, after many
142 trials and errors, a random forest tree was developed considering, 400 trees in the forest, a
143 maximum depth of 70 for each tree, minimum of 4 samples at a leaf, and a minimum number
144 of 10 samples required to split.



145

146 Fig. 2. Random forest architecture

147

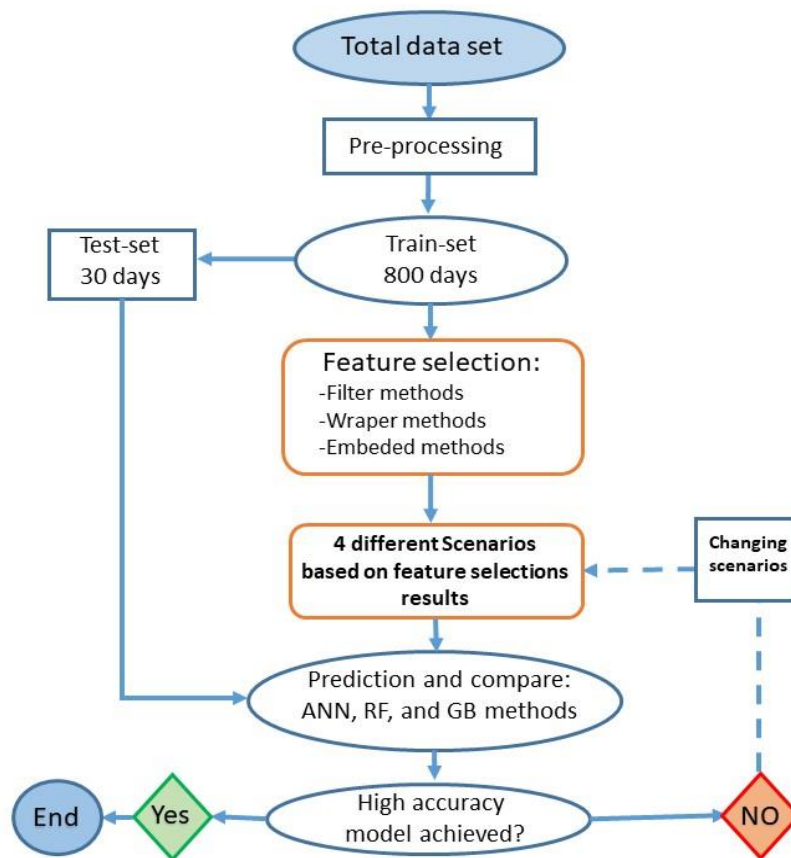
148 **2.3.3 Gradient boosting machine (GBM)**

149 Gradient boosting machine is a decision tree based ML algorithm similar to RF, but it has a
150 different constructive strategy of the ensemble formation. In a boosting approach, we add new
151 trees to the ensemble sequentially according to the error of the whole ensemble prediction. As
152 we add new trees with a constant learning rate, the estimation error regarding the dependent
153 variable shrinks continuously until reaching the maximum possible precision. Due to the nature
154 of GBM, hyper-parameters justification is extremely important [46]. In the current research,
155 after many trials, we used a gradient boosting machine with considering the learning rate of
156 0.05 for training, 2000 trees in the forest, subsampling of a total of 0.8, a min sample leaf of
157 50, a tree depth of 6, and 600 as minimum split samples.

158

159 2.4 Model construction

160 The total data set was divided into two groups: train data (almost 90% of total data-set) and
161 test data (10% of the total data set) as unseen data, followed by applying to preprocess, and
162 feature selection methods. Furthermore, four scenarios were defined to compare the feature
163 selection methods. Also, for the prediction of the target (TN), three prediction models
164 containing fully connected artificial neural network (ANN), random forest (RF), and state-of-
165 the-art gradient boosting machine (GBM) were selected and applied for all sub-data-sets. The
166 normalized data were used as input data for training and testing all models. After defining
167 different scenarios and model structures, the TN concentration was forecasted by noted models,
168 then the predicted values were compared to the real data to evaluate the model accuracy (Fig.
169 3).



170
171 Fig. 3. Modeling and prediction structure
172

173 **2.5 Model evaluation**

174 To measure the quality and performance of a model, several model metrics can be employed
175 depending on the model task, data types, and scenarios. In this article, models are scored based
176 on the coefficient of determination (R^2) (Eq.1), root mean square error (RMSE) (Eq.2), and
177 mean absolute error (MAE) (Eq.3) [47].

178
$$R^2 = 1 - \frac{\sum(a_i - p_i)^2}{(a_i - \mu_a)^2} \quad (1)$$

179
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (a_i - p_i)^2} \quad (2)$$

180
$$MAE = \frac{1}{n} \sum_{i=1}^n |a_i - p_i| \quad (3)$$

181

182 Where $i = 1, 2, \dots, n$ is the number of observations, and n is the total number of records.
183 Considering a_i for output, p_i as real values, and μ_a as mean value.

184 **3 Results and discussion**

185 **3.1 Data statistical information**

186 A brief demonstration of primary statistical properties (Min, Mean, Max, and standard
187 deviation) is represented in Table 1.

188 Table 1

189 Data set statistical properties

Parameters	Units	Min	Mean	Max	SD
pH	-	6.94	7.26	7.90	2.14
DO	mg/L	1.22	1.34	1.80	2.18
NH ₄ -N	mg/L	26.11	68.78	93.94	3.14
BOD	mg/L	205.40	505.83	858	4.21
COD	mg/L	367.36	1063.85	1881.6	4.87
MLSS	mg/L	148.80	554.17	1041.5	3.64
MLVSS	mg/L	99.7	366.39	697.87	3.51
TN	mg/L	35.81	91.60	125.73	3.16

190

191 3.2 Summary of selected feature procedure

192 Each FS method has a particular subset suggestion as described in Table 2. The variance
193 threshold revealed the redundant features. Filter methods (ANOVA and MI) suggesting very
194 similar subsets, and PC is indicating NH₄-N with the highest correlation to the target variable.

195 TN as the target of the prediction displayed a strong correlation with NH₄-N, COD, and BOD
196 respectively, and a weak correlation with pH and DO as can be seen in the Pearson correlation
197 result (supplementary file, Fig.S6). The highest correlation (~1.0) among input parameters
198 belonged to MLSS and MLVSS, and the lowest value was related to BOD and pH.

199 Generally, FS algorithms are pointing at (NH₄-N, COD, and BOD) as the best possible subset,
200 while LASSO is showing a different result. Besides, four scenarios as shown in Table 3 were
201 grouped based on FS suggestions and they were compared with ANN, RF, and GBM
202 techniques to introduce the best scenario. Details for the result of each FS process can be found
203 in the supplementary information file.

204 Table 2

205 Summary of feature selection application on the data set

Method	Subset	Description
Variance	COD, MLSS, MLVSS, BOD, NH ₄ -N	Dropping redundant features
ANOVA	NH ₄ -N, COD, BOD, MLSS	Ranking based on the importance level
Mutual Information	NH ₄ -N, COD, BOD, DO	Ranking based on the importance level
Pairwise correlation	NH ₄ -N	Choosing highly correlated features (> 0.8) with target
Backward	NH ₄ -N, COD, BOD	Choosing features that increase the regression model performance reasonably
Random Forest	NH ₄ -N, COD, BOD, MLSS	Ranking based on Gini importance level
LASSO	NH ₄ -N, MLSS, MLVSS	Ranking based on the importance level

206

207 Table 3
 208 Different scenarios defined in this study based on different feature selection methods

Scenario	Number of features	Suggested by	Name of features
I	1	Pairwise correlation	NH ₄ -N
II	3	LASSO	NH ₄ -N, MLSS, MLVSS
III	4	ANOVA, Random Forest, Backward Elimination	NH ₄ -N, COD, BOD, MLSS
IV	4	Mutual Information	NH ₄ -N, COD, BOD, DO

209

210 3.3 Prediction results

211 After training the models in various scenarios, the whole dataset was predicted by models (Fig.
 212 4), then model metrics were calculated (R^2 , RMSE, and MAE) for both training and test dataset.
 213 The model metrics of each scenario are described in both data sets (Table 4).

214 According to scenario I, with only one feature (NH₄-N), RF has the best performance on the
 215 training set, but its performance dropped significantly on the test data set which shows serious
 216 overfitting issues. Similarly, ANN and GBM lost their precision, but with a lower difference.
 217 These two models with $R^2=0.52$ had a better outcome for this scenario on the test data-set.

218 In scenario-II, although more features were introduced to the models and accuracy on the
 219 training set was increased, the performance on the test data-set was decreased. This result is
 220 indicating the introduced subset is not adding precision to the models, but it is causing
 221 overfitting issues. For these features, GBM has the best result on the test dataset with $R^2=0.52$.

222 In scenario-III, the RF has the highest accuracy on the training dataset and the lowest
 223 performance on the test dataset. In this scenario, ANN performance was increased slightly, so
 224 it is showing that this subset has a better outcome than the subset in scenario-II for neural
 225 network algorithms, while it is causing overfitting issues for decision tree algorithms (GBM
 226 and RF).

227 The last scenario (IV) showed better results compared to previous scenarios. As indicated in
 228 Table 4, the RMSE of test data of this scenario is 0.092, 0.095, and 0.095 for GBM, RF, and
 229 ANN respectively. In this scenario, RF had less overfitting, ANN performance was increased,
 230 and GBM had the best performance both on training set $R^2=0.88$ and test data-set $R^2=0.58$.

231 GBM has the highest precision followed by RF and ANN in scenario-IV. Also, among FS
 232 methods, Mutual Information has better performance, because it was the only technique that
 233 considered DO as effective variable.

234 The results revealed that how sensitive is ANN to selecting the wrong features. In scenario-II
 235 and scenario-III, with adding more features to the subset, the ANN metrics decreased on the
 236 test dataset. Similarly, RF with very high performance on the training dataset suffered from
 237 more overfitting issues due to introducing inefficient subsets in scenario-II and scenario-III. In
 238 contrast, GBM showed a more robust model. Introducing wrong features to GBM didn't change
 239 the model performance considerably, and more or less it kept the performance level similar to
 240 previous subsets, but with introducing the best subset (scenario-IV), GBM showed an
 241 exceptional improvement in model evaluation. So, generally, it can be said that decision tree
 242 algorithms (RF and GBM) showed better performance than the neural network model (ANN).

243

244 Table 4

245 Model Metrics (accuracy and errors) of each prediction models

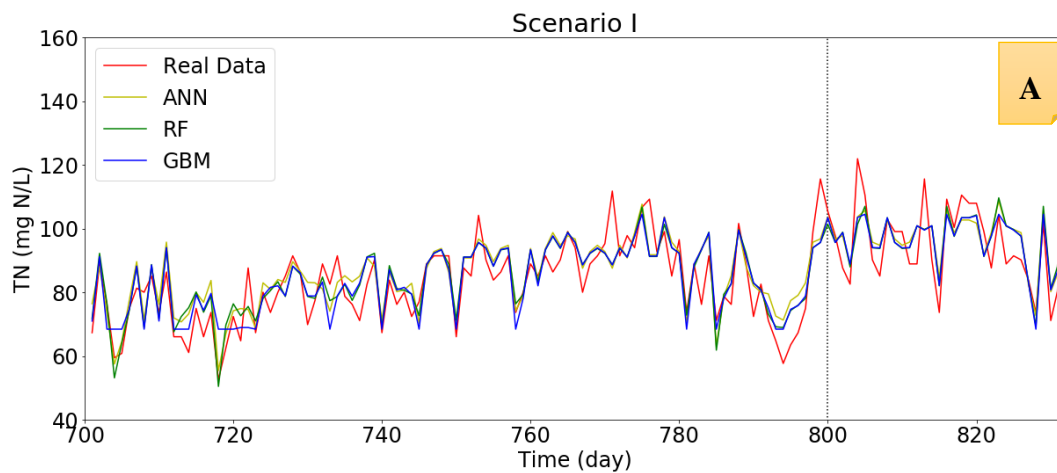
		Training Data			Test Data (unseen)			
		Model	R ²	RMSE	MAE	R ²	RMSE	MAE
Scenario I	ANN	0.77	77E-3	8E-5	0.52	94E-3	83E-4	
	GBM	0.76	78E-3	3.5E-5	0.52	95E-3	57E-4	
	RF	0.80	96E-3	5E-5	0.50	96E-3	104E-4	
Scenario II	ANN	0.75	79E-3	18E-3	0.42	104E-3	51E-3	
	GBM	0.81	72E-3	1E-3	0.52	95E-3	34E-3	
	RF	0.88	60E-3	1.4E-3	0.46	100E-3	34E-3	
Scenario III	ANN	0.79	74E-3	-5E-5	0.51	96E-3	28E-3	
	GBM	0.84	68E-3	11E-4	0.51	96E-3	31E-3	
	RF	0.89	55E-3	12E-4	0.48	98E-3	28E-3	

Scenario IV	ANN	0.81	73E-3	56E-4	0.55	95E-3	22E-3
	GBM	0.88	68E-3	4E-4	0.58	92E-3	17E-3
	RF	0.83	55E-3	9E-4	0.55	95e-3	19E-3

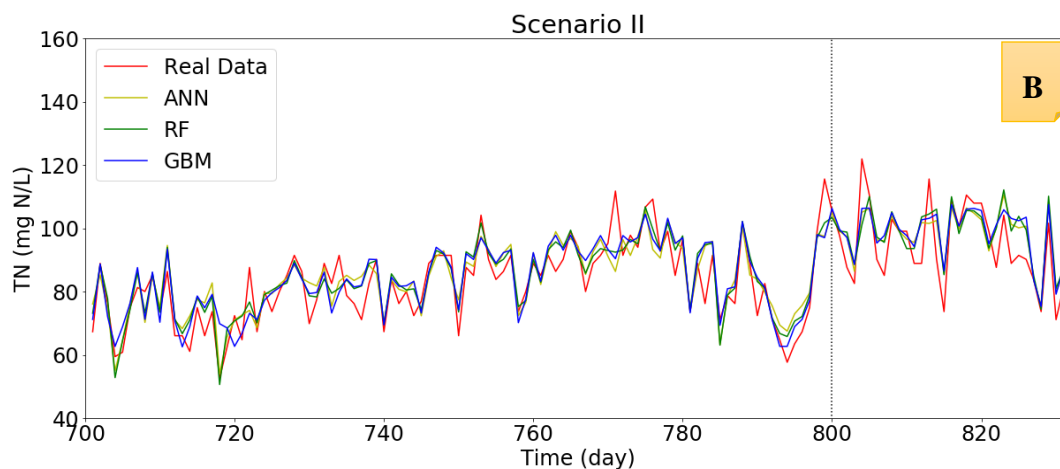
246

247 According to Table 4 and Fig.4, comparing Scenario-I with scenario II and III, the accuracy of
 248 models on capturing the complexity of the training dataset was increased, while the model
 249 performance on the test dataset was not improved. This means that the model learned the
 250 training dataset very well, but not able to generalize the patterns perfectly. Among all models,
 251 RF has had the best match with real data on the training data set and faced more with overfitting
 252 issues. In Scenario-IV, GBM showed the best matching with real data, as well as, a great
 253 improvement in generalizing the patterns for the unseen dataset.

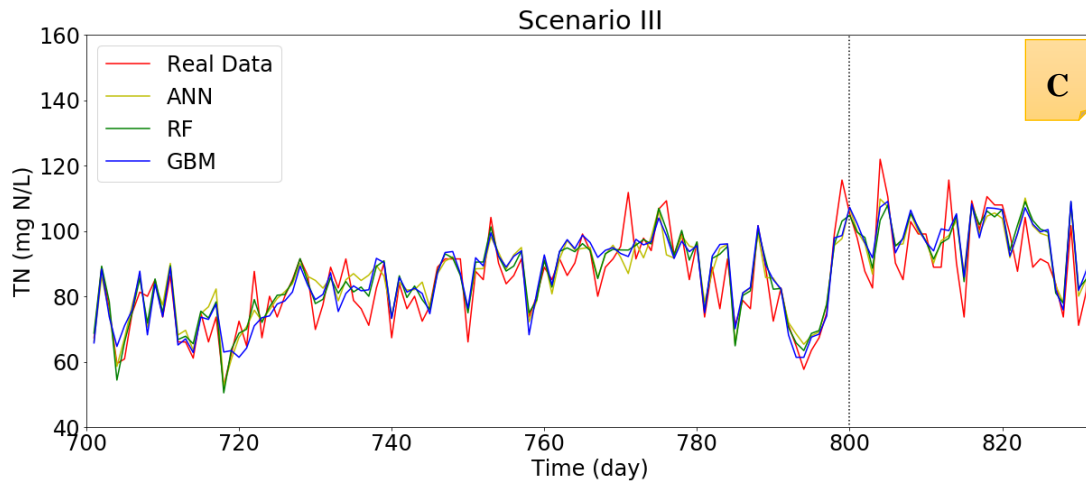
254



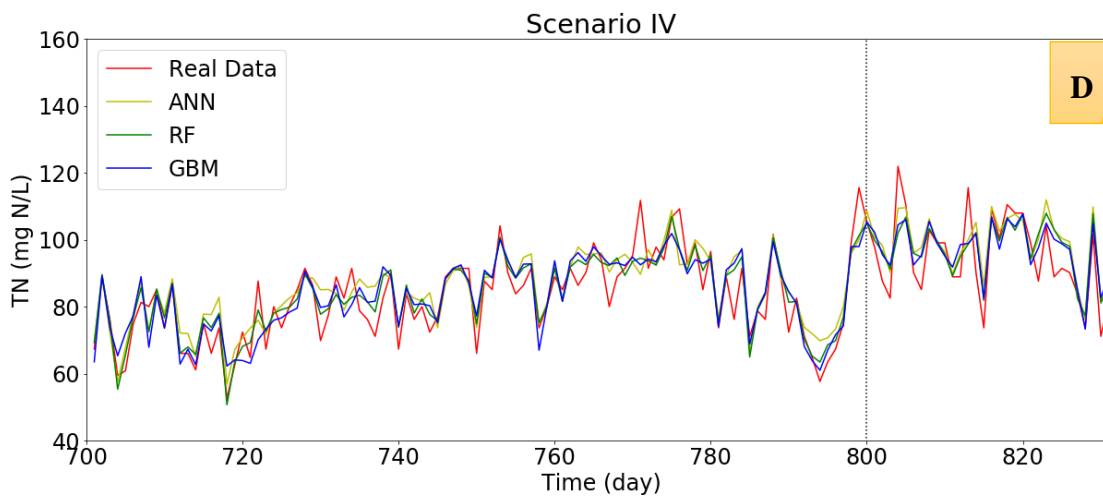
255



256



257



258

259 Fig. 4.- prediction of TN concentration for two sub-set of training (0-800 days) and unseen test data
 260 (800-830 days) for different scenarios, A) Scenario I, B) Scenario II, C) Scenario III, and D)
 261 Scenario IV

262 Prediction of critical characteristics like TN in the WWTP influent is a topic in which many
 263 researchers attempt to propose various methods to enhance precision. In recent approaches with
 264 ML methods and data-driven decision techniques, better results are demonstrated (Table 4).

265 It is noticeable that due to the sophisticated nature of different processes in WWTPs, there is
 266 no single adequate model for all types of similar issues. Consequently, this matter has required
 267 the improvement of more solid and effective models utilizing accessible information [48-52].
 268 Table 5 shows the summarized information of recently TN prediction studies in various
 269 WWTPs

270

271 Table 5

Feature selection methods	Prediction Algorithm	Model Accuracy (unseen data)	Remarks	References
Forward selection	parallel-serial hybrid	$R^2=0.81$, MSE=N/A	Combine ML models with mechanistic models (biological simulation) for increasing the model performance	Hvala et al. 2020 [50]
Latin Hypercube One factor At a Time (LH-OAT)	SVM, ANN	$R^2=0.47$, MSE=N/A	ANN showed better result rather than the SVM algorithm	Guo et al. 2015 [48]
Pearson Correlation	LSTM	$R^2 =N/A$ MSE=0.015	LSTM needs lower training time and has high performance for predicting unseen dataset.	Yaqub et al. 2020 [49]
Forward Selection, Genetic algorithms, Pearson Correlation	k-fold model	N/A	five-fold cross-validation caused an increase in the accuracy of the prediction	Tomperi et al. 2017 [5]
N/A	SDAE, SVR, BNN, GBM, SAE	$R^2=0.05$ MSE= 1.58	Stacked denoising auto-encoders (SDAE) showed the best performance for predicting TN	Shi and Xu 2018 [53]
Analysis of variance, Mutual Information, Backward Elimination, Pearson Correlation, LASSO	ANN, RF, and GBM	$R^2=0.58$, MSE=0.0084	GBM model showed the best performance on training and test data-set and less vulnerable to add or remove extra features. Also, the Mutual Information feature selection method suggested the best features.	This study

273

274 Based on table 5, Guo et al. [48] utilized SVM (support vector machine) and ANN to predict
 275 TN concentration in a WWTP. The models have trained 200 records and tested during the 90
 276 days. The model performance indicated the coefficient of determination 0.46 and 0.47 for
 277 SVM and ANN respectively. In a different approach, Tomperi [5] firstly, performed extensive
 278 feature selection methods such as stepwise selection, forward selection, and genetic algorithms,
 279 then they developed a k-fold model to predict TN with $R^2=0.69$ without testing on unseen data.
 280 Also, Yaqub et al. [49] proposed a prediction method for TN by developing a two-layered
 281 stacked long short-term memory (LSTM) network on a large data set (6000 training and 1876

282 testing) with a low average model error (MSE=0.015). In addition to that, hybrid simulation
283 can be helpful for accurate prediction of TN. For example, [50] designed a parallel-serial hybrid
284 model (machine learning and mechanistic models) on a data set (400 train and 250 test data-
285 set) with high accuracy ($R^2=0.81$). Combining external biological simulation (mechanistic
286 modeling) to ML algorithm caused high model precision. Hence, considering based on
287 standalone ML prediction, the proposed model by this study is a high accuracy model for TN
288 prediction among recent similar studies.

289 4 Conclusions

290 In the present study, the importance of using suitable FS as a booster of prediction was
291 evaluated. Also, the following conclusions were derived from this study as follows:

- 292 • Selecting a suitable feature selection for obtaining the best possible input-data increases
293 the prediction precision (up to 20%).
- 294 • Considering the outcome of recent literature for TN prediction in the influent/effluent
295 flow of WWTP, this study demonstrated high precision prediction by Mutual
296 Information FS model and GBM prediction algorithm.
- 297 • Scenarios III and IV declared a more reliable performance of the model predictions
298 which means that the wrapper feature selections (ANOVA, Random Forest, Backward
299 Elimination, and MI) can select the level of features importance better than commonly
300 used filter methods.
- 301 • Decision tree algorithms (RF and GBM) revealed better performance results in
302 comparison to neural network algorithm (ANN), and GBM has the highest accuracy
303 ($R^2=0.58$, RMSE=0.092, and MAE=0.017 for the test dataset respectively) followed by
304 RF and ANN in the best scenario (IV).
- 305 • GBM is less sensitive to add or remove features to the subset. In contrast, ANN
306 accuracy drops significantly, if redundant features are added.

307 5 References

- 308 1. Elawwad, A., et al., *Plant-wide modeling and optimization of a large-scale WWTP*
309 *using BioWin's ASDM model*. Journal of Water Process Engineering. 31 (2019)
310 100819.
- 311 2. WHO, *Guidelines for drinking-water quality*, World Health Organization. (2011).

- 312 3. Metcalf and Eddy, *Wastewater engineering: treatment and reuses recovery*. 5th ed.
313 ed. 2013, United States: McGraw-Hill Education - Europe.
- 314 4. Salgot, M. and M. Folch, *Wastewater treatment and water reuse*. *Current Opinion in*
315 *Environmental Science & Health*. 2 (2018) 64-74.
- 316 5. Tomperi, J., E. Koivuranta, and K. Leiviskä, *Predicting the effluent quality of an*
317 *industrial wastewater treatment plant by way of optical monitoring*. *Journal of Water*
318 *Process Engineering*. 16 (2017) 283-289.
- 319 6. Alighardashi, A. and M. Mehrani, *Survey and zoning of nitrate-contaminated*
320 *groundwater in Iran*. *Journal of Materials and Environmental Sciences*. 8 (2017) 10
321 2785-2794.
- 322 7. Jaramillo, F., et al., *Advanced strategies to improve nitrification process in*
323 *sequencing batch reactors - A review*. *Journal of Environmental Management*. 218
324 (2018) 154-164.
- 325 8. Liu, J., et al., *Advanced nutrient removal from surface water by a consortium of*
326 *attached microalgae and bacteria: A review*. *Bioresource Technology*. 241 (2017)
327 1127-1137.
- 328 9. Géron, A., *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*.
329 second ed. 2019: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol,
330 CA 95472.
- 331 10. Ye, Z., et al., *Tackling environmental challenges in pollution controls using artificial*
332 *intelligence: A review*. *Science of The Total Environment*. 699 (2020) 134279.
- 333 11. Mohammad, A.T., et al., *Modelling the chlorophenol removal from wastewater via*
334 *reverse osmosis process using a multilayer artificial neural network with genetic*
335 *algorithm*. *Journal of Water Process Engineering*. 33 (2020) 100993.
- 336 12. Poznyak, A., I. Chairez, and T. Poznyak, *A survey on artificial neural networks*
337 *application for identification and control in environmental engineering: Biological*
338 *and chemical systems with uncertain models*. *Annual Reviews in Control*. 48 (2019)
339 250-272.
- 340 13. Khatri, N., K.K. Khatri, and A. Sharma, *Artificial neural network modelling of faecal*
341 *coliform removal in an intermittent cycle extended aeration system-sequential batch*
342 *reactor based wastewater treatment plant*. *Journal of Water Process Engineering*. 37
343 (2020) 101477.
- 344 14. Khatri, N., K.K. Khatri, and A. Sharma, *Prediction of effluent quality in ICEAS-*
345 *sequential batch reactor using feedforward artificial neural network*. *Water Science*
346 *and Technology*. 80 (2019) 2 213-222.
- 347 15. Pisa, I., et al., *ANN-Based Soft Sensor to Predict Effluent Violations in Wastewater*
348 *Treatment Plants*. *Sensors*. 19 (2019) 6 1280.
- 349 16. Abu Qdais, H., K. Bani Hani, and N. Shatnawi, *Modeling and optimization of biogas*
350 *production from a waste digester using artificial neural network and genetic*
351 *algorithm*. *Resources, Conservation and Recycling*. 54 (2010) 6 359-363.

- 352 17. Giwa, A., et al., *Experimental investigation and artificial neural networks ANNs*
353 *modeling of electrically-enhanced membrane bioreactor for wastewater treatment.*
354 *Journal of Water Process Engineering.* 11 (2016) 88-97.
- 355 18. Ansari, M., F. Othman, and A. El-Shafie, *Optimized fuzzy inference system to*
356 *enhance prediction accuracy for influent characteristics of a sewage treatment plant.*
357 *Science of The Total Environment.* 722 (2020) 137878.
- 358 19. Abba, S.I., et al., *Emerging evolutionary algorithm integrated with kernel principal*
359 *component analysis for modeling the performance of a water treatment plant.* *Journal*
360 *of Water Process Engineering.* 33 (2020) 101081.
- 361 20. Jayaweera, C.D., M.R. Othman, and N. Aziz, *Improved predictive capability of*
362 *coagulation process by extreme learning machine with radial basis function.* *Journal*
363 *of Water Process Engineering.* 32 (2019) 100977.
- 364 21. Su, Y. and Y. Zhao. *Prediction of Downstream BOD based on Light Gradient*
365 *Boosting Machine Method.* in *2020 International Conference on Communications,*
366 *Information System and Computer Engineering (CISCE).* 2020.
- 367 22. Zhou, P., et al., *A random forest model for inflow prediction at wastewater treatment*
368 *plants.* *Stochastic Environmental Research and Risk Assessment.* 33 (2019) 10 1781-
369 1792.
- 370 23. Bunce, J.T. and D.W. Graham, *A Simple Approach to Predicting the Reliability of*
371 *Small Wastewater Treatment Plants.* *Water* 11 (2019) 2397.
- 372 24. Motoda, H.L., *Feature Selection for Knowledge Discovery and Data Mining.* 2012,
373 Springer Science & Business Media.
- 374 25. De Clercq, D., Z. Wen, and F. Fei, *Determinants of efficiency in anaerobic bio-waste*
375 *co-digestion facilities: A data envelopment analysis and gradient boosting approach.*
376 *Applied Energy.* 253 (2019) 113570.
- 377 26. Yu, Z., et al., *Efficient pyrolysis of ginkgo biloba leaf residue and pharmaceutical*
378 *sludge (mixture) with high production of clean energy: Process optimization by*
379 *particle swarm optimization and gradient boosting decision tree algorithm.*
380 *Bioresource Technology.* 304 (2020) 123020.
- 381 27. E.W. Rice, R.B. Baird, A.D. Eaton, *Standard Methods for the Examination of Water*
382 *and Wastewater,* Twenty, third ed., WEF, New York, 2017.
- 383 28. Ranjan, K.G., B.R. Prusty, and D. Jena, *Review of preprocessing methods for*
384 *univariate volatile time-series in power system applications.* *Electric Power Systems*
385 *Research.* 191 (2021) 106885.
- 386 29. Julián Luengo, et al., *Big Data Preprocessing.* 2020: Springer.
- 387 30. Solorio-Fernández, S., J.A. Carrasco-Ochoa, and J.F. Martínez-Trinidad, *A review of*
388 *unsupervised feature selection methods.* *Artificial Intelligence Review.* 53 (2020) 2
389 907-948.
- 390 31. Luíza da Costa, N., M. Dias de Lima, and R. Barbosa, *Evaluation of feature selection*
391 *methods based on artificial neural network weights.* *Expert Systems with*
392 *Applications.* 168 (2021) 114312.

- 393 32. Guozhu Dong and H. Liu, *Feature Engineering for Machine Learning and Data*
394 *Analytics*. 2018: Taylor and Francis group LLC.
- 395 33. Liu, H., M. Zhou, and Q. Liu, *An embedded feature selection method for imbalanced*
396 *data classification*. IEEE/CAA Journal of Automatica Sinica. 6 (2019) 3 703-715.
- 397 34. Scheffé, H., *The Analysis of Variance*. 1999: Wiley publisher.
- 398 35. Wilcox, R., *Modern Statistics for the Social and Behavioral Sciences*, ed. 2nd. 2017:
399 Chapman and Hall/CRC.
- 400 36. Gao, L. and W. Wu, *Relevance assignation feature selection method based on mutual*
401 *information for machine learning*. Knowledge-Based Systems. 209 (2020) 106439.
- 402 37. Gonzalez-Lopez, J., S. Ventura, and A. Cano, *Distributed multi-label feature*
403 *selection using individual mutual information measures*. Knowledge-Based Systems.
404 188 (2020) 105052.
- 405 38. Michalak, K. and H. Kwasnicka. *Correlation-based Feature Selection Strategy in*
406 *Neural Classification*. in *Sixth International Conference on Intelligent Systems Design*
407 *and Applications*. 2006.
- 408 39. Fernando Jimenez, et al., *Extreme Learning Machine Based Prediction of Soil Shear*
409 *Strength: A Sensitivity Analysis Using Monte Carlo Simulations and Feature*
410 *Backward Elimination*. Sustainability 12 (2020) 2339.
- 411 40. Masmoudi, S., et al., *A machine-learning framework for predicting multiple air*
412 *pollutants' concentrations via multi-target regression and feature selection*. Science
413 of The Total Environment. 715 (2020) 136991.
- 414 41. Sufi Karimi, H., et al., *Comparison of learning-based wastewater flow prediction*
415 *methodologies for smart sewer management*. Journal of Hydrology. 577 (2019)
416 123977.
- 417 42. Fonti, V. and E. Belitser. *Paper in Business Analytics Feature Selection using LASSO*.
418 2017.
- 419 43. Tosun, E., K. Aydin, and M. Bilgili, *Comparison of linear regression and artificial*
420 *neural network model of a diesel engine fueled with biodiesel-alcohol mixtures*.
421 Alexandria Engineering Journal. 55 (2016) 4 3081-3089.
- 422 44. P. Raut and A. Dani, *Correlation Between Number of Hidden Layers and Accuracy of*
423 *Artificial Neural Network*. 2020: Springer, Singapore.
- 424 45. Lakshmanaprabu, S.K., et al., *Random forest for big data classification in the internet*
425 *of things using optimal features*. International Journal of Machine Learning and
426 Cybernetics. 10 (2019) 10 2609-2618.
- 427 46. Natekin, A. and A. Knoll, *Gradient boosting machines, a tutorial*. Frontiers in
428 Neuroinformatics. 7 (2013) 21.
- 429 47. Bhagat, S.K., et al., *Prediction of sediment heavy metal at the Australian Bays using*
430 *newly developed hybrid artificial intelligence models*. Environmental Pollution. 268
431 (2021) 115663.

- 432 48. Guo, H., et al., *Prediction of effluent concentration in a wastewater treatment plant*
433 *using machine learning models*. Journal of Environmental Sciences. 32 (2015) 90-
434 101.
- 435 49. Yaqub, M., et al., *Modeling of a full-scale sewage treatment plant to predict the*
436 *nutrient removal efficiency using a long short-term memory (LSTM) neural network*.
437 Journal of Water Process Engineering. 37 (2020) 101388.
- 438 50. Hvala, N. and J. Kocijan, *Design of a hybrid mechanistic/Gaussian process model to*
439 *predict full-scale wastewater treatment plant effluent*. Computers & Chemical
440 Engineering. 140 (2020) 106934.
- 441 51. Hadi, S.J. and M. Tombul, *Forecasting Daily Streamflow for Basins with Different*
442 *Physical Characteristics through Data-Driven Methods*. Water Resources
443 Management. 32 (2018) 10 3405-3422.
- 444 52. Elkiran, G., V. Nourani, and S.I. Abba, *Multi-step ahead modelling of river water*
445 *quality parameters using ensemble artificial intelligence-based approach*. Journal of
446 Hydrology. 577 (2019) 123962.
- 447 53. Shuai, S., Guoren X., *Novel performance prediction model of a biofilm system*
448 *treating domestic wastewater based on stacked denoising auto-encoders deep*
449 *learning network*. Chemical Engineering Journal. (2018)
450 DOI: <https://doi.org/10.1016/j.cej.2018.04.087>