water

MDPI

Article

En-WBF: A Novel Ensemble Learning Approach to Wastewater Quality Prediction Based on Weighted BoostForest

Bojun Su [1], Wen Zhang [2,*], Rui Li [1,2], Yongsheng Bai [1] and Jiang Chang [1]

[1] Technology R&D Center of Beijing Drainage Group, Beijing 100044, China; subojun@bdc.cn (B.S.)
[2] School of Economics and Management, Beijing University of Technology, Beijing 100124, China
* Correspondence: zhangwen@bjut.edu.cn

Abstract: With the development of urbanization, the accurate prediction of effluent quality has become increasingly critical for the real-time control of wastewater treatment processes. The conventional method for measuring effluent biochemical oxygen demand (BOD) suffers from significant time delays and high equipment costs, making it less feasible for timely effluent quality assessment. To tackle this problem, we propose a novel approach called En-WBF (ensemble learning based on weighted BoostForest) to predict effluent BOD in a soft-sensing manner. Specifically, we sampled several independent subsets from the original training set by weighted bootstrap aggregation to train a series of gradient BoostTrees as the base models. Then, the predicted effluent BOD was derived by weighting the base models to produce the final prediction. Experiments on real datasets demonstrated that on the UCI dataset, the proposed En-WBF approach achieved a series of improvements, including by 28.4% in the MAE, 40.9% in the MAPE, 29.8% in the MSE, 18.2% in the RMSE, and 2.3% in the R^2. On the Fangzhuang dataset, the proposed En-WBF approach achieved a series of improvements, including by 8.8% in the MAE, 9.0% in the MAPE, 12.8% in the MSE, 6.6% in the RMSE, and 1.5% in the R^2. This paper contributes a cost-effective and timely solution for wastewater treatment management in real practice with a more accurate effluent BOD prediction, validating the research in the application of ensemble learning methods for environmental monitoring and management.

Keywords: wastewater treatment; soft measurement; biochemical oxygen demand; weighted boosting forest; ensemble learning

Citation: Su, B.; Zhang, W.; Li, R.; Bai, Y.; Chang, J. En-WBF: A Novel Ensemble Learning Approach to Wastewater Quality Prediction Based on Weighted BoostForest. *Water* **2024**, *16*, 1090. https://doi.org/10.3390/w16081090

Academic Editor: Zhenyao Shen

Received: 12 March 2024
Revised: 7 April 2024
Accepted: 9 April 2024
Published: 10 April 2024

1. Introduction

Wastewater treatment is of great significance in lowering urban pollution and promoting sustainable urban development with the development of industrialization [1,2]. Wastewater treatment plants, or WWTPs, are intricate industrial structures that function by utilizing a range of biological, chemical, and physical procedures in order to collect, recycle, and reuse wastewater for consistent compliance with regulations on the conservation of water environments [3,4]. Among these, the biochemical oxygen demand (BOD) is crucial for treating municipal wastewater since it is a dominant indicator of the effluent quality indicating the amount of organic contaminants in wastewater and is usually adopted in various water environment monitoring systems [5].

It is challenging for WWTPs to provide real-time monitoring and accurate effluent BOD measurements because the measurement of the BOD in effluent must be conducted under rigorous experimental circumstances [6], and the water quality sensor for BOD measurements is highly expensive [7]. To tackle this problem, the industry primarily uses the method of soft measurement [8] to construct the predictive relationship between the easily measured wastewater quality and the difficultly measured effluent BOD. In recent years, the success of data-driven methods across various domains has prompted a significant number of researchers to employ these approaches in wastewater treatment [9]. Currently, there are mainly two streams in predicting effluent BOD as the machine learning and the deep learning methods.

In the former, machine learning methods have demonstrated superior performance in the prediction of effluent BOD, with their advantages primarily manifesting in three key areas: interpretability, computational efficiency, and minimal data requirements. Firstly, machine learning models offer notable interpretability, simplifying the comprehension of the decision-making process. This is exemplified by Park et al. [10], who employed an interpretable machine learning approach to investigate the impact of input variable selection on model efficacy, utilizing Shapley additive explanations (SHAP) analysis to provide insightful explanations for model predictions. Secondly, in comparison to deep learning, machine learning models demand significantly fewer computational resources and exhibit rapid training capabilities, as illustrated by El-Rawy et al. [11] in their work on effluent quality prediction. Lastly, these models are adept at discerning effective patterns within relatively small datasets, a fact underscored by Zhang et al. [7] through the deployment of an enhanced gradient boosting regression tree (Miss-GBRT), thereby highlighting the supremacy of machine learning in scenarios of limited data availability. Given the susceptibility of individual machine learning models to noise, overfitting, and often limited prediction accuracy, researchers have increasingly employed ensemble learning approaches to mitigate these challenges [12,13]. However, it is notable that most ensemble learning strategies overlook the performance disparities among various base learners.

In the latter, deep learning has progressively taken the lead in solving industrial challenges in recent years due to its favorable accuracy [14]. The dynamic, nonlinear, and non-Gaussian behavior characteristics of wastewater quality can be successfully captured by deep learning, which is further used to reliably forecast the BOD [15]. For instance, the long short-term memory network (LSTM) [16] and the principal component analysis-enhanced nonlinear autoregressive network with exogenous inputs (PCA-NARX) [17] effectively capture these characteristics of wastewater quality but require extensive training data. Therefore, although deep learning models may be learned from the training data and can produce accurate predictions, the caliber and volume of the training data frequently affect how well they work and how broadly they can be applied [18]. Particularly, when there is scarce information available about the quality of the wastewater, it is challenging for researchers to identify potential patterns in the data. Thus, it is hard to train a model with small-size datasets to reliably predict the effluent BOD.

To address these problems, this paper makes use of ensemble learning to aggregate the prediction outcomes of a series of weighted base learners trained from sampled data to overcome the susceptibility of single machine learning models to noise, overfitting, and often limited prediction accuracy, the disadvantages of conventional ensemble learning methods, and the small sizes of training datasets in WWTPs [19], with the goal of enhancing the model's prediction performance. Compared to a single model, ensemble learning can better leverage the benefits of various models, lower the chance of overfitting, and increase model generalizability. In this paper, we propose a novel approach called En-WBF (**En**semble learning based on **W**eighted **B**oost**F**orest), using ensemble learning-based weighted BoostForest to predict the effluent BOD. First, the proposed En-WBF approach makes use of BoostTree [20] as the base model to exert its easily trained merits on a small-size dataset pertaining to the small sizes of training datasets in WWTPs. Furthermore, the proposed En-WBF approach can reduce the bias of the prediction model by iterative learning in the training process. In addition, the proposed En-WBF approach also uses the bootstrap aggregating algorithm (bagging) [20] to sample several independent datasets from the original training set, train multiple base models (i.e., BoostTrees), respectively, and produce the final BOD prediction outcome by the weighted summation of the prediction outcomes of the base models. Bagging ensemble learning can reduce the variance of the prediction model and can also solve the problem of small-sized training data to a certain extent.

The contributions can be summarized as follows: we propose a novel approach called En-WBF that uses weighted aggregation of base models (i.e., BoostTrees) to reliably predict the effluent BOD. The effluent BOD can achieve a satisfactory prediction result using the

boosting mechanism of BoostTree to decrease the model's prediction bias and the bagging mechanism of weighted BoostForest to reduce the model's prediction variance.

We propose using bootstrapping to sample independent training subsets for a series of BoostTrees to address the problem of small-sized datasets in WWTPs.

We conducted a comprehensive evaluation of our work on publicly available datasets and real wastewater treatment datasets. The results show that this method has good performance in predicting the effluent BOD.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 describes the problem as effluent BOD prediction. Section 4 proposes the novel En-WBF approach, including its overall architecture, BoostTree and its training process, as well as the research questions. Section 5 presents the conducted experiments. Section 6 concludes the paper.

2. Related Work

This section reviews two categories of related work on machine learning and deep learning methods to study prediction models for effluent BOD prediction. First, we summarize the research on machine learning methods for predicting the effluent BOD, such as the single model and ensemble model, and their drawbacks. Second, we present the studies of recent deep learning techniques and their applications in real practice.

2.1. Machine Learning Methods

Machine learning methods applied in wastewater quality prediction distinguish between two principal categories of approaches, i.e., single models and ensemble models. Due to the simplicity, interpretability, and low computational cost of single models, many studies have used them for predicting effluent BOD. For example, Wang et al. [21] developed a forward variable selection approach based on K-nearest-neighbor mutual information to eliminate redundant variables from the wastewater quality dataset and used support vector regression (SVR) to produce the effluent BOD. Zhang et al. [8] used an updated gradient boosting regression tree (GBRT) to impute the missing value of effluent BOD, accounting for the missing measurement of sewage indicators generated by anomalous sensors in the sewage treatment process. Wang et al. [22] employed random forest (RF) enhanced by latent Dirichlet allocation to reduce 12-dimension auxiliary feature vectors to 3-dimension feature vectors. Their experiments demonstrated that this approach can reduce data noise and redundant information while improving RF's prediction performance. Ching et al. [23] employed extreme gradient boosting (XG-Boost) to predict the concentration of effluent BOD in wastewater. A collection of weak regression trees was utilized to capture the dynamic and nonlinear behavior characteristics of the BOD in the wastewater. Liu et al. [24] used the enhanced relevance vector machine (RVM) to predict the effluent BOD concentration in wastewater treatment to address the complexity and uncertainty that come with wastewater quality.

Due to the vulnerability of single models to noise, overfitting, and their often lower predictive accuracy, several studies have turned to ensemble learning as a solution to these challenges. For instance, Sharafati et al. [12] introduced three novel integrated machine learning models: AdaBoost regression (ABR), gradient boosting regression (GBR), and random forest regression (RFR) to predict key effluent quality parameters, such as the total dissolved solids (TDS), the 5-day biochemical oxygen demand (BOD5), and chemical oxygen demand (COD), demonstrating that the efficacy of these models varies across different wastewater indices. Similarly, Nourani et al. [13] employed a composite artificial intelligence (AI) model comprising feedforward neural networks (FFNNs), support vector regression (SVR), and adaptive neuro-fuzzy inference system (ANFIS) for the prediction of effluent biological oxygen demand (BODeff) and chemical oxygen demand (CODeff) using daily data from real wastewater treatment plants (WWTPs). While ensemble learning has proven effective in effluent quality prediction, the differential performance of base learners has often been overlooked in the literature. Consequently, this paper proposes a weighted

approach to aggregating predictions from diverse base learners to enhance the accuracy of the final prediction outcome.

2.2. Deep Learning Methods

In order to accurately predict the effluent BOD, deep learning methods are proposed to characterize the complicated, non-dynamic, and non-Gaussian wastewater quality behavior by building a predictive relationship between the easily measured wastewater quality indicators and the difficultly measured wastewater quality indicators of the effluent BOD. In both academics and industry, the deep learning method has been widely adopted because of its favorable accuracy and real-time performance. Recently, many researchers have proposed various deep learning methods to predict the effluent quality during the wastewater treatment process. For instance, Foschi et al. [25] employed shallow artificial neural networks (ANNs) to predict the quality of sewage, and they reported that, although the shallow fully connected layer was unstable at mimicking the time dependency in the time series, the shallow ANN was superior at capturing the nonlinear correlations among the wastewater quality series. Based on time series analysis, Wongburi et al. [26] employed long short-term memory networks (LSTMs) to predict forecast sewage quality to resolve nonlinear and long-term dependent intricacies within the wastewater quality. Yang et al. [17] developed a principal component analysis–dynamic nonlinear autoregressive with exogenous inputs (PCA-NARX) model to forecast effluent quality. By employing various time delay parameters and training algorithms to refine the system performance, the model demonstrated exceptional predictive accuracy. Wang et al. [27] developed a prediction model that leverages a novel integration of multi-source data fusion, pattern decomposition, an enhanced Sparrow search algorithm (SSA), an attention (AT) mechanism, and gated recurrent unit (GRU) technology (i.e., GRU-AT). This model aims to effectively navigate the nonlinear, complex, and periodic nature of DO data sequences. In a related study, Satish et al. [28] combined climatic and geospatial factors to construct a model that delineates the causal relationships among urban land use elements. Building on this foundation, they introduced a stacked ANN ensemble model, demonstrating its superior performance over conventional single machine learning approaches.

While deep learning methods are pretty much effective in predicting effluent BOD, most of them overlook the issue of small-sized data in real practice. When dealing with small-sized datasets, there is large bias and small variation in the prediction outcomes with deep learning methods. Nevertheless, some deep learning models are sensitive to outliers, which may lead to model overfitting in the training phase. To tackle this problem, we employed a hybrid approach that integrates both bagging and boosting techniques, leveraging their respective strengths to achieve notable enhancements in model prediction performance. Specifically, bagging enhances model stability by aggregating multiple complex base learners, thereby reducing prediction variance. Conversely, boosting reduces model bias by sequentially combining multiple simple base learners. Furthermore, in the bagging approach, weighting base learners differently based on their performance further improves model efficacy.

3. En-WBF Model Development

This section proposes an ensemble learning approach called En-WBF to predict effluent BOD considering the dynamic, nonlinear, and non-Gaussian behavior of the wastewater quality.

3.1. Problem Statement

Accurate prediction of the biological oxygen demand (BOD) in wastewater is crucial for treatment outcome assessment and environment protection in operating the wastewater treatment plants. A large amount of data is needed to train deep learning models, which usually cannot be satisfied in WWTPs. Meanwhile, traditional statistics-based models are incapable of dealing with complex variations in sewage water quality when they are used to predict the BOD. For this reason, this paper proposes using ensemble learning to predict

the effluent BOD. Assume that the effluent BOD is a real-value scalar $y \in \mathbb{R}$, and the other wastewater quality is a vector $x \in \mathbb{R}^D$. The task of prediction is to learn a prediction model g, as shown in Equation (1).

$$y = g(x) \tag{1}$$

3.2. Overall Architecture of En-WBF

The proposed En-WBF approach combines boosting and bagging to predict the effluent BOD. The boosting algorithm can integrate simple base learners to reduce bias, and the bagging algorithm can sample different subsets of training data to train individual base learners to reduce variance. This combination of boosting and bagging has the potential to significantly address the issue of accurately predicting the effluent BOD with small-sized datasets in WWTPs [8].

To guarantee the mutual independence of each sample dataset, the proposed En-WBF approach, as illustrated in Figure 1, firstly samples K subsets randomly from the original training dataset D as $\{D_1, \ldots, D_K\}$ using the bootstrap procedure. Secondly, the proposed En-WBF approach uses each sampled subset D_i to train a BoostTree T_i. The basic idea of BoostTree is to apply gradient boosting to a single decision tree, which can be well adapted to the dynamic and nonlinear characteristics of effluent BOD. The goal of the proposed En-WBF approach is to take into account the performance difference between each base learner (i.e., BoostTree) in an ensemble learning manner. By feeding each validation set $\overline{D}_i$ into the trained BoostTree T_i, it obtains the goodness of fit R_i^2 and the prediction outcome $\hat{y}_i$. Then, the weight w_i is obtained by Equation (2). Ultimately, the proposed En-WBF approach weights each prediction $\hat{y}_i$ using weight w_i and sums them to produce the final prediction $\hat{y}$, as shown in Equation (3).

$$w_i = \frac{R_i^2}{\sum_{j-1}^{K} R_j^2} \tag{2}$$

$$\hat{y} = \sum_{i=1}^{K} w_i \hat{y}_i \tag{3}$$

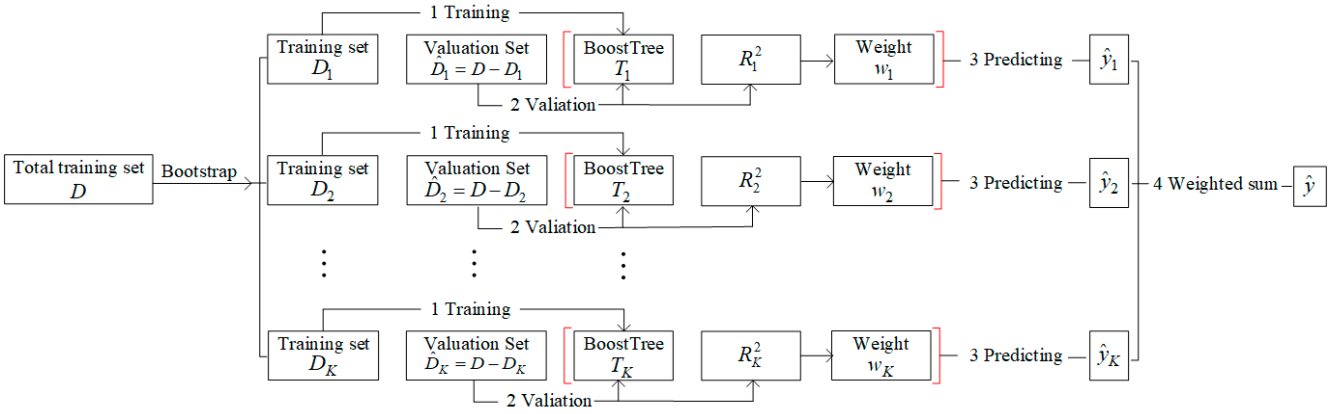

Figure 1. The overall architecture of En-WBF.

3.3. BoostTree

BoostTree makes use of boosting gradients and node functions to train a linear or nonlinear model on each node. The given input is first sorted into a leaf node by BoostTree, as seen in Figure 2, and the output of each node model along the path from the root node to that leaf node is then summarized to determine the final predicted outcome. Distinct from gradient-boosted regression trees (GBRTs) [8], which enhance model performance by sequentially incorporating new trees to address errors within the ensemble, BoostTree

innovatively integrates gradient boosting principles directly within the growth of a single decision tree. This integration is facilitated by the random selection of cut-points during node splitting, a strategy that significantly increases model diversity (i.e., randomness), contrasting with the ensemble-level boosting characteristic of GBRT.

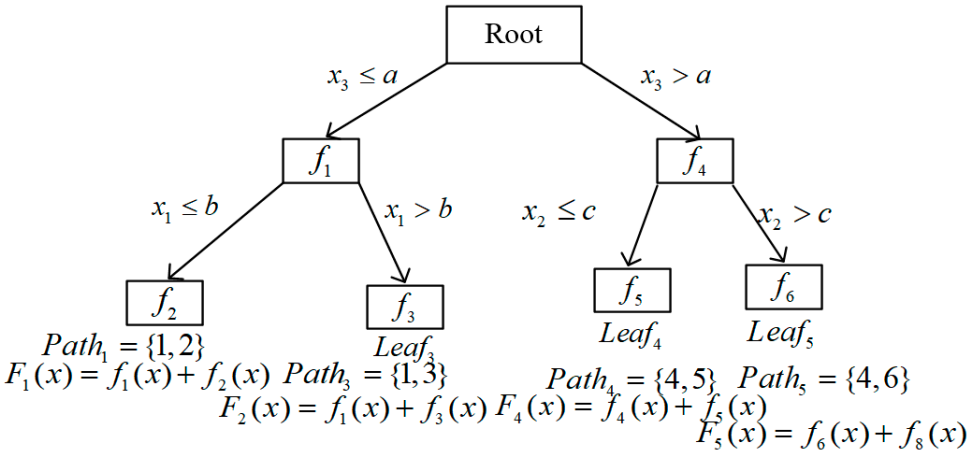

Figure 2. An example of BoostTree.

Assume that a BoostTree has M nodes except the root node. For the mth node (m $\in [1, M]$), BoostTree will train a node function f_m and predict its output using Equation (4), as follows.

$$\hat{y}_n = F_m(x_n) = \sum_{m \in Path_{q(x_n)}} f_m(x_n) \tag{4}$$

where $Path_{q(x_n)}$ is the set of node indexes of the sample along the path from the root node to the leaf node $q(x_n)$. To minimize the difference between the real value and predicted outcome during the training phase, BoostTree minimizes the objective function (i.e., the loss function) with Equation (5).

$$Loss(F) = \sum_{n=1}^{N} l(y_n, \hat{y}_n) + \sum_{m=1}^{M} \lambda \Omega(f_m) \tag{5}$$

The difference between the real value and the predicted outcome is measured by the error term, which is the first term in Equation (5). The regulation in the second term in Equation (5) keeps BoostTree from overfitting. The regulation coefficient is denoted by λ. In this paper, we set $\lambda = 1$ in accordance with the suggestion documented in the literature [20], and Ω is the complexity of the tree, which is controlled by the maximum number of leaves on the tree.

In general, it is impossible to explicitly optimize the objective function in Equation (4). Therefore, BoostTree minimizes Equation (5) in an additional way, which is inspired by the gradient-boosting algorithm. Assume that a BoostTree has $T(T \ge 2)$ leaf nodes after a $T-1$ round iteration and the number of non-leaf nodes should be $M = 2T-2$. As a result, Equation (5) can be changed to Equation (6), as follows.

$$Loss(F) = \sum_{m=1}^{T} LeafLoss_m + \sum_{m=1}^{2T-2} \Omega(f_m) \tag{6}$$

where

$$LeafLoss_m = \sum_{n \in I_m} l(y_n, \sum_{i \in Path_m} f_i(x_n)) \tag{7}$$

$$I_m = \{n | q(x_n) = m\} \tag{8}$$

Here, I_m is the collection of all training samples that are associated with the leaf node m, and $LeafLoss_m$ is a measure of that node's loss. A greedy learning technique is used by BoostTree to add branches to the leaf node (i.e., split leaf node) that has the greatest loss in each iteration.

Presume node m is the leaf node with the greatest loss. After the node splitting, BoostTree splits I_m into two subsets: the left node's sample set I_L, and the right node's sample set I_R. If we assume that f_L and f_R are the left and right node functions that were trained by I_L and I_R, respectively, then the decreased loss of Equation (5) is shown in Equation (9), as follows.

$$\delta_{Loss} = C - Loss(f_L) - Loss(f_R) \tag{9}$$

where

$$C = \sum_{n \in I_m} l(y_n, F_m(x_n)) \tag{10}$$

$$Loss(f_L) = \sum_{n \in I_L} l(y_n, F_m(x_n) + f_L(x_n)) + \Omega(f_L) \tag{11}$$

$$Loss(f_R) = \sum_{n \in I_R} l(y_n, F_m(x_n) + f_R(x_n)) + \Omega(f_R) \tag{12}$$

where F_m is the collection of models along the path from the root node to the leaf node, and C is a constant. The left and right child nodes' respective loss functions are denoted by $Loss(f_L)$ and $Loss(f_R)$. Meanwhile, Equation (9) can be optimized by minimizing $Loss(f_L)$ and $Loss(f_R)$, with node splitting and gradient lifting. One can refer to [20] for more details.

3.4. Weighted BoostForest

Multiple BoostTrees were integrated into a forest by using Weighted BoostForest (En-WBF). Prior to obtaining the final prediction outcome by weighting (see Equation (3)), the proposed En-WBF approach firstly sampled K training subsets using bootstrap. Next, it trained a BoostTree on each sampled subset. Finally, it input each validation set into the trained BoostTree to produce the weight of each base learner.

The issue of limited wastewater quality data can be successfully resolved with this weighted integration method, which takes into account the performance variations of each base learner to improve the accuracy of the produced prediction outcome. The training pseudo-code that designates the improved forest is provided by Algorithm 1, as follows.

Algorithm 1 Training process of WBF.

Input: $Data = \{(x_n, y_n)\}_{n=1}^{N}$, N is the number of training samples, $x_n \in \mathbb{R}^{D \times 1}$; *n_estimators* is the number of BoostTree, *Bootstrap_rate* is the sampling proportion;
Output: *WeightedBoostForest*
1: Initialize *WeightedBoostForest* = {}
2: **For** i = 1: *n_estimators* **do**
3: Sample $Data'$ from *Data* according to *Bootstrap_rate*
4: Train $BoostTree_i$ on $Data'$
5: Add $BoostTree_i$ to *WeightedBoostForest*
6: **End**

3.5. Research Questions

We developed four research questions (RQs) to direct an examination of how well the proposed En-WBF approach predicted the effluent BOD, as listed in Table 1. The proposed En-WBF approach has an advantage over conventional machine learning techniques in that it requires fewer training data. As a result, the first RQ investigates the performances of the proposed En-WBF approach and the traditional machine learning techniques in predicting BOD. One benefit of the proposed En-WBF approach is that it aggregates the base learners' expected outcomes in a weighted manner. In order to investigate the performance comparison between the proposed En-WBF approach and the anticipated outcomes of the

average aggregation (i.e., AveragedBoostForest, ABF), ablation studies were devised for the second RQ.

Table 1. Research questions.

ID	Question
RQ1	How are the performances of the proposed WBF approach compared with that of traditional machine learning methods in predicting effluent BOD?
RQ2	Is weighted polymerization better than average polymerization for BoostForest?
RQ3	What are the optimal parameter settings for the proposed En-WBF approach?
RQ4	How does the proposed En-WBF approach perform on actual wastewater treatment data?

Furthermore, there are strong correlations among the parameters used during the training and the performance of the proposed En-WBF approach. The purpose of the third RQ was to facilitate our analysis of the ideal parametric configuration for the proposed En-WBF approach. Lastly, the fourth RQ investigates the practical application of the proposed En-WBF approach on actual wastewater treatment plant datasets in order to decide whether the proposed En-WBF approach can be used in real wastewater treatment operations.

4. Experiment

This section provides an overview of the dataset, experimental settings, and evaluation measures used. It further presents and discusses the experimental results addressing the four research questions (RQs).

4.1. Dataset

In our experiments, we evaluated the performances of the proposed En-WBF approach using a dataset that was acquired from a WWTP. This dataset has 527 instances and 38 water quality variables that were taken from the UCI machine learning repository (UCI WWTP dataset, https://archive.ics.uci.edu/dataset/106/water+treatment+plant, accessed on 20 January 2024). Daily sensor measurements from an urban WWTP make up the data. In order to predict the effluent BOD using soft measurements, as indicated in Table 2, 1379 complete data were collected by eliminating rows with missing values. The best auxiliary input variable set was then filtered out using the *K*-nearest-neighbor mutual information forward variable selection approach, as per the literature [21], as shown in Table 2. All experiments were performed on a computer equipped with AMD Ryzen 5800H, 3.20 GHz CPU, NVIDIA GeForce RTX 3060 GPU (Nvidia, Santa Clara, CA, USA), and the Windows 10 operating system.

Table 2. Optimal auxiliary input variable set and its description on UCI dataset.

Notation	Description
DBO-D	Input biological demand of oxygen to secondary settler
SS-D	Input suspended solids to secondary settler
CONE-D	Input conductivity to secondary settler
DQO-S	Output chemical demand of oxygen
SED-S	Output sediments
RD-DBO-S	Performance input biological demand of oxygen to secondary settler
RD-SS-G	Global performance input suspended solids

4.2. Evaluation Metrics

This research used seven evaluation metrics, including the mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), root mean squared error (RMSE), coefficient of determination (R^2), Akaike information criterion (AIC) [29], and

Bayesian information criterion (BIC) [29], as shown in Equations (13)–(19), to evaluate the performances of the proposed En-WBF approach in comparison with the baseline methods. In order to mitigate the impact of random sampling, we repeated each experiment ten times to average its performance. Ultimately, the average performance values of the ten repetitions were used to derive these seven metrics.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{13}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{14}$$

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{17}$$

$$AIC = n \log(\frac{RSS}{n}) + 2K \tag{18}$$

$$BIC = n \log(\frac{RSS}{n}) + K \log(n) \tag{19}$$

where y_i is the predicted value of the ith sample, $\hat{y}_i$ is the corresponding true value of the ith sample, n is the number of samples, RSS is the residual sum of squares, K is the number of parameters. MAE quantifies the average magnitude of prediction errors, providing a straightforward measure of prediction accuracy without directionality. MAPE offers an intuitive percentage-based assessment of prediction errors, facilitating comparability across different scales of data. MSE captures the average of the squares of the errors, emphasizing larger errors, which is particularly useful for highlighting significant prediction deviations. RMSE provides error magnitudes in the same units as the predicted values, making their interpretation more tangible. R^2 indicates the proportion of variance in the dependent variable predictable from the independent variables, reflecting the model's explanatory power. AIC affords a measure of the relative quality of statistical models for a given dataset, penalizing model complexity to deter overfitting. BIC expands upon AIC by introducing a stronger penalty for the number of parameters, balancing the model fit against its complexity in a Bayesian context.

4.3. Experimental Results

4.3.1. RQ1: How Does the Performance of the Proposed WBF Approach Compare with That of Traditional Machine Learning Methods in Predicting Effluent BOD?

The proposed En-WBF approach was evaluated in comparison with seven baseline methods in predicting effluent BOD, including SVR [21], GBRT [8], RF [22], XG-Boost [23], RVM [24], ANN [25], and LSTM [16], in order to compare its prediction performance. The parameter descriptions and values of the above seven baseline methods are shown in Table 3. In addition, Figure 3 shows the performance validation of En-WBF on the prediction of effluent BOD. It depicts an elementwise comparison of the measured BOD values (denoted by blue circles connected with a line) and the corresponding predicted BOD values (indicated by red triangles connected with a line). The plotted data points show the fluctuations in the BOD concentration over the samples. Both the measured

and predicted values appear to follow a similar trend across the dataset, illustrating the En-WBF's predictive capability.

Table 3. Parameter settings and description of the baseline methods.

Method	Parameter	Description
SVR	SVR_c = 1, SVR_{Kernel} = 'rbf'	SVR_c is the regularization parameter of SVR; SVR_{Kernel} is the kernel type of SVR;
GBRT	n_estimators = 100, min_samples_leaf = 5	n_estimators is the number of estimators of GBRT; min_samples_leaf is the minimum sample number of GBRT leaf nodes;
RF	n_estimators = 110, min_samples_leaf = 5	n_estimators is the number of estimators of RF; min_samples_leaf is the minimum sample number of RF leaf nodes;
XG-Boost	n_estimators = 100, max_depth = 4, max_leaves = 5	n_estimators is the number of estimators of XG-Boost; max_depth is the maximum depth of XG-Boost; max_leaves is the maximum number of leaf nodes per tree of XG-Boost;
RVM	RVM_{Kernel} = 'linear'	RVM_{Kernel} is the kernel type of RVM;
ANN	n_{first} = 32, n_{second} = 16	n_{first} is the number of hidden units in the first dense layer; n_{second} is the number of hidden units in the second dense layer;
LSTM	n_{lstm} = 32	n_{lstm} is the number of hidden units in the LSTM layer.

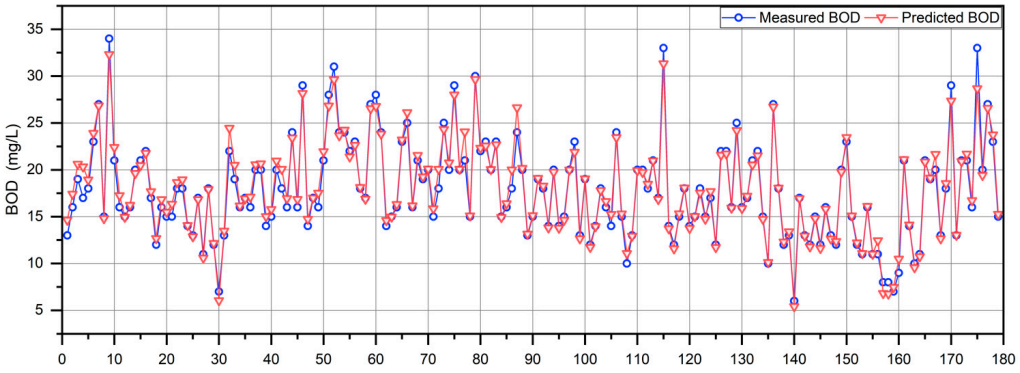

Figure 3. Performance validation of En-WBF on the prediction of effluent BOD.

Table 4 shows the seven metrics of the effluent BOD predicted by the proposed En-WBF approach and the seven baseline methods. It can be seen that the performance of the proposed En-WBF approach for five metrics is better than that of the baseline methods of effluent BOD prediction, with an MAE of 0.2573, MAPE of 0.0171, MSE of 0.1869, RMSE of 0.4323, and R^2 of 0.9938. The proposed En-WBF approach outperformed the SVR model, which achieved the second-best prediction performance, by 28.4% in the MAE, 40.9% in the MAPE, 29.8% in the MSE, 18.2% in the RMSE, and 2.3% in the R^2. This reveals that the proposed En-WBF approach, by repeatedly sampling the training sets, can effectively address the problem of the small size of the UCI dataset when compared to the other baseline effluent BOD prediction methods. Although the proposed EN-WBF approach

exhibited a slightly lower performance in terms of the AIC and BIC metrics compared to the SVR model, it nonetheless secured the second-best position, achieving an AIC of 147.7845 and a BIC of 861.7589. This indicates that the En-WBF model demonstrates a robust fit to the data.

Table 4. Comparison results between the proposed En-WBF approach and the baseline methods.

Method	MAE	MAPE	MSE	RMSE	R^2	AIC	BIC
En-WBF	**0.2573**	**0.0171**	**0.1869**	**0.4323**	**0.9938**	147.7845	861.7589
SVR	0.3667	0.0241	0.2610	0.5109	0.9713	**99.5610**	**641.4165**
GBRT	0.8821	0.0586	1.6393	1.2803	0.9452	572.4742	1343.8216
RF	1.0001	0.0778	2.9375	1.7139	0.9018	1128.8830	2620.5796
XG-Boost	0.8734	0.0559	1.7999	1.3416	0.9398	1029.2039	2501.7761
RVM	1.2383	0.0771	2.5810	1.6066	0.9137	183.7237	206.0354
ANN	0.8454	0.0491	1.5282	1.1577	0.9467	1677.9122	4231.0082
LSTM	0.8373	0.0447	1.4913	1.2211	0.9501	2469.7556	6348.8042

It is possible to alleviate the impact of some overfitting base learners on the final prediction by weighting the base learner's contribution. Furthermore, the proposed En-WBF approach does a good job in capturing the dynamic and nonlinear behaviors of effluent BOD. The findings demonstrate that the proposed En-WBF approach can mitigate the noise issue in wastewater quality by integrating different learning models, reduce the model prediction variance, and alleviate the possible overfitting of each base learner.

4.3.2. RQ2: Is Weighted Polymerization Better Than Average Polymerization for BoostForest?

The authors of this paper removed the weighting in the proposed En-WBF approach and devised a comparative experiment to see whether there was any improvement in order to further investigate the efficacy in weighting the base learners. Equation (20) illustrates how average BoostForest (ABF) produced the final prediction result by averaging the predictions made by each base learner.

$$\hat{y} = \frac{1}{K}\sum_{i=1}^{K}\hat{y}_i \tag{20}$$

The result of the ablation study is shown in Table 5. It is evident that the proposed En-WBF approach yielded a better MAE and MSE than the ABF technique. The proposed En-WBF approach increased the MAE by 1.49%, the MAPE by 1.73%, the MSE by 3.01%, the RMSE by 1.53%, the R^2 by 4.95%, the AIC by 2.57%, and the BIC by 2.04% when compared to the ABF technique. These outcomes demonstrate the effectiveness of the weighting strategy for BoostForest and its ability to accommodate the variations in the performance of each base learner while alleviating the impacts of both possible overfitting and underfitting within BoostTree.

Table 5. The results of the ablation study.

Method	MAE	MAPE	MSE	RMSE	R^2	AIC	BIC
En-WBF	**0.2573**	**0.0171**	**0.1869**	**0.4323**	**0.9938**	147.7845	861.7589
ABF	0.2612	0.0174	0.1927	0.4390	0.9889	151.6813	879.7123

4.3.3. RQ3: What Are the Optimal Parameter Settings of the Proposed En-WBF Approach?

The training process of the proposed En-WBF approach involves three key parameters influencing its performance, including the number of learners at the learning base *n_estimators*, the sampling ratio *Bootstrap_rate*, and the minimal sample number leaf nodes min_*samples_leaf*. *n_estimators* dictates the total count of base learners that constitute the

ensemble learning. Multiple learners are trained to solve the same problem, and their predictions are combined in some manner (e.g., averaging or weighted averaging) to produce the final output. The number of learners is crucial because it directly influences the model's ability to capture complex patterns in the data. *Bootstrap_rate* controls the proportion of the training dataset to be used for training each base learner within the ensemble. This is a strategy often associated with bagging (bootstrap aggregating) techniques, where each learner is trained on a random subset of the data. In tree-based models, min_$samples_leaf$ specifies the minimum number of samples a leaf node must have. This threshold acts as a constraint on tree growth, preventing the model from creating leaves that only contain a small number of samples.

We employed MAE and MSE as performance metrics for parameter tuning, subsequently applying a greedy algorithm [30] to iteratively tune the above principal parameters. First, we adjusted *n_estimators* to the desired value and set the other two parameters to their default settings. Second, we changed *Bootstrap_rate* so that *n_estimators* was set to the ideal value and $min\_samples\_leaf$ was the default value. Lastly, we found the ideal value for the remaining two parameters by adjusting $min\_samples\_leaf$.

The outcomes of the parameter adjustment are shown in Tables 6–8. As can be seen, the best MAE and MSE were 0.2718 and 0.2108, respectively, when $n\_estimators = 40$. When the sample ratio *Bootstrap_rate* was varied and *n_estimators* was set to 40, it was found that the best MAE and MSE were 0.2675 and 0.1996, respectively, with $Bootstrap\_rate = 0.75$. When $min\_samples\_leaf$ was set as 5, the proposed En-WBF approach performed the best, with an MAE of 0.2573 and an MSE of 0.1869, with settings of $n\_estimators = 40$ and $Bootstrap\_rate = 0.75$ in this case.

Table 6. Parameter tuning of *n_estimators* with *Bootstrap_rate* = 0.75 and *min_samples_leaf* = 5.

n_estimators	MAE	MSE
10	0.3263	0.3134
20	0.3085	0.2819
30	0.2910	0.2612
40	**0.2718**	**0.2108**
50	0.2744	0.2176
60	0.2789	0.2178
70	0.2884	0.2223

Table 7. Parameter tuning of *Bootstrap_rate* with *n_estimators* = 40 and *min_samples_leaf* = 5.

Bootstrap_rate	MAE	MSE
0.60	0.3478	0.3896
0.65	0.2909	0.2713
0.70	0.3027	0.2408
0.75	**0.2675**	**0.1996**
0.80	0.2775	0.2018
0.85	0.2847	0.2040
0.90	0.2886	0.2060

Table 8. Parameter tuning of *min_samples_leaf* with *Bootstrap_rate* = 0.75 and *n_estimators* = 40.

min_samples_leaf	MAE	MSE
1	0.4268	0.3449
2	0.3892	0.2905
3	0.3384	0.2453
4	0.3032	0.2209
5	**0.2573**	**0.1869**
6	0.2738	0.2184
7	0.2828	0.2223

4.3.4. RQ4: How Does the Proposed En-WBF Approach Perform on Real Wastewater Treatment Data?

We conducted experiments on a real wastewater quality dataset, which included a total of 414 instances and 11 variables after data preprocessing, collected by the Beijing Drainage Group's Fangzhuang Wastewater Treatment Plant in order to confirm the efficacy of the proposed En-WBF approach. Using the K-nearest neighbor mutual information forward variable selection method, we identified an optimal set of seven auxiliary input variables, detailed in Table 9. Table 10 shows the effluent BOD prediction performance of the proposed En-WBF approach. The proposed En-WBF approach significantly performed better than SVR, GBRT, RF, XG-Boost, RVM, ANN and LSTM, with its MAE of 3.0647, MAPE of 0.2272, MSE of 17.9724, RMSE of 4.2392, and R^2 of 0.6838. The proposed En-WBF approach outperformed the second-best SVR by 8.8%, 9.0%, 12.8%, 6.6%, and 1.5%. respectively, on these five metrics. In addition, En-WBF achieved an AIC of 500.2910 and a BIC of 1113.9190, second only to SVR. This demonstrates that the proposed En-WBF approach, which can be used in wastewater treatment, has a considerable advantage over the baseline methods for predicting effluent BOD.

Table 9. Optimal auxiliary input variable settings and descriptions on Fangzhuang dataset.

Notation	Description
Q-E	Effluent flow rate
MLSS	Mixed liquor suspended solids
COD-I	Influent biological demand of oxygen
BOD-I	Influent chemical demand of oxygen
SS-I	Influent suspended solids
P-I	Influent phosphorus
P-E	Effluent phosphorus

Table 10. Model validation of En-WBF on real wastewater treatment dataset.

Method	MAE	MAPE	MSE	RMSE	R^2	AIC	BIC
En-WBF	**3.0647**	**0.2272**	**17.9724**	**4.2392**	**0.6838**	637.1084	1113.9190
SVR	3.3431	0.2494	20.5808	4.5366	0.6738	**500.2910**	**835.2406**
GBRT	3.3529	0.2502	20.8585	4.5671	0.6529	697.0014	1225.0396
RF	3.3678	0.2498	21.2829	4.6133	0.6501	1010.0689	1845.4727
XG-Boost	3.4259	0.2548	21.5569	4.6429	0.6233	1138.0689	2099.5714
RVM	3.6679	0.2749	23.7362	4.8720	0.5991	171.1084	188.8410
ANN	3.3542	0.2496	20.9333	4.5752	0.6317	1763.1911	3341.3949
LSTM	3.4317	0.2559	21.7285	4.6613	0.6091	2597.1671	4995.0124

Table 10 reveals that the En-WBF model exhibited suboptimal performance across the MAPE, R^2, and other metrics. The impact of outliers or noise was pronounced in the MAPE, where percentage errors can be significantly inflated by outliers. These outliers, if unaligned with predominant trends, can also depress the R^2 values. Consequently, to enhance the efficacy of En-WBF, future research will consider incorporating a refined outlier and noise management module.

5. Discussion

This study introduced a novel ensemble learning approach, En-WBF, aimed at enhancing the prediction accuracy of effluent biochemical oxygen demand (BOD) in WWTPs. The En-WBF approach was designed to address several prevalent challenges in the field, including the propensity for single models to under-fit and the variability in performance among traditional ensemble learners, especially when confronted with small-scale wastewater quality datasets.

One of the core strengths of En-WBF lies in its innovative use of the bootstrap procedure to create independent data subsets, which mitigates the limitations imposed by

small datasets. Furthermore, the training of multiple BoostTrees, aimed at overcoming the under-fitting commonly associated with single-model approaches, and the strategic weighting of these BoostTrees based on performance, significantly enhances the model's prediction accuracy. These methodological choices underscore the potential of En-WBF as a robust solution for the BOD prediction of WWTPs.

The evaluation of En-WBF through established research questions revealed its superior performance in comparison to conventional machine learning and deep learning techniques. Specifically, an ablation study on the weighting strategy confirmed its effectiveness, and a sensitivity analysis on key hyper-parameters helped in identifying the optimal model configuration. Validation of the model using the Fangzhuang dataset further demonstrated its proficiency across various metrics.

Despite these strengths, this study acknowledges certain limitations of the En-WBF approach. Primarily, while the model performed well with small datasets, its efficacy in contexts with extensive data, particularly when compared against advanced deep learning techniques, remains uncertain [16,17]. Additionally, its reliance on basic methods for missing value imputation and outlier detection may undermine the model's performance [8]. These limitations highlight the necessity for further research and refinement of the En-WBF approach.

In summary, when assessed against other state-of-the-art BOD prediction models documented in the literature, En-WBF exhibits notable advantages, particularly in terms of adaptability and precision. However, the scalability of En-WBF and its ability to manage the complexities of diverse wastewater treatment scenarios warrant further investigation.

6. Conclusions and Future Work

The En-WBF model represents a significant advancement in the prediction of effluent BOD in WWTPs, addressing critical challenges in the field through an innovative ensemble learning approach. The model's development, centered around the bootstrap procedure and the strategic weighting of BoostTrees, marks a step forward in enhancing the accuracy of BOD predictions. In conclusion, the En-WBF model offers a promising solution to the challenges of effluent BOD prediction in WWTPs. Through continuous improvement and expansion, it has the potential to significantly contribute to the field of environmental science and engineering.

Future work of this research will aim to extend the capabilities of the En-WBF model. This includes the integration of advanced modules for outliers and noise removal to further improve performance, especially for larger and more complex datasets. Additionally, considering the significant costs associated with effluent BOD prediction, its scope will be expanded to include more wastewater quality indicators, such as the chemical oxygen demand (COD), total nitrogen (TN), and ammonium nitrogen (NH_4-N). These expansions will not only enhance the model's capability but also contribute to more comprehensive environmental monitoring and management.

Author Contributions: Conceptualization, J.C.; Methodology, W.Z. and Y.B.; Writing—original draft, B.S. and R.L.; Writing—review & editing, B.S. and R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Beijing Natural Science Fund grant number 9222001; National Natural Science Foundation of China grant numbers 72174018 and 71932002.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Authors Bojun Su, Rui Li, Yongsheng Bai and Jiang Chang were employed by the company Beijing Drainage Group. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Huang, J.; Qian, R.; Gao, J.; Bing, H.; Huang, Q.; Qi, L.; Huang, J. A novel framework to predict water turbidity using Bayesian modeling. *Water Res.* **2021**, *202*, 117406. [CrossRef] [PubMed]
2. Saravanan, A.; Kumar, P.S.; Jeevanantham, S.; Karishma, S.; Tajsabreen, B.; Yaashikaa, P.R.; Reshma, B. Effective water/wastewater treatment methodologies for toxic pollutants removal: Processes and applications towards sustainable development. *Chemosphere* **2021**, *280*, 130595. [CrossRef] [PubMed]
3. Tang, W.; Pei, Y.; Zheng, H.; Zhao, Y.; Shu, L.; Zhang, H. Twenty years of China's water pollution control: Experiences and challenges. *Chemosphere* **2022**, *295*, 133875. [CrossRef] [PubMed]
4. Fathollahi-Fard, A.M.; Ahmadi, A.; Al-e-Hashem, S.M. Sustainable closed-loop supply chain network for an integrated water supply and wastewater collection system under uncertainty. *J. Environ. Manag.* **2020**, *275*, 111277. [CrossRef] [PubMed]
5. Luo, L.; Dzakpasu, M.; Yang, B.; Zhang, W.; Yang, Y.; Wang, X.C. A novel index of total oxygen demand for the comprehensive evaluation of energy consumption for urban wastewater treatment. *Appl. Energy* **2019**, *236*, 253–261. [CrossRef]
6. Zhu, J.J.; Kang, L.; Anderson, P.R. Predicting influent biochemical oxygen demand: Balancing energy demand and risk management. *Water Res.* **2018**, *128*, 304–313. [CrossRef] [PubMed]
7. Wang, G.; Jia, Q.S.; Zhou, M.; Bi, J.; Qiao, J.; Abusorrah, A. Artificial neural networks for water quality soft-sensing in wastewater treatment: A review. *Artif. Intell. Rev.* **2022**, *55*, 565–587. [CrossRef]
8. Zhang, W.; Li, R.; Zhao, J.; Wang, J.; Meng, X.; Li, Q. Miss-gradient boosting regression tree: A novel approach to imputing water treatment data. *Appl. Intell.* **2023**, *53*, 22917–22937. [CrossRef]
9. Bahramian, M.; Dereli, R.K.; Zhao, W.; Giberti, M.; Casey, E. Data to intelligence: The role of data-driven models in wastewater treatment. *Expert Syst. Appl.* **2023**, *217*, 119453. [CrossRef]
10. Park, J.; Lee, W.H.; Kim, K.T.; Park, C.Y.; Lee, S.; Heo, T.Y. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. *Sci. Total Environ.* **2022**, *832*, 155070. [CrossRef] [PubMed]
11. El-Rawy, M.; Abd-Ellah, M.K.; Fathi, H.; Ahmed, A.K.A. Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. *J. Water Process Eng.* **2021**, *44*, 102380. [CrossRef]
12. Sharafati, A.; Asadollah, S.B.H.S.; Hosseinzadeh, M. The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. *Process Saf. Environ. Prot.* **2020**, *140*, 68–78. [CrossRef]
13. Nourani, V.; Asghari, P.; Sharghi, E. Artificial intelligence based ensemble modeling of wastewater treatment plant using jittered data. *J. Clean. Prod.* **2021**, *291*, 125772. [CrossRef]
14. Zhan, C.; Zhang, X.; Yuan, J.; Chen, X.; Zhang, X.; Fathollahi-Fard, A.M.; Wang, C.; Wu, J.; Tian, G. A hybrid approach for low-carbon transportation system analysis: Integrating CRITIC-DEMATEL and deep learning features. *Int. J. Environ. Sci. Technol.* **2024**, *21*, 791–804. [CrossRef] [PubMed]
15. Yang, C.; Zhang, Y.; Huang, M.; Liu, H. Adaptive dynamic prediction of effluent quality in wastewater treatment processes using partial least squares embedded with relevance vector machine. *J. Clean. Prod.* **2021**, *314*, 128076. [CrossRef]
16. Yang, B.; Xiao, Z.; Meng, Q.; Yuan, Y.; Wang, W.; Wang, H.; Feng, X. Deep learning-based prediction of effluent quality of a constructed wetland. *Environ. Sci. Ecotechnol.* **2023**, *13*, 100207. [CrossRef]
17. Yang, Y.; Kim, K.R.; Kou, R.; Li, Y.; Fu, J.; Zhao, L.; Liu, H. Prediction of effluent quality in a wastewater treatment plant by dynamic neural network modeling. *Process Saf. Environ. Prot.* **2022**, *158*, 515–524. [CrossRef]
18. Abou Houran, M.; Bukhari SM, S.; Zafar, M.H.; Mansoor, M.; Chen, W. COA-CNN-LSTM: Coati optimization algorithm-based hybrid deep learning model for PV/wind power forecasting in smart grid applications. *Appl. Energy* **2023**, *349*, 121638. [CrossRef]
19. Cui, S.; Yin, Y.; Wang, D.; Li, Z.; Wang, Y. A stacking-based ensemble learning method for earthquake casualty prediction. *Appl. Soft Comput.* **2021**, *101*, 107038. [CrossRef]
20. Zhao, C.; Wu, D.; Huang, J.; Yuan, Y.; Zhang, H.T.; Peng, R.; Shi, Z. BoostTree and BoostForest for ensemble learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8110–8126. [CrossRef] [PubMed]
21. Wang, W.; Yang, C.H.; Han, J.; Yi, Y.G. Forward variable selection method based on k-nearest neighbor mutual information and its application on soft sensor modeling of water quality parameters. *Syst. Eng. Theory Pract.* **2022**, *42*, 253–261. (In Chinese)
22. Wang, Y.; Lu, W.; Zuo, C.H.; Bao, M.Y. Research on Water Quality BOD Prediction Based on Improved Random Forest Model. *Chin. J. Sens. Actuators* **2021**, *34*, 1482–1488. (In Chinese)
23. Ching PM, L.; Zou, X.; Wu, D.; So, R.H.Y.; Chen, G.H. Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine. *Environ. Res.* **2022**, *210*, 112953. [CrossRef] [PubMed]
24. Liu, H.; Yang, C.; Huang, M.; Yoo, C. Soft sensor modeling of industrial process data using kernel latent variables-based relevance vector machine. *Appl. Soft Comput.* **2020**, *90*, 106149. [CrossRef]
25. Foschi, J.; Turolla, A.; Antonelli, M. Soft sensor predictor of E. coli concentration based on conventional monitoring parameters for wastewater disinfection control. *Water Res.* **2021**, *191*, 116806. [CrossRef] [PubMed]
26. Wongburi, P.; Park, J.K. Prediction of Wastewater Treatment Plant Effluent Water Quality Using Recurrent Neural Network (RNN) Models. *Water* **2023**, *15*, 3325. [CrossRef]
27. Wang, Z.; Wang, Q.; Liu, Z.; Wu, T. A deep learning interpretable model for river dissolved oxygen multi-step and interval prediction based on multi-source data fusion. *J. Hydrol.* **2024**, *629*, 130637. [CrossRef]
28. Satish, N.; Anmala, J.; Rajitha, K.; Varma, M.R. A stacking ANN ensemble model of ML models for stream water quality prediction of Godavari River Basin, India. *Ecol. Inform.* **2024**, *80*, 102500. [CrossRef]

29. Ha, N.T.; Manley-Harris, M.; Pham, T.D.; Hawes, I. The use of radar and optical satellite imagery combined with advanced machine learning and metaheuristic optimization techniques to detect and quantify above ground biomass of intertidal seagrass in a New Zealand estuary. *Int. J. Remote Sens.* **2021**, *42*, 4712–4738. [CrossRef]
30. Zhang, W.; Wang, S.; Wang, Q. KSAP: An approach to bug report assignment using KNN search and heterogeneous proximity. *Inf. Softw. Technol.* **2016**, *70*, 68–84. [CrossRef]