2021 Intelligent Water Systems Challenge Solution

Data-Driven Optimization of Neutralizer pH Control in a Wastewater Treatment Plant

1. Team

i. Zhen (Jason) He, Professor, Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis. Email: zhenhe@wustl.edu (**Team leader**)
   a. Specializes in the research and development of wastewater treatment processes and technologies.
   b. Responsible for managing the project, coordinating the team, defining the problem, project progress, applied solutions and desired results.

ii. Yanran Xu, Ph.D. student, Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis. Email: xuyanran@wustl.edu
   a. Specializes in wastewater research, data analysis and machine learning modeling.
   b. Responsible for data analysis, modeling development, and result data visualization.

iii. Sandy Bernard, Process Engineer, American Bottoms Wastewater Treatment Facility. Email: sandyb@americanbottoms.com
   a. Specializes in wastewater treatment operations, online process control, and data collection systems.
   b. Responsible for collecting data, building datasets, and wastewater treatment processes control.

iv. Xuhui Zeng, Civil Engineer, Greeley and Hansen. Email: xzeng@greeley-hansen.com
   a. Specializes in machine learning and computer programming language.
   b. Responsible for machine learning guidance & support, model review, and quality control.

2. The Problem Statement

2.1 Objectives

The Physical-Chemical Wastewater Treatment Plant (PChem Plant) in the Village of Sauget, Illinois mainly provides preliminary and primary treatment for industrial wastewater. The treated water then is conveyed to American Bottoms Biological Wastewater Treatment Plant for secondary biological activated sludge treatment. The industrial wastewater entering the PChem Plant typically has a low pH of less than 3. Due to the extremely low pH of the influent wastewater, a large quantity of lime is required to neutralize pH; otherwise, it will result in corrosions of facility equipment and severe damage to the downstream biological treatment

process. The existing neutralization process includes a lime-dosing system and three neutralizing tanks in series to increase the pH. The flow rate of lime slurry delivered to each tank is controlled by the position of the lime addition valve, which is determined by a proportional-integral-derivative (PID) controller. One of the problems is that the existing PID controller fails to provide an accurate and robust control of downstream pH due to the complexity of the system thermodynamics and kinetics. Outliers of extremely low or high pH were observed, threatening the stability of downstream biological treatment. The second problem is that the PID settings are fixed with the slow response of the chemical system, failing to comply with the occurrence of various influent wastewater. This indicates a high requirement for the skills and experience of operators and results in uncertainty of the effluent pH. Consequently, this causes lime waste and thus higher costs. The first objective of our work is to predict the pH of three neutralizer tanks using well-trained machine learning (ML) models based on the current datasets of the wastewater treatment system; the results will serve as a reference for PID settings adjustment. In addition, an alert for pH control can be provided when an extreme pH is predicted. The second objective is to recommend lime addition valve position, which can serve as a good reference for manual control of valve position. For future work, a real-time supervisory system is expected to be developed. The well-trained "dose-response" ML model will serve as a Virtual Work Space to run the Digital Twin simulation. A data pipeline will be developed to automatically provide real-time data for the ML model. Real-time monitoring of the system performance will be visualized on a data dashboard. Also, recommended valve position is provided as a reference for better control of the pH. The solution shall achieve high accuracy in ML models with minimum impacts on the whole wastewater treatment plant.

2.2 The Intelligent Water System

2.2.1 Data Source

The data sets used in this challenge were provided by AB PChem Plant for a period of three years. The data sets are summarized as shown in Table 1.

Online Monitoring Data: data were collected in the influent, and three neutralizers, including 1) daily data: temperature; and 2) data at a 5-minute interval: flow rate, pH, and lime addition valve output from the supervisory control and data acquisition (SCADA) system.

Water Quality Analysis Data: routine plant monitoring daily data in the influent, and three neutralizers, including biological oxygen demand (BOD), total suspended solids (TSS), and total Kjeldahl nitrogen (TKN).

Table 1. Summary of data sets

Parameter	Unit	Time interval
Temperature	℃	1 day

Influent	PC Influent Flow	MGD	5 minutes
	PC Influent pH	/	5 minutes
	Influent BOD	mg L^{-1}	1 day
	Influent TSS	mg L^{-1}	1 day
	Influent TKN	mg L^{-1}	1 day
Neutralizer	Neutralizer #1 pH	/	5 minutes
	Neutralizer #2 pH	/	5 minutes
	Neutralizer #3 pH	/	5 minutes
	Neutralizer TSS	mg L^{-1}	1 day
	Neutralizer #1 lime valve	% closed	5 minutes
	Neutralizer #2 lime valve	% closed	5 minutes
	Neutralizer #3 lime valve	% closed	5 minutes
Effluent	PC Effluent Flow	MGD	5 minutes
	Effluent pH	/	5 minutes
	Effluent BOD	mg L^{-1}	1 day
	Effluent TSS	mg L^{-1}	1 day
	Effluent TKN	mg L^{-1}	1 day

2.2.2 Lime Control System

The lime slurry is used for the neutralization of wastewater pH. It is recirculated in the lime system after slaking the ½ inch to ¾ inch pebble lime in the lime slake tank and stored in the lime storage tank. The programmable logic controller named P-CHEM PLC was initially installed as an industrial digital computer for data monitoring points and control of the lime system. The lime system consists of lime pumps, lime addition valves, and neutralizer gates. The pumps are used to circulate lime slurry at a specified flow rate. Three automatically controlled valves are used to control pH to a setpoint in each of the three neutralization chambers. The valve positions are controlled by the PLC and shown in % closed. The valves can also be manually controlled. When the valve is in the Auto mode, the PID controller calculates a position signal of lime addition valves based on the pH setpoint, the current pH, and the PID settings. Two neutralizer slide gates separate the three neutralizer chambers. These gates are opened automatically when the plant goes into a rain mode to permit higher flows. Each neutralizer chamber has an agitator. The lime dosing can be sufficient to control wastewater pH, but the lime dosing amount is based on a conservative setpoint, which could cause lime overdosing. One of the outcomes of this work is to develop a lime dosing strategy to minimize overdosing and thus to save the operational costs.

2.3 Work Design

2.3.1 Work Plan

We use online monitoring data and water quality analysis data to predict neutralizer pH. The lime addition valve position can be recommended with well-trained ML models, which will serve as a reference parameter for the PID control. The well-trained model, as a "dose-response" ML model, is built to simulate the pH response concerning different wastewater characteristics data and lime dosing amounts. This simulation dataset will serve as a Virtual Work Space to run the Digital Twin simulation. A wide range of valve position values from 0 to 100 are applied to the ML models, resulting in a set of predicted pH values produced. The recommend lime addition valve position is obtained when the predicted pH is equal to the target pH. Finally, a real-time supervisory system is developed where a well-trained ML model is applied and a data pipeline is built to feed data into the model. And a data dashboard for real-time data visualization and an alarming system for abnormal influent occurrence are included.

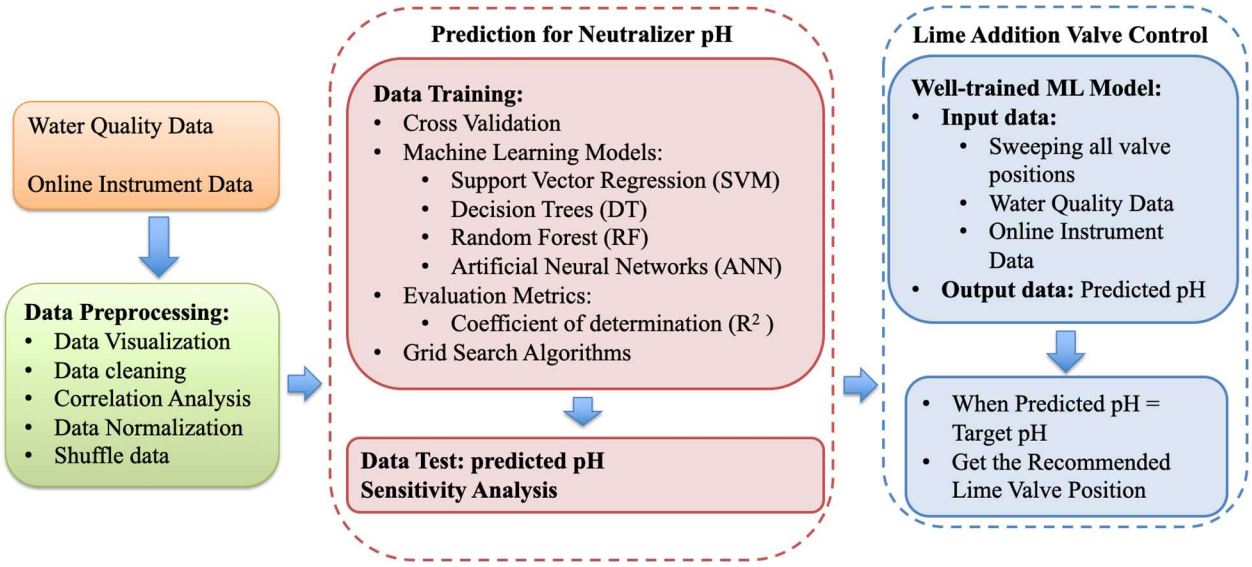


Figure 1. Work plan for the final solution

2.3.2 Method

i. Correlation analysis

The correlation analysis was used to identify the relevant attributes of the features to the output feature in the dataset. The correlation matrix was generated showing the positive or negative correlations between two features. The input features were decided due to their obvious attributes to the output feature.

ii. Data normalization

Since values of input feature varied in different ranges, normalization was applied to change the values to a common scale without distorting differences in the value distribution. Several scaling

methods were compared and applied on input data of ML models, including Min_max scaler, Standard scaler, and Normalizer. In general, standard scalers performed best among the three methods.

iii. Machine learning models

The commonly used ML models, such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Ada Boost (AB), Artificial Neural Network (ANN) and XGBoost were chosen for regression modeling.

iv. Cross-validation and hyperparameter tuning

For all ML models, 80% of data was selected for training while the remaining 20% of data was for testing. Cross-Validation was applied to test the ability of the model to predict new data. It was used to avoid problems like overfitting or selection bias. And the Grid Search was used to tune hyperparameters and choose the best-performed ML models.

v. Sensitivity analysis

The sensitivity analysis is considered a powerful way to understand an ML model. It examined the impact each feature has on the model's prediction. The impacts of each feature on neutralizer pH were illustrated with this analysis.

3. Solutions

3.1 Benchmarking the Existing System Performance

The performance of the existing system was evaluated based on the frequency and severity of overdosing as well as the condition under which overdosing would happen. These overdosing events can be vividly illustrated through data visualization.

The influent pH fluctuates in a wide range between 0 and 13, with most remained at low pH values between 1 and 2.5 (Fig. 2a). The influent pH was required to be accurately adjusted, and expected to reach a target pH of 8 after the neutralization process. The neutralization was designed to increase the pH stepwise with setpoints of 5.6, 6.8 and 8.2 among three neutralizers. Here, the results of pH control were illustrated through data visualization (Fig. 2b, 2c & 2d). The occurrence of real pH values was counted, and the dashed lines illustrated the target pH. The extreme pH conditions should be avoided as much as possible, or it will disrupt the downstream biological process. In addition, the varying alkalinity condition will increase the difficulty to O&M. More improvements can be applied to avoid outliers of extreme pH and control the process more accurately. For example, pH outliers at low pH of 2.5 and high pH of 10 can be further eliminated in neutralizer 1 (Fig. 2b). It can be avoided that the pH distribution of neutralizer #2 was divided into two peaks in Fig. 1c. And pH outliers at 10 and 11 can be prevented in neutralizer #3 (Fig. 2d). The control alerts will be provided if the pH is predicted with ML models, and these outliers can be prevented with in-time lime dosing adjustment.

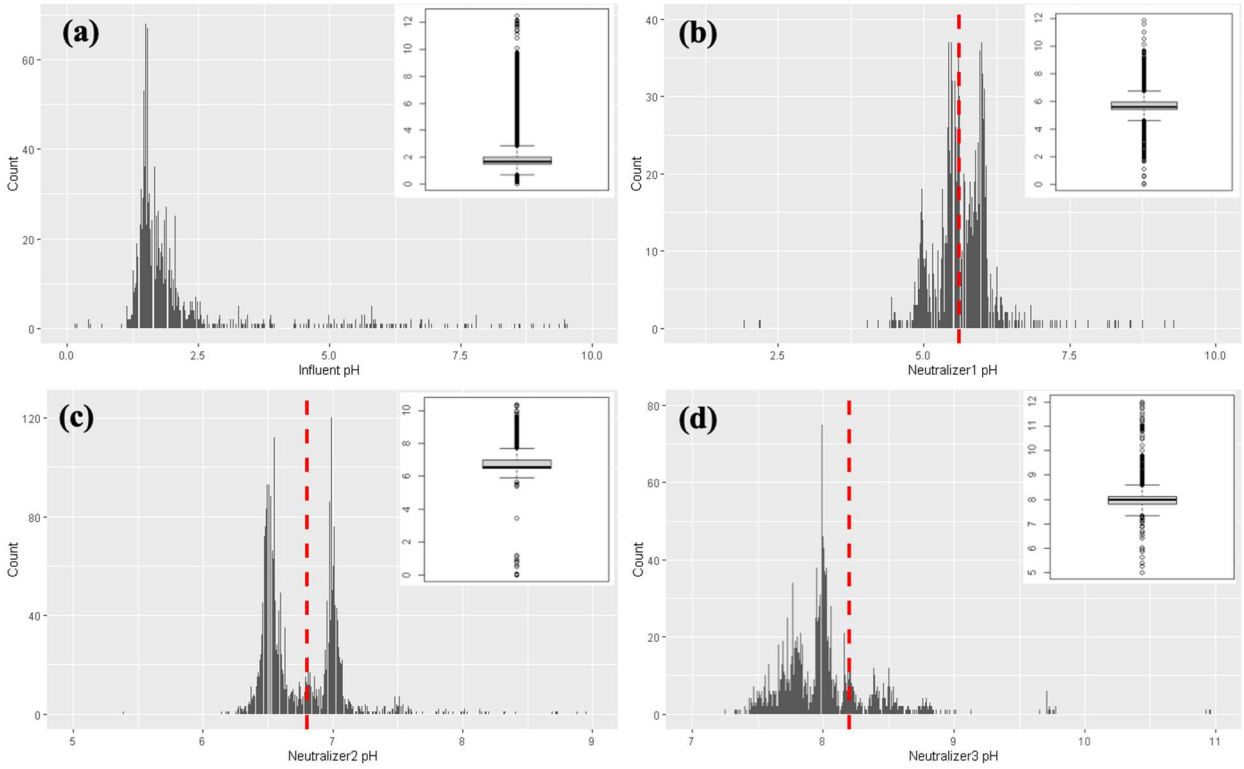


Figure 2. The pH performance of the existing neutralization process. The boxplots show the maximum, minimum, median, first quartile and third quantile values for influent and three neutralizer pH.

3.2 Data Preprocessing

The raw datasets were collected and preprocessed as follows:

3.2.1 Data Cleaning

The raw data were provided with time intervals of one day or five minutes (Table 1). So, all data were unified into two intervals as two separate datasets, i.e., one dataset at a daily interval and the other at a five-minute interval: 1) Daily dataset: the average values were used as daily data; 2) Five-minute interval dataset: the daily data were replicated to align time interval of 5 minutes. Besides, the observations with missing values were moved out.

3.2.2 Correlation Analysis

The correlation analysis was used to evaluate the strength of the relationship between two variables (or features in ML models). The values of correlation coefficients are between -1 and 1, showing a strong positive correlation with 1, a strong negative correlation with -1, and no correlation with 0. For example, as shown in Fig. 3, the neutralizer #1 pH may be correlated with valve #1 position, influent pH, influent TKN and temperature. And these features were chosen as inputs to predict neutralizer #1 pH in ML models. Appendix A shows the correlation matrix for all features, which served as a reference for choosing input features in ML models.

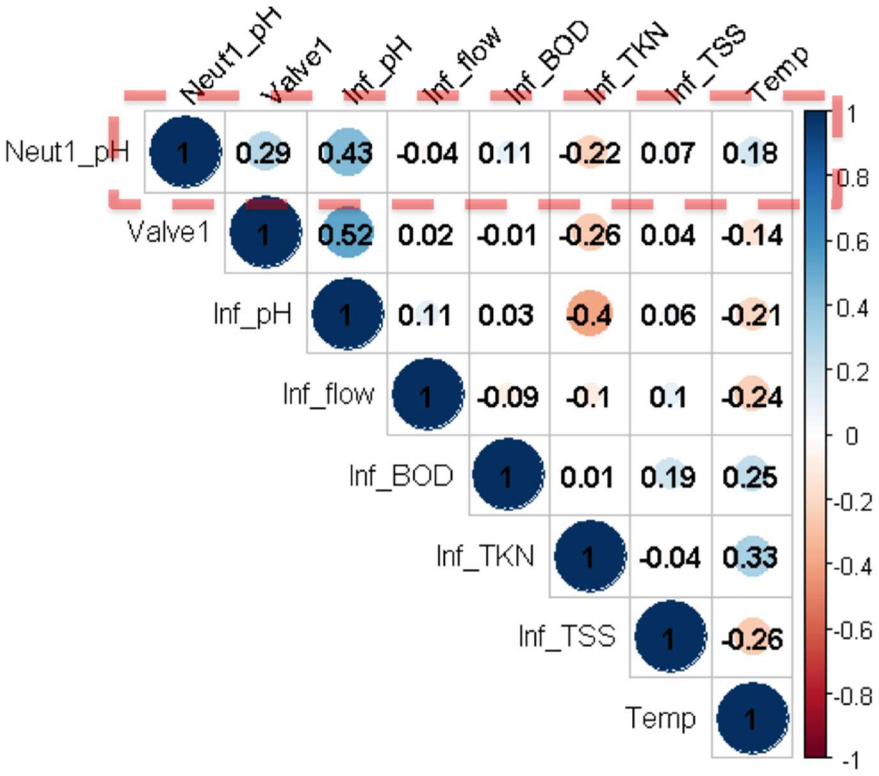


Figure 3. Correlation matrix among features for neutralizer #1. The color and size of circles, and values indicate the correlation between two features.

3.3 Module One: pH Prediction in Neutralizer #1 & #2 & #3

The existing lime control system mainly relies on PID controller to determine lime dosing valve position. However, some outliers of extreme pH values were still found. This issue can be addressed by extra operations when pH is predicted, and an outlier alert is provided.

3.3.1 Data, Modeling, and QA/QC

In this module, online monitoring data (pH, flow rate, and valve position) and water quality analysis data (temperature, BOD, TKN, and TSS) were used. The results of the five-minute interval dataset outperformed the daily dataset. So, all data were unified with a time interval of five minutes. Because the water quality analysis data were collected daily, all daily data were replicated to align with 5-minute data. The lime dosing valve position of three neutralizers was started to be recorded on march 30th 2021, and collected on June 7th 2021. After data cleaning and removing missing data, 14,395 data were used in ML models.

Because the range of pH was narrow, the regression models were conducted to predict pH. The ML models including SVM, DT, RF and ANN were used and compared.

Based on the correlation matrix shown in Appendix A, correlations between neutralizer pH and other features were illustrated, which provided a good reference to choose input features in ML models (as shown in Table 2). For neutralizer #1 pH, obvious positive correlations with influent

pH and lime addition valve #1 were found. Besides, neutralizer #2 pH was positively correlated to influent pH, neutralizer #1 pH, lime addition valve #2 and temperature. Finally, neutralizer #3 pH was mainly related to neutralizer #2 pH and no obvious correlation was found with lime addition valve #3. It was indicated that lime addition made a difference in the first two neutralizers, but affected pH less in the last neutralizer. All input features were normalized since the input features fell in various ranges. All data were shuffled and split into training and testing sets.

Table 2. Input features for the prediction of three neutralizer pH

	Input features
Neutralizer #1 pH	Influent pH, lime addition valve #1 position, influent flow, influent BOD, influent TKN, influent TSS, and temperature.
Neutralizer #2 pH	Influent pH, neutralizer #1 pH, lime addition valve #2, temperature, influent flow, lime addition valve #1, influent BOD, influent TSS, and influent TKN.
Neutralizer #3 pH	Neutralizer #2 pH, Influent pH, neutralizer #1 pH, lime addition valve #2, lime addition valve #3, temperature, influent flow, lime addition valve #1, influent BOD, influent TSS, and influent TKN

3.3.2 The pH Prediction of Neutralizers

The 80% of data were first used to training ML models. The data training was implemented with Cross-Validation and selected with the Grid Search technique. The training results were evaluated based on R^2 score, and more details were shown in Appendix B. The rest 20% of data were applied for model testing, and the results were concluded in Table 3. The models were proved to predict pH with R^2 score of at least 0.7. The RF model performed best, and predict pH of three neutralizers with 0.718, 0.714 and 0.895 in R^2. The performance of the RF model for neutralizer #1 pH was visualized as shown in Fig. 4, where the predicted and actual pH values were compared. The performance of RF models for the other two neutralizers was introduced in Appendix C. And it was illustrated that the outliers caused challenges for ML model prediction.

Table 3. Machine learning regression model testing results for Module one: pH prediction of three neutralizers

R^2 score	SVM	DT	RF	ANN
Neutralizer #1 pH	0.637	0.619	0.718	0.703
Neutralizer #2 pH	0.616	0.634	0.714	0.689
Neutralizer #3 pH	0.672	0.812	0.895	0.859

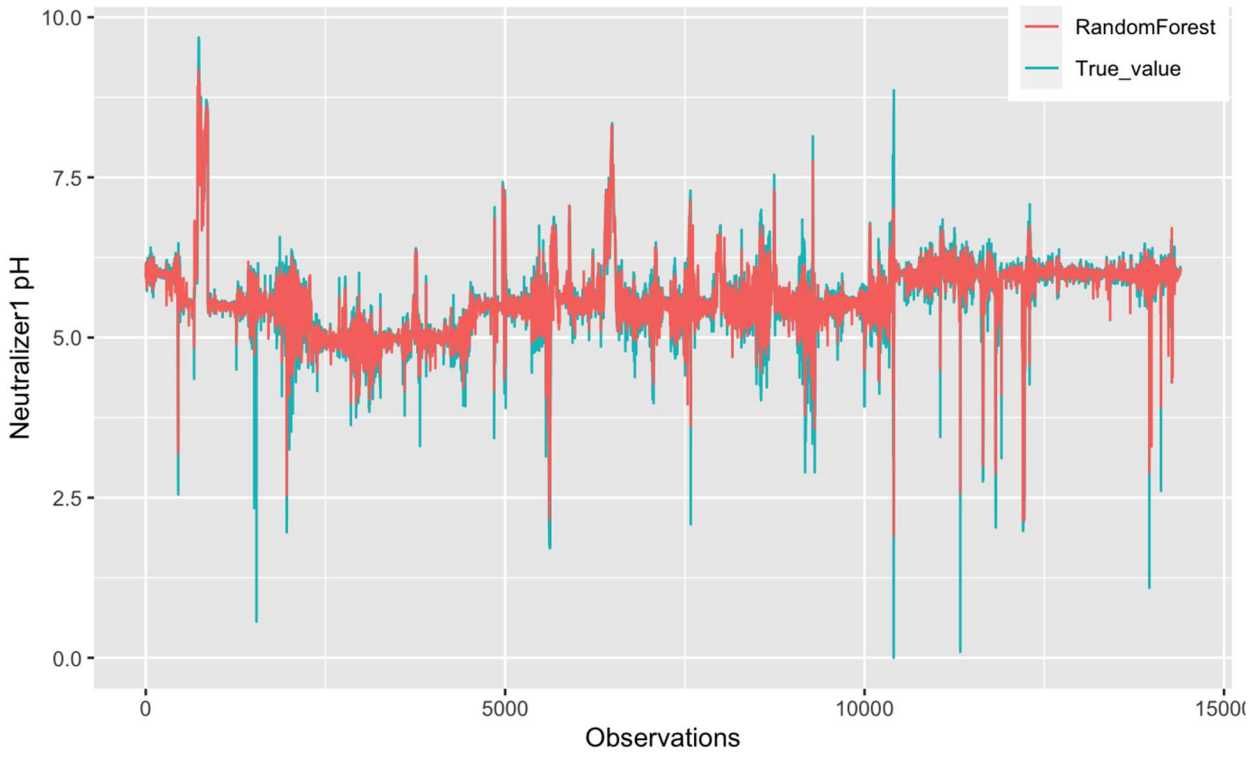


Figure 4. Predicted (red line) and actual (green line) pH of neutralizer #1 with the Random Forest regression model.

3.3.3 The Sensitivity Analysis

Based on the RF models, the impacts of other features on neutralizer pH were examined through sensitivity analysis. To achieve target pH, the impacting features need to be precisely controlled. In the neutralization process, the pH was mainly adjusted by lime addition, whose dosage was determined by the valve position. So, the sensitivity analysis was conducted to check the impact of lime addition valve position on pH. As shown in Fig. 5, the impacts of features on neutralizer #1 pH were ranked. The temperature showed the largest impact but it was not manually adjustable. The lime addition valve #1 ranked second, confirming its feasibility and effectiveness in pH adjustment. So, in Module two, recommended valve positions were calculated with ML models to optimize pH control, which could serve as a good reference for the lime system control. The sensitivity analysis for the other two neutralizers was mentioned in Appendix D. For neutralizer #2 pH, it was mainly impacted by neutralizer #1 pH and temperature, which were followed by lime addition valve #2 position. However, neutralizer #3 pH was slightly affected by lime addition valve #3 (ranked 5), but mainly by influent BOD. This indicated a challenge for the following valve position prediction based on pH in neutralizer #3.

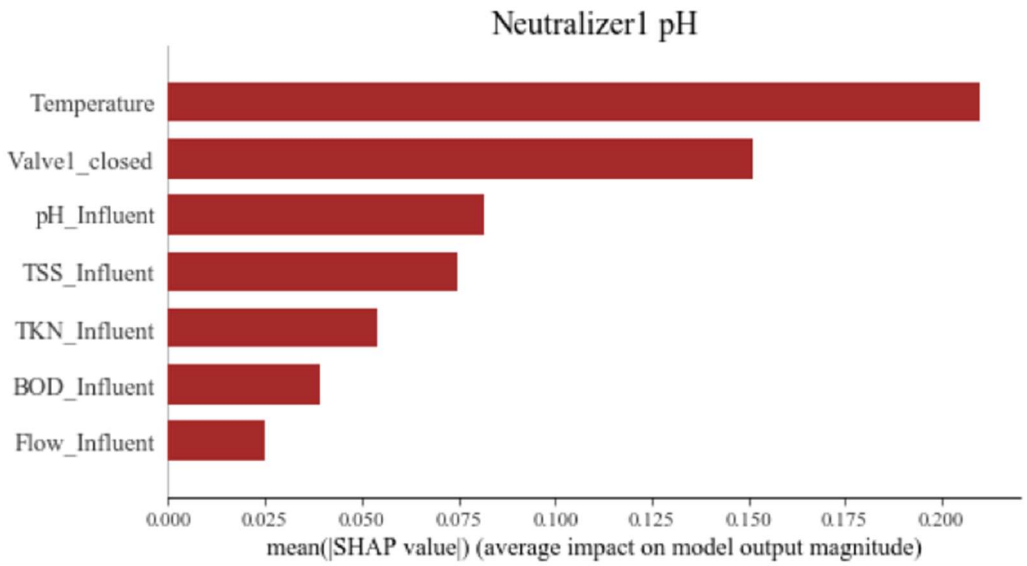


Figure 5. Sensitivity analysis of neutralizer #1 pH with the Random Forest regression model

3.4 Module two: recommended lime addition valve position

The lime dosing is represented by the lime addition valve position. The impacts of valve position on neutralizer pH were confirmed in Module one by sensitivity analysis. Besides, the existing lime control system is operated with PID control and is usually in auto mode. However, sometimes manual mode is used when extreme pH is observed, which shows a high requirement for skilled and experienced operators. So, the recommended valve position for lime dosing can also serve as a reference to optimize pH control.

3.4.1 Data, modeling and QA/QC

Both online monitoring data and water quality analysis data were used as input features. The water quality analysis data were aligned to the time interval of five minutes by duplicating. The well-trained RF regression models (from Module one) were used to predict the neutralizer pH. The difference compared to Module one was that the input valve position swept from 0 to 100. And the corresponding predicted pH values were collected, and compared with the target pH. The recommended valve position was chosen when the predicted pH was closest to the target pH.

Based on the correlation matrix shown in Appendix A, correlations between valve position and other features were demonstrated. The lime addition valve #1 was positively correlated with neutralizer #1 pH. For lime addition valve #2, obvious positive correlations were observed with pH of neutralizer #2. However, lime addition valve #3 pH was mainly related to neutralizer #2 pH and no obvious correlation was found with neutralizer #3 pH, indicating a challenge to adjust neutralizer #3 pH through lime addition valve #3.

3.4.2. The Recommendation of lime addition valve position

For example, different values of lime addition valve #1 (from 0 to 100) were input to the well-performed RF model, and the neutralizer #1 pH was predicted as shown in Fig. 6. According to the target pH of 5.6 for neutralizer #1, the recommended valve #1 position was in the range between 45 and 70.

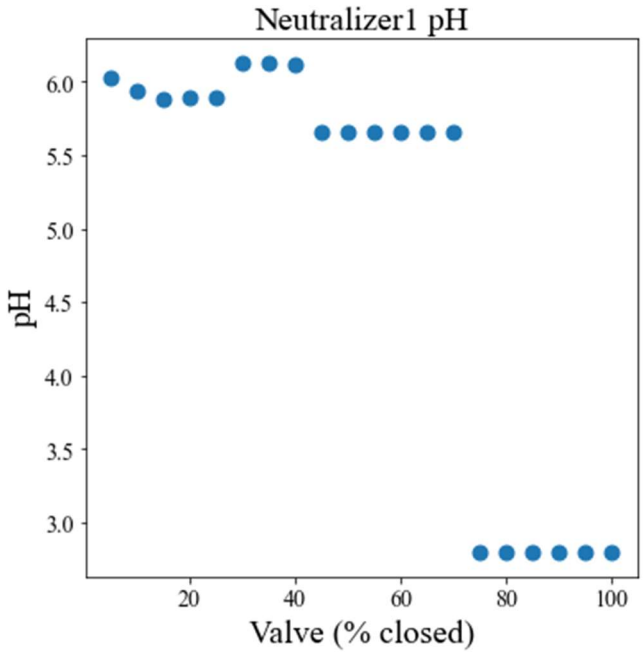


Figure 6. The recommended lime addition valve #1 position for neutralizer #1 pH

3.5 Future work: development of a real-time supervisory system

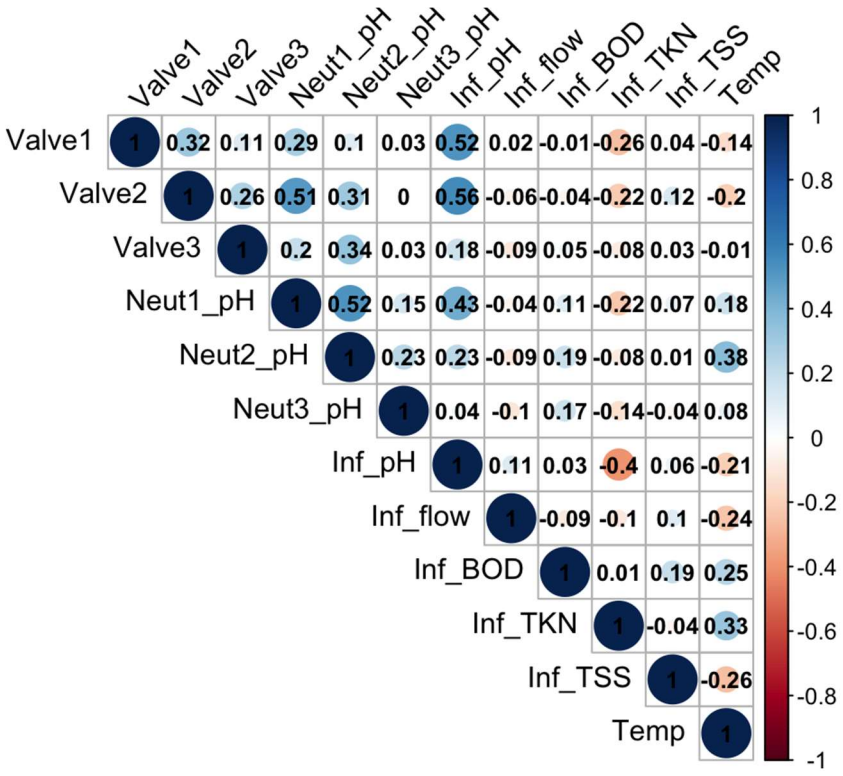
A real-time supervisory system is developed where well-trained "dose-response" ML models are included, and a data pipeline is built to feed real-time data into the model. A data dashboard is built for real-time data visualization and an alarming system for abnormal influent occurrence. Also, recommended lime addition valve position can be provided through ML models, serving as a reference for pH control.

4. Conclusions

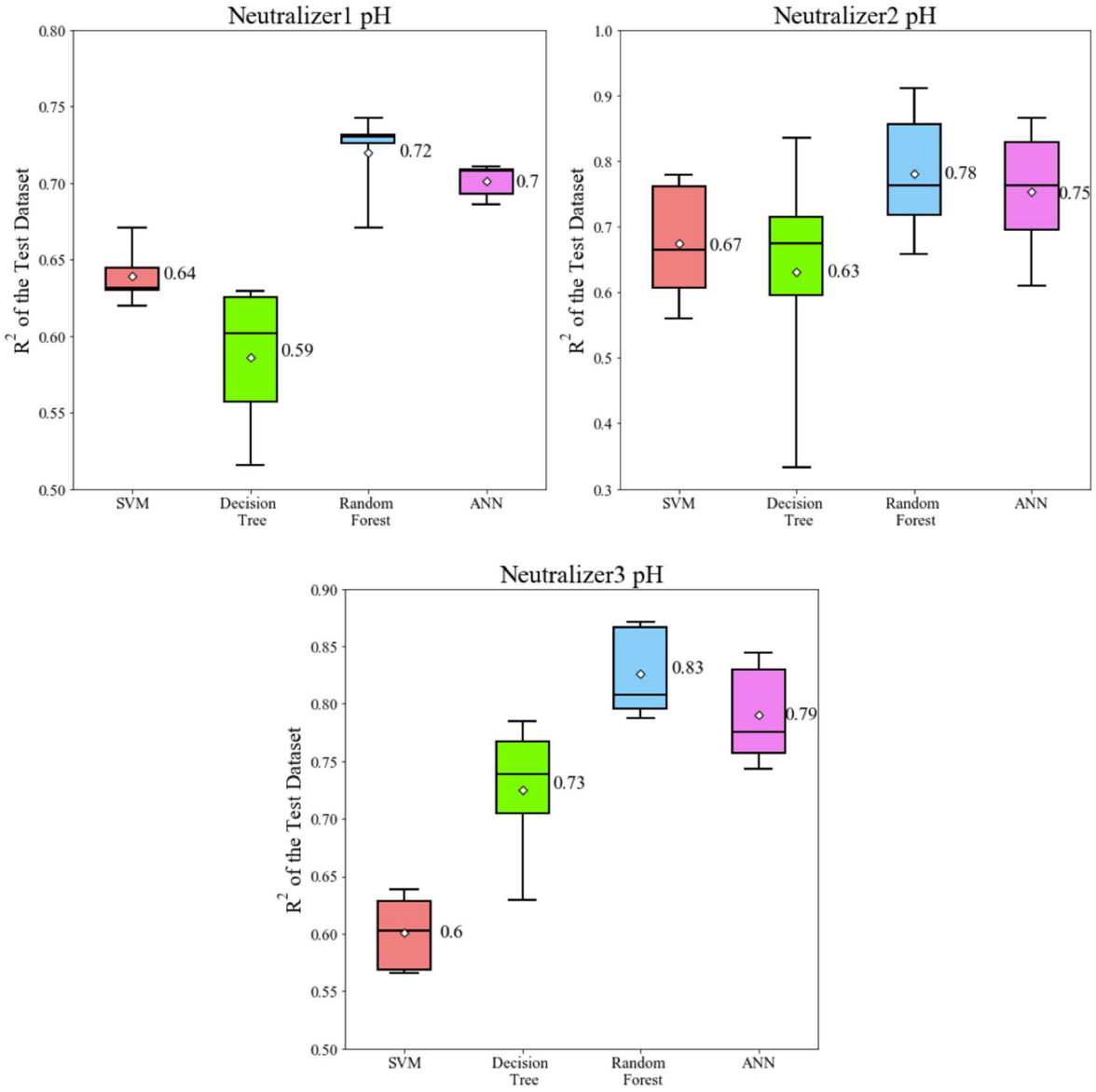
This work aimed to apply data-driven methods to solve real-life problems in wastewater treatment plants and optimize the facility performance with big data calculation. It is a good example and case study for the application of machine learning models in wastewater plant management. The machine learning models can assist to prevent the extreme pH conditions in PChem Plant, which will benefit biological treatment downstream and decrease the O&M related cost. The well-trained machine learning models are capable of predicting pH in each neutralizer, and address the drawbacks of PID control for lime dosing system. The recommend lime dosing

can be provided by choosing the lime addition valve position to reach the target pH. This work will serve as a good reference for PID control of lime dosing. The performance of the neutralization process will be further optimized. We are confident that those data-driven methods can be employed to address pH and chemical dosage problems in other wastewater treatment facilities.

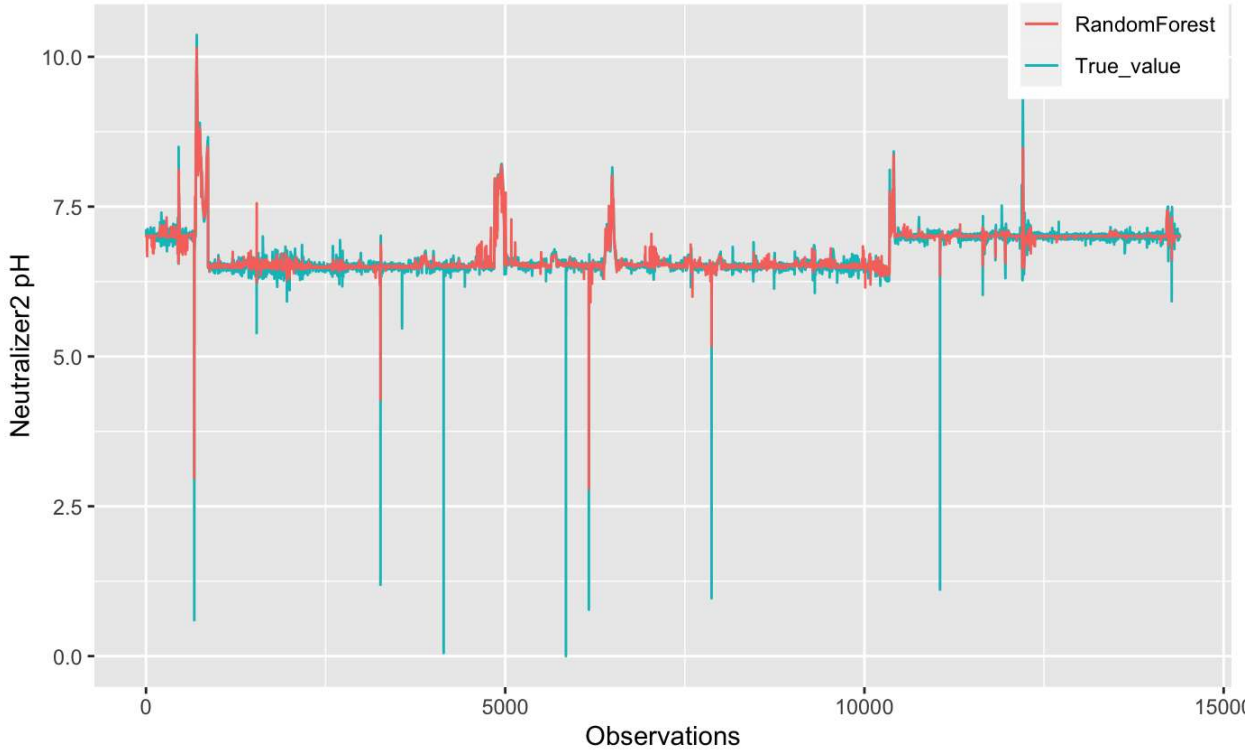
Appendix A – Correlation matrix for all features

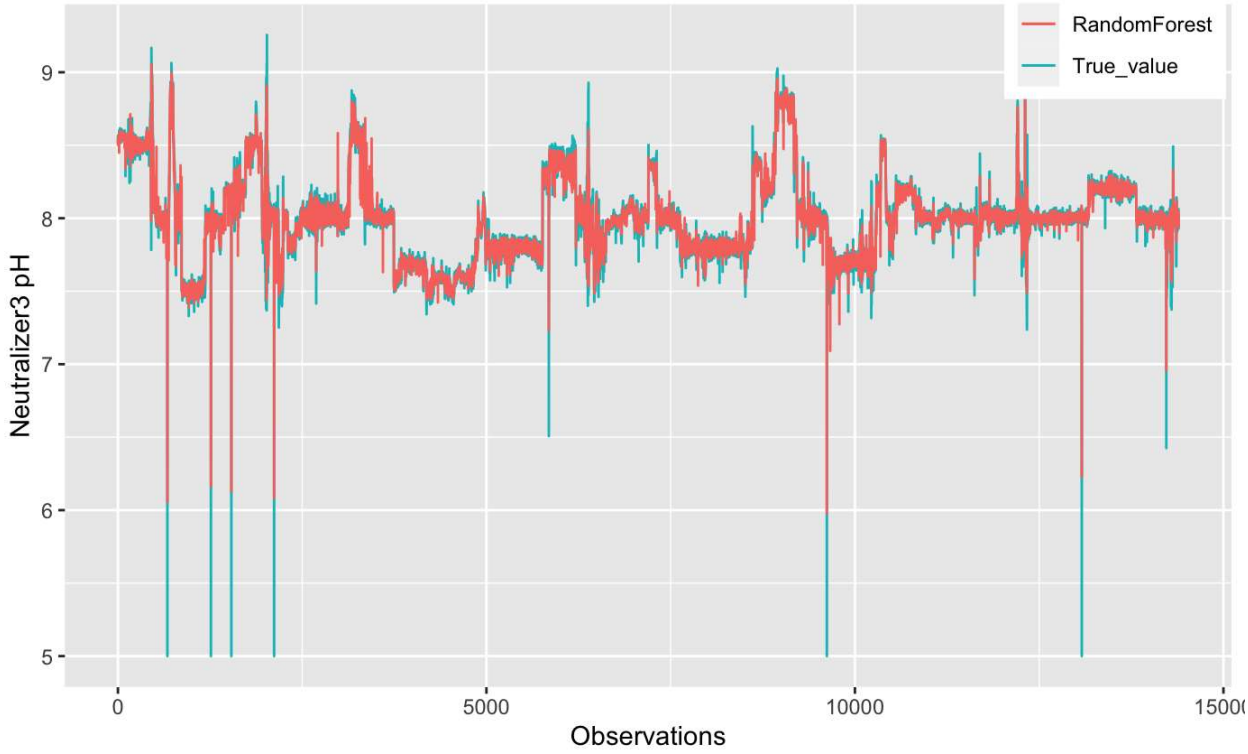


Appendix B – ML model training results for Module One



Appendix C – ML model prediction for neutralizer 2 & 3 pH with the Random Forest model





Appendix D – Sensitivity analysis of neutralizer 2 & 3 pH with the Random Forest model