Forecasting the Ammonia Concentration and Color Level in Reclaimed Water using Machine Learning

by

Ting Hsi LEE

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in the Department of Civil and Environmental Engineering

August 2022, Hong Kong

Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Ting Hsi LEE

August 2022

Forecasting the Ammonia Concentration and Color Level in Reclaimed Water using Machine Learning

by

Ting Hsi LEE

This is to certify that I have examined the above MPhil thesis
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by
the thesis examination committee have been made.

Prof. Chii SHANG, Thesis Supervisor

Prof. Meimei Han, Head of Department

Department of Civil and Environmental Engineering
August 2022

Acknowledgments

First of all, I am truly grateful for being one of the first PhD students supervised by Prof. Li. He was full of passion and patience when helping me build the know-how for this degree. It has been a great pleasure for me to be part of this team and grow together with the lab during the last four years. Furthermore, I would like to thank all of the members of the thesis examination committee for their careful examination of my thesis.

Finally, I would not stand at this current point without their endless love and unconditional support for all these years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Forecasting the Ammonia Concentration and Color Level in Reclaimed Water using Machine Learning

by Ting Hsi LEE

Department of Civil and Environmental Engineering

The Hong Kong University of Science and Technology

Abstract

Water scarcity is a global challenge. One of the promising ways to mitigate the water resource crisis is via wastewater reclamation. Chlorine is commonly used for reclaimed water disinfection and requires precise dosing to satisfy endorsed quality standards. Ammoniacal nitrogen (NH_3N) and colour exist in the reclaimed water at concentrations between 0.23 – 5.44 mg N/L and 80 – 150 Hazen units, respectively, and can affect the chlorine demand. Forecasting the reclaimed water quality enables a feedback control system over the disinfection process by predicting the exact chlorine dose required which secures sufficient time to respond to sudden surges in color and ammonia levels. This study developed time-variant models based on machine learning to predict the NH_3N concentration and colour three hours into the future in the reclaimed water. The NH_3N data was collected by an online analyzer, and colour data was collected by a customized auto-sampling spectrophotometer, both are installed in the reclaimed water treatment plant in Hong Kong. Long Short-Term Memory (LSTM) was found to be the most effective architecture for training NH_3N and colour forecasting models. In the training processes, we applied data pre-processing methods and feature engineering, a technique to select or create relevant variables in raw data to enhance predictive model performance. From feature engineering, we discovered that the daily fluctuation in NH_3N and colour has correlations with the urban water consumption patterns. This finding further enhanced the NH_3N and colour forecasting model performance by 4.9% and 5.4% compared to baseline models. This research work offers novel methods and feature engineering pro-

cesses for NH_3N concentration and colour forecasting in reclaimed water for treatment optimization.

CHAPTER 1

INTRODUCTION

1.1 Background

AI technologies have been successfully applied to different DWT processes, such as the prediction of the coagulant dosage, discrimination of the DBP formation potential, advanced control of membrane fouling, membrane preparation and optimization, and water quality prediction. Li et al. (2021)

Forecasting models play an important roles in water quality control in drinking water treatment plants (DTPs) and wastewater treatment plants (WWTPs). The need of using forecasting models are becuase the unpredictable nature of water quality, and the treatment operations are subjected to the change of water quality to prodcue effluent complied the government regulation Chen et al. (2003)

Forecasting models can also be called time series model becuase the data is consisted of the values and the time (need to be further revised). For the well-know time series models are for example, RNN, ... These are used to replace the theory-based models, for example Activated Sludege Model (ASM). The difference between these two models are, machine learning based models require to learn from historic data, while the thoery-based models only need to enter the basic operational parameters (e.g., influent flow, tempearture, and pH, etc).

Despite the promising usage and performance of machine learning models, the collection of the data became the most difficult tasks. Many small scale or old treatment plants do not have the capital or the available environment for the set-ups of the online sensors to collect data. Although these are the major issues, it's still possible to train a forecasting model with one input, which is also called a self-prediction model. Although the accuracy or stability compared to multi-input models, the forecasted results can be used at some cases. To increase the model performance, there are several ways. Paper included weather data, or perform data-preprocessing methods to improve the model performance.

These solutions (data preprocessing, feature engineering) are not well discussed in this field, also the potential of using univariate models are under estimated.

1.2 Objectives

The specific objectives of this thesis work are:
(1) To build baseline univariate forecasting models using machine learning and deep learning models.
(2) To develop data preprocessing methods for enhancing model forecasting performance.
(3) To extract features and hidden relations of water parameters in MBR effluent by analyzing the wastewater collected upstream of the WWTPs.
(4) To develop methods for improving performance of forecasting models using the hidden features and relations of the water parameters.

1.3 Organization of the thesis

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to water quality control

2.1.1 Automated system for water quality control

Programmable logic controller (PLC) is an industrial computer system designed for any process requiring a series of devices and equipment operates cohesively to achieve multiple purposes in manufacturing or treatment processes. The main components of PLC include a center process unit (CPU), input modules and output modules (I/O). CPU is responsible to process digital signals from input modules and send commands through output modules based on the control logics programmed on the PLC. For chemical dosing control in water treatment plants (WTPs), PLC system receives readings from turbidity and pH sensors and uses pumps to dose aluminum solution automatically (Andhare and Palkar, 2014). The PLC system with the capability of producing real-time output commands in response to the input sigals also makes it widely used in the wastewater treatment plants (WWTPs). For oxygen concentration control in the aeration tank, PLC system receives signals of dissovled oxygen (DO) detectors and transmits signals to open or close the electric butterfly valves to further alter the DO concentration (Zhu and Qiu, 2017). Although PLC systems are the most used system across industries for its easy programming and reliable control, PLC system is merely a device that can be programmed to control operative devices with on-off logic (i.e., a logic control with two states) and the capability of complex control is compromised. In reality, many WTPs or WWTPs have the need of precise control of the treatment processes. Being awared of the limitations of the PLC systems, a more advanced controller called proportional–interal–derivative (PID) controller for receiving analog signals was developed to obtain more sophisticated controls over the operative devices.

To react to rapidly-changing process conditions, a PID controller generates an output value based on continuous calculation of an error value e(t) as the difference between a desired setpoint (SP) and a measured process variable and applies a correction based on

proportional, integral, and derivative terms. The use of the "P", "I", and "D" allows the system to quickly reach steady state with a feedback control system (i.e., the system output is returned to the system input which is included in the decision making process in PID controller). Generally speaking, a PID controller is a technology (i.e., a specialist algorithm) for contorlling a single device with more complex logics, while a PLC system is a physical system consists of different modules and capable of controlling dozens of devices only with two-state logic. In addition, A PID controller can be designed to operate on PLC device and provide a more precise control strategy to a designated device. In WWTPs, a single-variable feedback analog control loop in PID can be used to control the temperature in the activated sludge treatment by stablizing the system temperature in a shorter time (Bados and Morejon, 2020). The feedback control scheme is also applied in WTPs to adjust the addition of chlorine dosage (i.e., also known as the disinfection process, chlorination, or postchlorination) to reach the target concentration of free chlorine residual (FRC) (Wang and Xiang, 2019). Disinfection process is carried out in a chlorine contact tank which provides sufficient time for chlorine to disinfect pollutants. Since the chlorine added by the dosing device requires time to travel from the entry to the exit, the system output can only reflect the changes of water quality in a delayed time of 30 minutes (i.e., the designed time for water to travel in chlorine contact tank is usually 30 minutes or longer). In the case of chlorination, the lag of time makes feedback control difficult (Kobylinski et al., 2006) as the system is delayed in responding to any sudden surge of the pollutants when it can only receive output at the end of the disinfection process. PID controllers in WWTPs also encounter similar challenges as the increasing complexity of water quality and stricter regulations on the discharged water quality.

To tackle the difficulties encountered in process control system, many control strategies are proposed, such as feed forward-feedback control, linearized and optimal control, model-predictive control, and fuzzy control, etc (Demir and Woo, 2014). Among the algorithms used in control strategies, Artificial Intelligence (AI) modeling has gained the most attentions in recent years compared to modeling based on mathematical models or empirical formulas. In WTPs or WWTPs, to fully understand the physical, biological, and chemical interactions in the treatment plants is very difficult. The unpredictable behaviors during the water treatment can be the significant changes of influent flow rate, flucutations of water quality, the complexity of biological treatment process, and the large time delay exists between this control variable and the process input, etc. Therefore, AI

modeling shows a great potential in dealing with the highly complex conditions in the treatment process (Li et al., 2021). In the next sections, the applications of different AI modeling methods will be discussed.

2.1.2 Artificial Intelligence

Artificial intelligence (AI) can perform cognitive tasks with the development of computational solutions. The concepts of AI are usually confused, in fact, AI is a very broad term and any kind of algorithms or models which involved in decision-making with computation fall in the domain of AI. For example, fuzzy logic and optimization algorithm are formulated with human design and computer decision making process. There are another subset of AI called machine learning (ML), but the process of generating a ML model is different to generating a fuzzy logic model. ML uses learning algorithms to generate a model via learning from historical or large amount of data without being explicitly programmed. ML algorithms can be classified into three categories, which are Supervised, Unsupervised, and Reinforcement learning. In the training process of supervised learning, input variable (x) and output variable(Y) we will provided, and model will learn from the provided dataset to map the x to the Y. A trained supervised model can generate a prediction for the response to the new data (i.e., also called the unseen data). Unsupervised learning is when the dataset is not labelled, the model can learn to infer patterns in the dataset without reference to the known outputs. This type of algorithm can find similarities and differences in the data. In reinforcement learning, models are designed to constantly interact with the environment in a try-and-error way and recieved rewards and punishments based on the purpose of the tasks. Generating a optimal action to achieve lowest penalties is the main function of a reinforcement learning model. In process control, supervised learning are frequently used in many senarios.

Regression is a supervised machine learning technique used to predict continuous values. A regression model can estimate the relationship between the input variables in the system and the output target from a given dataset, and then use the nonlinear relationship to map the unseen input data to a predicted output data. This type of application is sutiable for water quality prediction (Librantz et al., 2018), and sensor fault detection (Cecconi and Rosso, 2021), etc.

Fuzzy logic (FL) control is still an effective strategy for process control, and this type

of AI modeling is called reasoning. Fuzzy logic is described as an interpretative system in which objects or elements are related with borders not clearly defined, granting them a relative membership degree and not strict, as is customary in traditional logic. The typical architecture of a fuzzy controller, shown in Figure 3, consists of a fuzzifier, a fuzzy rule base, an inference engine, and a defuzzifier Santín et al. (2015) proposed a hybrid control system comprised of FL controller and model predictive control using optimizaion model to control the chlorine dosing in a WTP. FL controller and optimzation model fall in the domain of AI, which is excluded from the subset of ML.

Fuzzy logic (FL), a method based on multi-valued logic, uses fuzzy sets to study fuzzy judgement, which allows FL-based fuzzy inference systems to simulate the human brain to implement natural inference [40].The adaptive fuzzy neural inference system (ANFIS) composed of FL and ANN with an inference mechanism has high interpretability compared to common ANN. The combined model has been used to control coagulant dosing systems [41,42].

2.1.3 Machine learning and deep learning

In machine learning, popular models which are frequently used by the researchers for training predictive models are Supporting Vector Machine (SVM), Random Forest (RF), and Artificial Neural Networks (ANN). Librantz et al. (2018) trained a RF model to predict the free residual chlorine concentration (FRC) in a WTP, and Xu et al. (2021) built a RF-based model to predict total nitrogen concentration in water bodies. Guo et al. (2015) compared the reliability and accuracy of an ANN model and a SVM model in predicting 1-day interval T-N concentration in a WWTP, and the results showed that RF model has higher accuracy while ANN model is more reliable for assisting decision-making process.

As the the computing power doubled every 18 months according to Moore's law. A subset of ML, Deep Learning (DL) becomes more accessible for sovling everyday issues. In simplicity, DL models can be defined as neural networks with more than two hidden layers (i.e., the model complexity increased and required more computing power to calculate). In DL, there are various types of architectures designed based on the type of problems. For image processing, Convolutional Neural Network (CNN) is designed to extract important features from the image vectors. Another popular DL architecture is Recurrent Neural

Network (RNN), which is powerful in solving time series-related applications and Natural Language Processing (NLP) tasks (Li et al., 2018). Although each architecture has their strengh in tackling different types of problems, both architectures can be used for a single task Li et al. (2022) built a regression CNN-RNN model for rainfall-runoff prediction. DL can be extremely powerful when multiple architectures ared fused into a single model to perform a specific task, which cannot be realized by machine leraning models. That being said, DL can achieve higher model performance in terms of the prediction accuracy compared to ML.

2.2 Water quality control with machine learning

2.2.1 Drinking water treatment plants

A drinking water treatment plant (DWTPs) produces potable (i.e., drinking water) water for human consumptions by removing contaminants from the source water, such as lake or stream, or from an underground aquifer. The raw water enters DWTPs and goes through treatmet units of coagulation, flocculation, sedimentation, filtration, and disinfection in sequence as the primary treatment scheme in the conventional DWTPs (Li et al., 2021).

1.8.1 Drinking Water Treatment Drinking water treatment plant could be classified into: – Disinfection plant which is used for high-quality water source to ensure that water does not contain pathogens – Filtration plant: this is usually used to treat surface water – Softening plant which is used to treat groundwater Typical filtration plant is shown in Fig. 5 which is designed to remove odors, color, and turbidity as well as bacteria and other contaminants. Filtration plant employs the following steps: a.Rapid mixing : where chemicals are added and rapidly dispersed through the water b.Flocculation : Chemicals like alum (aluminum sulfate) are added to the water both to neutralize the particles electrically and to make them come close to each other and form large particles called flocs that could more readily be settled out c.Sedimentation : During sedimentation, floc settles to the bottom of the water supply, due to its weight d.Filtration: Once the floc has settled to the bottom of the water supply, the clear water on top will pass through filters of varying compositions (sand, gravel, and charcoal) and pore sizes in order to remove fine particles that were not settled, such as dust, parasites, bacteria, viruses, and chemicals

e.Disinfection : involves the addition of chemicals in order to kill or reduce the number of pathogenic organisms

During the treatment processed, colloids, suspended matter, pathogenic microorganisms and organic matter are removed to meet the regulated standard. However, the quality of raw water isn't always stable, and corresponding actions are required to be promptly adopted when events like the surge of pollutants or the large variability of the influent flow. In any event, the treated water from DWTPs should generate drinking water which complies the World Health Organization's Guidelines (WHO's guideline) for drinking water quality. Otherwise, the treated drinking water should either be discharged and result in the short term outage of water supply to the downstream cities or the users will receive contaminated drinking water which can potentially transmit diseases and cause illness.

Turbidity is one of the critical water quality indicators, which can be defined as the "optical quality" of water, and the unit to decribe the turbidity is called Nephelometric Turbidity Unit (NTU). High levels of turbidity in raw water can impede the effectiveness of filtration and chlorination processes, and potentially cause short-term outages of water supply. Heavy rainfall and fissures within the aquifer can also lead to turbidity events are mostly likely to cause high turbidity (World Health Organization, 2017). The challenge in event of high turbidity in raw water is it occurs rapidly and mitigating activities must be actionable immediately. To address sudden event of such, Stevenson and Bravo (2019) trained forecasting models based on general linear model (GLM) and RF to predict the time when the turbidiy reaches higher than 7 NTU. The results indicate both model can successfully predict the events (i.e., with accuracy between 0.81 and 0.86), and RF model is found to have higher precision due to it's ability to capture the nonlinear relationship between rainfall (mm) and turbidity (NTU).

To maintain operational costs and water quality in the coagulation process, the amount of coagulant, which is mainly subject to the turbidity and alkalinity in the raw water, is traditionally determined thourgh manually sampling and analysis. Jar test is designed to find out the optimal chemical dosage for coagulation to remove the turbidity in raw water, and the entire process includes on-site sampling and up to more than 40 minutes of laboratory works (Gani et al., 2017). To replace the laborious procedure of jar tests, Wang et al. (2022) proposed using principal component regression (PCR), support vector regression (SVR), and long short-term memory (LSTM) neural network to build predictive

models for outputing daily estimated chemical dosage. Compared with linear PCR model, nonlinear SVR and LSTM models captures more relationship between the chemical dose (e.g., ferric sulfate) and the raw water quality based on a higher R-squared value of 0.70.

Disinfection is the last step of water treatment processes in drinking water treatment plants to generate safe potable water. In this step, one or more chemical disinfectants like chlorine, chloramine, or chlorine dioxide are added into the water to inactivate any remaining pathogenic microorganisms. However, the chlorination process requires precise dosing of disinfectant—too high will lead to the formation of disinfection byproducts (DBPs), and too low will result in insufficient levels of the residaul disinfectant concentration. In both senarios, the treated drinking water can pose health threats to the end users. The aforementioned PID controller can achieve automatic dosing of disinfection, however, Wang et al. (2020) found out that the accuracy of the predicted disinfectant dosage using (i.e., chlorine is used in this paper) a Support Vector Regression (SVR) model outperformed a PID controller in both simulation and experimental conditions. An Artificial Nerual Network based model also shows a more satisfied cost reduction in a chlorination dosing control system comapred to PID controller (Librantz et al., 2018).

The invariability of the raw water quality is always a big issue for disinfection. For instance, chlorine dose can be excessive dosed when the treated water contains less pollutants (e.g., non-organic matters and ammonia nitrogen). Exessive addition of chlorine results in the problem of wasting chemicals which is reflected on the increase operational cost and potentially generate undesired disinfection by-products (e.g., trihalomethanes (THMs), which are carcinogenic to human) due to the chemcial reaction between pollutants and overly dosed chlorine. Xu et al. (2022) trained an ANN model for predcting the occurrence of THMs in tap water using simple and easy water quality parameters (e.g., pH, temperature, UVA_{254} and residual chlorine (Cl_2)). Despite the results showed a good model accuracy in predicting for THMs (i.e., T-THMs, TCM and BDCM), the applications of the model is largely limited in reality due to the lack of dataset regarding the quantity and quality . In fact, lack of high quality dataset for trianing ML models is a common issue, which explains up until recently, mathametical or empirical based AI models like fuzzy logic (Gamiz et al., 2020; Godo-Pla et al., 2021) is still widely used for process control in WTPs.

2.2.2 Wastewater treatment plants

Human activies produce wastewater and discharge from homes, businesses, factories and commercial activities to the sewage systems which connect to wastewater treatment plants (WWTPs). The function of a WWTP is to remove contaminants from sewage and water so that the treated water can be returned to the natural water body without dangering any living beings reside in the ecosystem. Undertreated wastewater can lead to harmful algal blooms or cause oxygen deficit in the water (i.e., low oxygen content can kill the fishes). The steps for treating municipal wastewater involve three major categories—primary treatment, secondary treatment and tertiary treatment. The pollutants which will either float or settle will be removed in primary treatment; next, secondary treatment is mainly responsible for removing BOD_5 in the biological processes; in the final tertiary treatment, membrane filtration, adsorption by activated carbon and addition of disinfectant can be applied optionally to futher eliminate the undesired pollutants in the water.

Wastewater can be defined as the flow of used water discharged from homes, businesses, industries, commercial activities and institutions which is transported to treatment plants via pulbich sewer system or engineered network of pipes. This wastewater is further categorized and defined according to its sources of origin. Domestic wastewater refers to water discharged from residential sources generated by kitchen wastewater, cleaning and personal hygiene. Industrial/commercial wastewater is generated and discharged from manufacturing and commercial activities, such as textile industry and food and beverage processing wastewater. Institutional wastewater characterizes wastewater generated by large institutions such as hospitals and educational facilities. Regardless of the source of the wastewater, WWTPs have to achieve at least three sustainability targets: environmental protection (i.e., low pollutants discharge), social acceptance (i.e., human sanitary protection) and economic development (i.e., feasible operational and management costs) (Mannina et al., 2019). To effectively achieve these goals, process control is requried to reduce energy consumption, improve on effluent quality, and save costs in plant operation and management. The focus of this study is on discussing the development of using process control for treatment operation and management.

Under known operational conditions of a WWTP, machine learning models can be trained to assist the plant operators optimize treatment processes to improve effluent

quality . Wang et al. (2021) proposed a machine learning framework, utilizing a model based on Random Forest to predict the effluent Total Suspended Solid (TSS) and phosphate (PO_4). This study features using collected data from six on-line sensors (i.e., flow rate, TSS, pH, PO_4, temperature, and total solids (TS) meters) across the treatment line to train the RF model. The results indicated that the influent temperature is the most influential variable for both TSS and PO_4 in the effluent, and PO_4 depends strongly on the TSS in aeration basins, etc. It has been suggested that the combined use of RF model and analytical tools allows the author to pinpoint the critical factors influencing on the effluent quality, and this seems to be a innovative approach. However, there are severl major drawbacks hindering such model developments using on-line sensors to collect training data. Many of the existing WWTPs and DWTPs are not equipped with on-line sensors, and lack of automation and instrumentation is common. One of the examples that lack of data from on-line sensor is an emerging techology called aerobic granular sludge (AGS) in secondary treatment (i.e., biological treatment). In addition, Wilén et al. (2018) claimed that the complex nonlinear relationships between the sludge, wastewater quality and operational conditions makes the operation and management of AGS difficult. Awaring the high complexity of the AGS and the unavailabilities of on-line sensors, Zaghloul et al. (2021) attempted to address the issues by collecting data from lab-based reactors and training machine learning models. Considering the intricacy of operation coditions and the AGS system, the author claimed that with the use of feature selection and ensemble model, which is train with three different ML models, overfitting can be prevented. Given that the findings in this study provided good model performance in predicting Chemical Oxygen Demand (COD) and other sludge-related parameters, the results stating the fact of reducing overfitting using ensemble learnings should be treated with caution. Similar to the AGS system, electrocoagulation reactor is also an complex system that the operation and management are based on pH value, the current density, flow rate and the initial concentration of heavy metal ions, etc. Interestingly, instead of using an ensemble model to prevent the overfitting issue claied by Zaghloul et al. (2021), Zhu et al. (2021) used a deep learning Long and Short-term model (LSTM) and a error compensate Autoregressive Integrated Moving Average model (ARIMA) to predict the removal rate of heavy metal ion concentration in wastewater. A LSTM-ARIMA model has strengthed the model performance compared to solely used LSTM or ARIMA model in predicting removal rate shown by the Results. A possible rationalization of using as

LSTM model without worrying model overfitting is that deep learning is sophsiticated enough for learning the nonlinear patterns in complex system while machine learning model like RF might fail to capture the intricate relationships, resulting in overfitting.

The advancement in technology allows the easy access to real-time water quality data via on-line sensors. The collected real-time data can be used to train predictive models and assist the plant operation and management. Depite the advantages of what on-line sensors are capable of, the pitfalls can jeopardize the quality of predictives models or even induce wrong decisions for plant operation, ultimately deteriorate treatment efficiency in WWTPs. Haimi et al. (2015) suggested that reliable and moderately-priced real-time sensors are not always available, in addition, sensor malfunctions (i.e., fouling or erroneous measurement) can cause the down-time of the sensors. For the unavaible sensors (i.e., "hard-to-measure" or expensive sensors), many research works have proposed building "soft sensors". Instead of using hardware sensors to measure the water paratmers, soft sensor generates real-time values through a machine learning model, which is trained by other easy-to-measure water quality data. In the works of Wang et al. (2019), easy-to-measure variables such as, pH, flow rate, TSS, and ammonium nitrate (NH_4-N) are input to machine learning models to predict hard-to-measure water quality paratmers of COD and total phosphate (TP). Pattnaik et al. (2021) also used DO, pH, conductivity, turbidity, and temperature to train a model to predict BOD. It's beleived that both research works can solve the issues of the uavailability of certain water quality sensors.

The automated treatment operation and management heavily relies on the reliability of the on-line sensors, thus, preventing and the early detection of when the sensors are malfunctioned is the upmost concern to the plant operators. Sensor fault detections can be catogorized into three groups, which are (1) individual faults—an outlier data which can be distinguished with the respect to others data points; (2) contextual faults—an anomalous instance in a specific context and normal in another; (3) collective faults—a cluster of irregular intances with respect to other data trends (Chandola). Many research papers have proposed using machine learning models to help identify the sensor fouling.

Two main types algorithms, which is regression and classification can be used for finding fouling signals. A regression algorithm can identify fouling signals by comparing model predicted outputs (e.g., ammonium or COD concentration) to the actual signals; a classification algorithm can distinguish fouling signals through the direct outputs of the

model (i.e., the model outputs 2 class labels, one can be assigned as normal and the other is abnormal signal). Cecconi and Rosso (2021) proposed a ammonium fault detection mechanism, utilizing a regression ANN model, along with principal component analysis (PCA) and Shewhart monitoring charts (i.e., statistical control chart). The remarkable idea from this study is to analyze the residaul between the predicted ammonium and the real ammonium snesor signal and identify the individual and contextual faults with the help of statistical tools. Despite the accuracy of fault detection mechanism can reach R^2 value of 0.87, the method comes with great limitations. The author points out to maintain the high accuracy of the predictive model, the quality of the input data needs to be carefully attended by performing manual cleaning procedures on a weekly basis.

Classification algorithms as an alternative solution is used to address fault detection problem in the works of Mamandipoor et al. (2020). It is believed that this is the first research paper using a LSTM network to achieve a fully automatic fault detection method in WWTPs. Contrast to others works, input variables for model training heavily relies on the manual selection of the experts before inputing into models like PCA and fuzzy nerual networks. The significance of using a LSTM network, which is a deep learning model, is it's capability of capturing long-term temperol dependencies from a large dataset compared to machine learning models (i.e., PCA-SVM model). The results showed that the accuracy (i.e., F1-score) from LSTM model is 92%, outperformed the PCA-SVM model of 87%. This finding can suggest DL models fits more in solving classification problems for fault detection compared to ML models.

2.2.3 Reclaimed water system and water body

2.3 Tools and techniques for enchancing the performance of machine learning modeling

2.3.1 Programming languages

(Mamandipoor et al., 2020)

2.3.2 Data preprocessing

Poor connection, sensor failures, or fading signal strength, are some of the causes. There are a number of techniques in the literature of time series data to deal with missing values, such as simply ignoring the whole data point with a missing value, filling it with statistically related data, or using more complicated methods to estimate the missing value. S

2.3.3 Feature engineering

The main factors affecting the removal rate of the electrocoagulation reactor are comprehensively considered and analyzed, and find out the input variables of the prediction model, and the input variables of the prediction model are found out. In addition, in order to facilitate the network to analyze and learn the overall trend of the data, and to predict the reactor removal rate more accurately, the change gradient value of the historical removal rate at two adjacent times of the reactor is extracted as the feature value to reflect the change trend of the removal rate. (Zaghloul et al., 2021)

To help with the analysis of ammonia, several statistical measures have been extracted from this feature, such as mean, maximum, minimum, variance and standard deviation, which increase the total number of features to 16. The data are segmented to a maximum window size to create the sequences for the LSTM neural network. The LSTM network would learn the proper amount of information from this window. (Mamandipoor et al., 2020)

CHAPTER 3

METHODS AND MATERIALS

3.1 Wastewater treatment plant description

3.1.1 Treatment processes

TheMBRisaprocessthatintegratesbiodegradationofcontaminants byactivatedsludge,withdirectsolid-liquidseparationbymembrane filtration,i.e.throughaMForUFmembrane.TheMBRtechnologyiscurrentlywi wideuseofMBRshasbeenattributedtoitsnotableadvantages,such ashighqualityofproducedwater,highbiodegradationefficiencyof contaminants,andanoverallsmallerfootprint(JuddandJudd,2011). Thistechnologypermitsbioreactoroperationwithconsiderably highermixedliquorsuspended-solids(MLSS)concentrationthanCAS systems,whicharelimitedbysludgesettlingphenomena.Theprocess inMBRsistypicallyoperatedatMLSSintherangeof8–12g/L,while CASisoperatedintherangeof2–3g/L(Melin generationofslow-growingbacteria,whichhavetheabilitytodegrade certainbiologically-recalcitrantorganica (Clouzotetal.,2011).Therefore,despitenotbeendesignedtoremove organicandinorganicmicrop-ollutants,MBRsmayprovideeffectiveremovalofsomeoftheCEC.EarlystudiesreportedimprovedCECremoval withMBRscomparedtoCAS,asMBRsoperateatahigherSRTthanCAS, thusenhancingcontam-inantbiodegradability(Holbrooketal.,2002; Stephensonetal.,2007).However,whenMBRsandCASwerecomp undersimilaroperatingconditions(i.e.,SRT,temperature)intheremovalofCEC,nosignificantdifferenceswere 2006;Boujuetal.,2009;WeissandReemtsma,2008;Abegglenetal., 2009).Therefore,itwaspostulatedthatMBR performsimilaraslongasthesameoperatingconditionsareprovided, althoughMBRsmayoutper-formCASathigherSRT.Thisisbecause CECaregenerallyhighlysolubleandrelativelysmallcom-pounds,typicallybelow1000Da,whichcanfreelypassthroughthemembranes usedinMBRsystem-stherebyindicatingthatthosemembraneshave nodirectimpactontheremovalofCEC(Snyderetal.,2007).Othe (Radjenovicetal.,2009;Luoetal.,2014).(Krzeminski et al., 2019)

3.1.2 Reclaimed water standard

3.2 Data collection and preparation

Most AI techniques were modeled using experimental data to simulate, predict confirm, and optimize contaminant removal in wastewater treatment processes. Experimental data set were either divided into three parts (training, validation, and testing) or two parts (training and testing). The training set was used to develop the model, the validation data set was used to optimize the model, and the testing data set was used to test the model in the prediction stage.

3.2.1 Ammonia data monitoring and collection

Ammonium and Potassium Probe for IQ Sensor Net System The WTW ISE sensor AmmoLyt®Plus 700 IQ is a rugged, ion selective electrode (ISE) probe for the IQ Sensor Net process monitoring system. The AmmoLyt provides continuous and reagentless monitoring of ammonium and potassium using the most precise and reliable electrodes on the market. The individually replaceable electrodes have a typical lifetime of 18 to 24 months and are warrantied for 1 year minimizing maintenance effort and ownership cost.

3.2.2 Color data monitoring and collection

3.2.3 Metrics for model evaluation

AI methods have been demonstrated to be effective in controlling chlorination, while ML models are effective in modeling DBP concentrations, as well as modeling important parameters for adsorption and membrane-filtration processes. The results are often evaluated using various statistical measures including the coefficient of correlation (R), the coefficient of determination (R2), the mean average error (MAE), the mean square error (MSE), the root mean square error (RMSE), and relative error (RE).

3.2.4 Data cleaning and pre-processing

However, the raw high-resolution data from each meter were compressed by averaging over 10-minute periods to obtain time series with temporal resolutions of 10 min.

he original data were embedded in multiple matrices and were very messy, with missing values, bad data cells, and unnecessary information. Therefore, the Python modules Numpy (Oliphant, 2006) and Pandas (McKinney, 2010) were used to prepare an organized 'clean' dataset for analysis. This dataset contained 105,861 samples (data points) with 34 variables, giving a matrix size of 105,861 $\times$ 34. The samples were organized in time series with 10min intervals.

3.2.4.1 Data smoothing with Savitzky-Golay filter

3.2.4.2 Exponentially Weighted Moving Average

3.2.4.3 Outlier Removal

3.2.5 Data transformation

Split of Train/valid/test dataset

3.3 Architecture design of the selected baseline models

3.3.1 Random Forest

F can be described as an ensemble method in which the final result is obtained by aggregating (through averaging in the case of regression) results from multiple weak learners known as Classification and Regression Trees (CARTs) (Breiman, 2017). Each weak learner (tree) is trained on the bootstrap set, which is obtained by sampling with replacement from the original training set. For trees, the input variables are used to generate nodes. These variables are selected partially and randomly as a subset in every split, then the variable contributing to the smallest sum of impurity of two child nodes at a certain split point is chosen as the split variable. This is done repeatedly until the trees don't need to split anymore. The regression impurity of a particular node is defined by Eqs. (2), (3) and (4), (Wang et al., 2021)

3.3.2 LSTM

Recently, deep recurrent neural networks (RNN) such as long short-term memory networks (LSTM) have shown breakthrough results over state-of-the-art machinelearning methods in many applications with non-linear temporal data, including robotics, high-energy physics and computational geometry (Goodfellow et al. 2016). These methods can successfully engineer appropriate long-term temporal dependencies and variable length features, significantly lessening the need to pre-process data with respect to traditional machine-learning methods or statistical approaches. It is the ability to capture the long-term dependencies that make LSTM networks particularity fitting for the problem at hand.

Fig. 1 The general schema of a RNN unit versus a LSTM one (adapted from Olah 2015)

3.3.3 RNN

3.3.4 GRU

3.4 Implementation of regularization

3.4.1 Scheduler

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Baseline performance of ammonia concentration and colour level forecasting models

4.1.1 Machine learning vs deep learning

RF is used in this work as a representative tree-based modeling strategy because RF models have some major advantages over alternative tree-based models; notably, they require fewer hyperparameters for tuning, their performance is robust to hyperparameter changes, and they are less likely to suffer from overfitting (Breiman, 2001;Breiman, 2002 ;Chen and Guestrin, 2016;Fawagreh et al., 2014 ;Ke et al., 2017).

4.2 Improved performance on forecasting models using data pre-processing techniques

4.3 Data enrichment via feature engineering based on effluent quality pattern

4.4 Design of model architecture through analyzing wastewater composition in sewer system

CHAPTER 5

CONCLUSION

Bibliography

Sunil L Andhare and Prasad J Palkar. SCADA a tool to increase efficiency of water treatment plant. *Asian Journal of Engineering and Technology Innovation*, page 8, 2014.

Jhon Stalin Figueroa Bados and Iralmy Yipsy Platero Morejon. Design of a PID Control System for a Wastewater Treatment Plant. In *2020 3rd International Conference on Robotics, Control and Automation Engineering (RCAE)*, pages 31–35, Chongqing, China, November 2020. IEEE. ISBN 978-1-72818-638-2. doi: 10.1109/RCAE51546.2020.9294199.

Francesca Cecconi and Diego Rosso. Soft Sensing for On-Line Fault Detection of Ammonium Sensors in Water Resource Recovery Facilities. *Environmental Science: Water Research and Technology*, 2021. doi: 10.1021/acs.est.0c06111.

Varun Chandola. Anomaly Detection : A Survey. page 72.

J.C. Chen, N.B. Chang, and W.K. Shieh. Assessing wastewater reclamation potential by neural network model. *Engineering Applications of Artificial Intelligence*, 16(2): 149–157, March 2003. ISSN 09521976. doi: 10.1016/S0952-1976(03)00056-3.

Feridun Demir and Wilbur W. Woo. Feedback control over the chlorine disinfection process at a wastewater treatment plant using a Smith predictor, a method of characteristics and odometric transformation. *Journal of Environmental Chemical Engineering*, 2(2):1088–1097, June 2014. ISSN 22133437. doi: 10.1016/j.jece.2014.04.006.

Javier Gamiz, Ramon Vilanova, Herminio Martinez-Garcia, Yolanda Bolea, and Antoni Grau. Fuzzy Gain Scheduling and Feed-Forward Control for Drinking Water Treatment Plants (DWTP) Chlorination Process. *IEEE Access*, 8:110018–110032, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3002156.

Paran Gani, Norshuhaila Mohamed Sunar, Hazel Matias-Peralta, and Ab Aziz Abdul Latiff. Effect of pH and alum dosage on the efficiency of microalgae harvesting via flocculation technique. *International Journal of Green Energy*, 14(4):395–399, March 2017. ISSN 1543-5075, 1543-5083. doi: 10.1080/15435075.2016.1261707.

Lluís Godo-Pla, Jose Javier Rodríguez, Jordi Suquet, Pere Emiliano, Fernando Valero, Manel Poch, and Hèctor Monclús. Control of primary disinfection in a drinking water treatment plant based on a fuzzy inference system. *Process Safety and Environmental Protection*, 145:63–70, January 2021. ISSN 09575820. doi: 10.1016/j.psep.2020.07.037.

Hong Guo, Kwanho Jeong, Jiyeon Lim, Jeongwon Jo, Young Mo Kim, Jong pyo Park, Joon Ha Kim, and Kyung Hwa Cho. Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *Journal of Environmental Sciences (China)*, 32:90–101, 2015. doi: 10.1016/j.jes.2015.01.007.

Henri Haimi, Francesco Corona, Michela Mulas, Laura Sundell, Mari Heinonen, and Riku Vahala. Shall we use hardware sensor measurements or soft-sensor estimates? Case study in a full-scale WWTP. *Environmental Modelling and Software*, 72:215–229, 2015. doi: 10.1016/j.envsoft.2015.07.013.

Edmund A. Kobylinski, Gary L. Hunter, and Andrew R. Shaw. On Line Control Strategies for Disinfection Systems: Success and Failure. *Proceedings of the Water Environment Federation*, 2006(5):6371–6394, January 2006. ISSN 1938-6478. doi: 10.2175/193864706783761716.

Pawel Krzeminski, Maria Concetta Tomei, Popi Karaolia, Alette Langenhoff, C. Marisa R. Almeida, Ewa Felis, Fanny Gritten, Henrik Rasmus Andersen, Telma Fernandes, Celia M. Manaia, Luigi Rizzo, and Despo Fatta-Kassinos. Performance of secondary wastewater treatment methods for the removal of contaminants of emerging concern implicated in crop uptake and antibiotic resistance spread: A review. *Science of The Total Environment*, 648:1052–1081, January 2019. ISSN 00489697. doi: 10.1016/j.scitotenv.2018.08.130.

Lei Li, Shuming Rong, Rui Wang, and Shuili Yu. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chemical Engineering Journal*, 405:126673, February 2021. ISSN 13858947. doi: 10.1016/j.cej.2020.126673.

Peifeng Li, Jin Zhang, and Peter Krebs. Prediction of Flow Based on a CNN-LSTM Combined Deep Learning Approach. *Water*, 14(6):993, March 2022. ISSN 2073-4441. doi: 10.3390/w14060993.

Zhe Li, Caiwen Ding, Siyue Wang, Wujie Wen, Youwei Zhuo, Chang Liu, Qinru Qiu, Wenyao Xu, Xue Lin, Xuehai Qian, and Yanzhi Wang. E-RNN: Design Optimization for Efficient Recurrent Neural Networks in FPGAs, December 2018.

André Felipe Librantz, Fábio Cosme Rodrigues dos Santos, and Cleber Gustavo Dias. Artificial neural networks to control chlorine dosing in a water treatment plant. *Acta Scientiarum. Technology*, 40(1):37275, September 2018. ISSN 1807-8664, 1806-2563. doi: 10.4025/actascitechnol.v40i1.37275.

Behrooz Mamandipoor, Mahshid Majd, Seyedmostafa Sheikhalishahi, Claudio Modena, and Venet Osmani. Monitoring and detecting faults in wastewater treatment plants using deep learning. *Environmental Monitoring and Assessment*, 192(2), 2020. doi: 10.1007/s10661-020-8064-1.

Giorgio Mannina, Taise Ferreira Rebouças, Alida Cosenza, Miquel Sànchez-Marrè, and Karina Gibert. Decision support systems (DSS) for wastewater treatment plants – A review of the state of the art. *Bioresource Technology*, 290:121814, October 2019. ISSN 09608524. doi: 10.1016/j.biortech.2019.121814.

Bhawani Shankar Pattnaik, Arunima Sambhuta Pattanayak, Siba Kumar Udgata, and Ajit Kumar Panda. Machine learning based soft sensor model for BOD estimation using intelligence at edge. *Complex & Intelligent Systems*, 7(2):961–976, 2021. doi: 10.1007/s40747-020-00259-9.

I. Santín, C. Pedret, and R. Vilanova. Fuzzy Control and Model Predictive Control Configurations for Effluent Violations Removal in Wastewater Treatment Plants. *Industrial & Engineering Chemistry Research*, 54(10):2763–2775, March 2015. ISSN 0888-5885, 1520-5045. doi: 10.1021/ie504079q.

Matthew Stevenson and Cristián Bravo. Advanced turbidity prediction for operational water supply planning. *Decision Support Systems*, 119:72–84, April 2019. ISSN 01679236. doi: 10.1016/j.dss.2019.02.009.

Dong Wang, Sven Thunéll, Ulrika Lindberg, Lili Jiang, Johan Trygg, Mats Tysklind, and Nabil Souihi. A machine learning framework to improve effluent quality control in wastewater treatment plants. *Science of The Total Environment*, 784:147138, August 2021. ISSN 00489697. doi: 10.1016/j.scitotenv.2021.147138.

Dongsheng Wang and Hao Xiang. Composite Control of Post-Chlorine Dosage During Drinking Water Treatment. *IEEE Access*, 7:27893–27898, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2901059.

Dongsheng Wang, Jingjin Shen, Songhao Zhu, and Guoping Jiang. Model predictive control for chlorine dosing of drinking water treatment based on support vector machine model. *DESALINATION AND WATER TREATMENT*, 173:133–141, 2020. doi: 10.5004/dwt.2020.24144.

Hui Wang, Tirusew Asefa, and Jack Thornburgh. Integrating water quality and streamflow into prediction of chemical dosage in a drinking water treatment plant using machine learning algorithms. *Water Supply*, 22(3):2803–2815, March 2022. ISSN 1606-9749, 1607-0798. doi: 10.2166/ws.2021.435.

Xiaodong Wang, Knut Kvaal, and Harsha Ratnaweera. Explicit and interpretable nonlinear soft sensor models for influent surveillance at a full-scale wastewater treatment plant. *Journal of Process Control*, 77:1–6, 2019. doi: 10.1016/j.jprocont.2019.03.005.

Britt-Marie Wilén, Raquel Liébana, Frank Persson, Oskar Modin, and Malte Hermansson. The mechanisms of granulation of activated sludge in wastewater treatment, its optimization, and impact on effluent quality. *Applied Microbiology and Biotechnology*, 102(12):5005–5020, June 2018. ISSN 0175-7598, 1432-0614. doi: 10.1007/s00253-018-8990-9.

World Health Organization. Water quality and health - review of turbidity: Information for regulators and water suppliers. Technical report, World Health Organization, Geneva, 2017.

Jianlong Xu, Zhuo Xu, Jianjun Kuang, Che Lin, Lianghong Xiao, Xingshan Huang, and Yufeng Zhang. An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies. *Water*, 13(22): 3262, November 2021. ISSN 2073-4441. doi: 10.3390/w13223262.

Zeqiong Xu, Jiao Shen, Yuqing Qu, Huangfei Chen, Xiaoling Zhou, Huachang Hong, Hongjie Sun, Hongjun Lin, Wenjing Deng, and Fuyong Wu. Using simple and easy water quality parameters to predict trihalomethane occurrence in tap water. *Chemosphere*, 286:131586, January 2022. ISSN 00456535. doi: 10.1016/j.chemosphere.2021.131586.

Mohamed Sherif Zaghloul, Oliver Terna Iorhemen, Rania Ahmed Hamza, Joo Hwa Tay, and Gopal Achari. Development of an ensemble of machine learning algorithms to model aerobic granular sludge reactors. *Water Research*, 189:116657–116657, 2021. doi: 10.1016/j.watres.2020.116657.

Hongqiu Zhu, Qiling Wang, Fengxue Zhang, Chunhua Yang, and Yonggang Li. A prediction method of electrocoagulation reactor removal rate based on Long Term and Short Term Memory - Autoregressive Integrated Moving Average Model. *Process Safety and Environmental Protection*, 152:462–470, 2021. doi: 10.1016/j.psep.2021.06.020.

Huijun Zhu and Xinglei Qiu. The Application of PLC in Sewage Treatment. *Journal of Water Resource and Protection*, 09(07):841–850, 2017. ISSN 1945-3094, 1945-3108. doi: 10.4236/jwarp.2017.97056.