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

METHODS AND MATERIALS

1.1 Wastewater treatment plant description

1.1.1 Process and data sources in SWHEPP

Shek Wu Hui Effluent Polish Plant (SWHEPP) is a secondary sewage treatment plant, which treats the municipal wastewater of the Sheung Shui, Fanling Districts and adjacent areas, and treated leachate effluent from North East New Territories (NENT) leachate treatment plant. The plant is designed for 300,000 population equivalents (PE) in 2001, and in 2009, the daily treatment capacity has been expanded from 80,000 m^3/day to 93,000 m^3/day. SHWEPP is operated and maintained by Drainage Services Department (DSD), and the plant will be upgraded to teritary treatment level to increase the treatment capacity of 190,000 m^3/day by the end of 2025. As shown in Fig. 1.1, the treatment plant is mainly comprised of primary sedimentation, secondary biological treatment, and final sedimentation followed by a membrane bioreactor (MBR), which provides an advanced level of organic and suspended solids removal. To monitor the effluent quality in real-time, low volume of the MBR effluent is pumped to an effluent container near by the MBR location. Two on-line meters, ammoniacal nitrogen on-line sensor and colour level on-line analyzer are installed in the effluent container, which are indicated as (a) and (b) in Fig. 1.1.

1.2 Data collection and preparation

1.2.1 On-line data monitoring and collection

To enable us to perform on-line monitoring of ammonium concentration (NH_3-N) in the MBR effluent, a Ammonium and Potassium Probe, AmmoLyt®Plus 700 IQ (Xylem Company) is installed as in Fig. 1.3a in the effluent container, as shown in Fig. 1.2. The operation was commenced on 27 April 2021 and completed on 27 March 2022. The

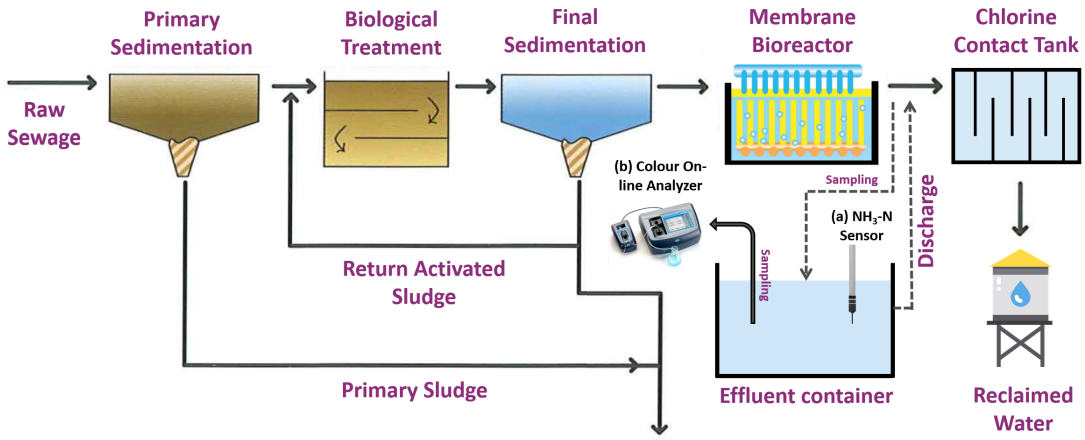


Figure 1.1: Sewage treatment process flowchart at SWHEPP (adapted from Drainage Services Department 2020)

ion-selective electrode (ISE) probe provides continuous and reagentless monitoring of ammonium and potassium at the configured interval of one measurement per minute. Due to the ISE probe cannot differentiate the potentials difference cause by ammonium and potassium ions in the electrodes, the on-line monitoring of ammonium concentration requires the continuous calibration using potassium concentration.

The instrument records ammonium concentration as NH_4-N mg/L, a form to express the sum of nitrogen found in reduced nitrogen (III) form. Ammonia has a reported pKa of 9.25 (National Center for Biotechnology Information, 2022), meaning ammonium is a primary species under the pH of 9.25 in water. In WWTPs, the pH in water normally ranges from pH of 7–8, making the NH_4-N concentration the dominant species. Both ammonia and ammonium contain one nitrogen atom, 1 mg/L NH_3-N is the same as 1 mg/L NH_4-N. Thus, to prevent confusion, in the following paragraph the unit of NH_4-N will be expressed by NH_3-N, which is the unit used in the water quality standard. The collection of on-line ammonia data is achieved through downloading csv files from the website connected to the IQ Sensor Controller (Xylem Comapny), as shown in Fig. 1.3b.

An hourly monitoring of the colour levels of MBR effluent was conducted from 5 October 2021 to 26 February 2022 by using a custom-made on-line colour analysis system. Originally, the spectrophotometer as Fig. 1.4a and a peristaltic pump as Fig. 1.4b can only initiate a single measurement of colour level by pressing the "READ" buttom on the DR3900 panel. To realize continuously sampling and analyzing colour level without human intervention, an actuator with programmable time function was mounted on the panel of DR3900, as shown in Fig. 1.4c.

Figure 1.2: Colour levels and ammonia concentration are measure in the effluent container (i.e., on the right of the image.) A water pump transports MBR effluent to the effluent container continuously at real-time. The black vault on the left of the image contains a laptop and a colour spectrophotometer.

The automatic sampling and analzying of the colour level begins with the action of the actuator, by clicking on the "READ" button to initiate the colour analysis at a fixed interval of 30 minutes. 3 mL of sample was collected from the effluent container and delivered to the spectrophotometer cell. Then, the sample was subsequentially analysed by the spectrophotometer with the data transmitted to an automatic data acquisition and storage software pre-installed in the laptop. The DR3900 device is connected to a laptop, which receives the real-time data and stores on a data management software from Hach company. To access the real-time data from the laptop, Google Remote Desktop is used to operate the laptop via Internet cloud services using any devices having access to the Internet. The entire process is illustrated in Fig. 1.5. After the measurement, the sample will be discharged to the effluent container and the online colour monitoring system is left idle until the next measurement.

The maintenance and calibration of the DR3900 spectrophotometer is performed on a weekly basis. During the maintenance, the DR3900 device was shut off, and chlorine solution at the concentration of 100 mg/L was pumped into the sampling tubes and the plastic cuvette for disinfection and cleansing. The cleanse of the tubes and cuvette were

(a) AmmoLyt®Plus 700 IQ, Xylem.

(b) DIQ/S 284-EF controller, Xylem.

Figure 1.3: Instrument of on-line ammonium monitoring system.

manually inspected with eyes to make sure no foreign objects were stuck inside. Deionized water was brought to the site to perform the spectrophotometer calibration after the reboot of DR3900.

Based on the proposed model training methods, which ammonia and colour data are used as the second features of training colour and ammonia forecasting models, the size and time of the ammonia and colour datasets should be the same. In addition, abnormal data caused by sensor downtime should also be excluded. Thus, we chose the ammonia and colour data from 23 December 2021 to 22 January, as shown in Fig. 1.6.

1.2.2 Loss function for model evaluation

Loss functions are used to determine the error between the model outputs (i.e., prediction or forecasting values) and the given target value (DeepAI, 2022). The bigger the difference between the ground truth $\boldsymbol{y}$ and the model outputs $\hat{\boldsymbol{y}}$, the higher the value of the loss function is, meaning the model perfomred poorer. A low value for the loss means the modle perfomred well. The selection of the types of the loss function is essential for training the model to perform specific tasks. In this study, Mean Squared Error (MSE) is used for evaluating the regression models. The values of MSE will never be negative, and is formally defined by the following equation:

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \tag{1.2.1}$$

(a) SIP10 peristaltic pump, Hach

(b) DR3900 spectrophotometer, Hach

(c) Customized clicker/actuator

Figure 1.4: Instruments of on-line colour analysis system.

1.2.3 Data cleaning and pre-processing

In this study, ammonia concentration and colour level forecasting models will be trained, and the model training steps are shown in Fig. 1.7. The training processes are split into two sections, one is the baseline model training steps, the other is proposed model training steps. The training steps of the first section used cleaned data to train forecasting models and generated baseline model performance, which will be further compared with the model perfromance generated from the second section. The second section includes using pre-processed datasets (i.e., data smoothing) and feature engineering enhanced datasets to train the forecasting model. In machine learning, the data used for training models is refered to model inputs, features and variables.

The raw data embedded in the original csv files exists many issues, such as missing values, having extreme low or high values, and unreadable texts, etc. Thus, the data cleaning and pre-processing are necessary for more effective process of model training. Python programming language and related modules of Numpy and Pandas were used

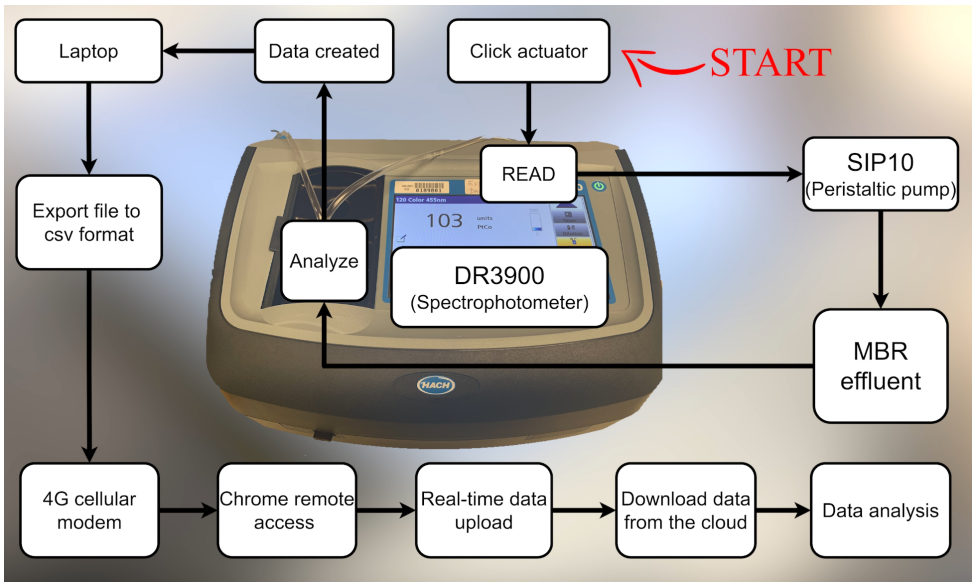


Figure 1.5: Schematic diagram of the custom-made on-line colour anlaysis system.

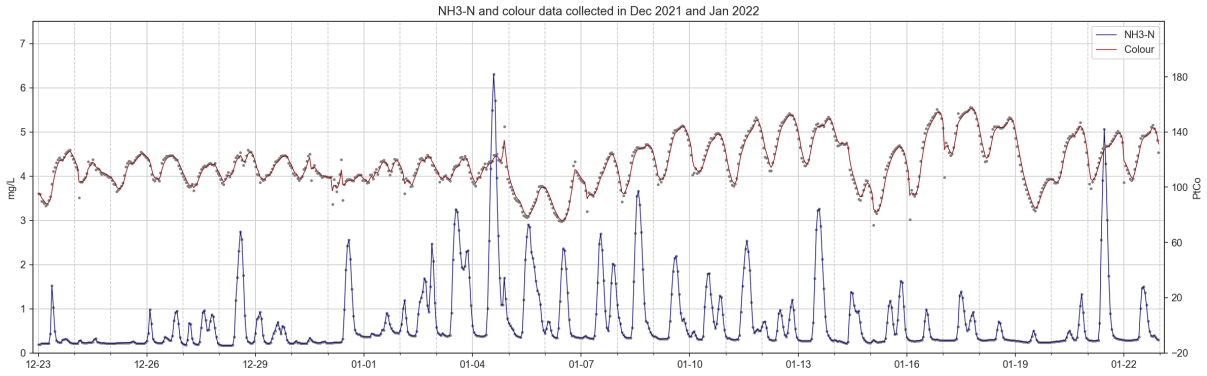


Figure 1.6: Ammonia and colour data collected from 23 December 2021 to 22 January 2022.

to clean and pre-process the raw dataset for further usage. The ammonia raw dataset contained 44,640 samples (data points) with 8 variables, giving a matrix size of 44,640 x 8, and the samples were collected in time series at 1 minute interval. The colour level raw dataset contained 1488 samples with 34 variables, giving a matrix size of 1488 x 34, and the samples were collected in time series at 30 minute interval.

Before the high-resolution data from colour and ammonia datasets were compressed into time series data at 1 hour interval via averaging, extreme values were manually removed. For ammonia dataset, we replaced the values higher than 7.0 mg/L with NaN (i.e., Not a number), and futher use interpolation to fill up the NaN along with the missing values in the dataset. For colour dataset, we manually took out the relatively low data points on the days when the maintenance and calibration tasks were performed;

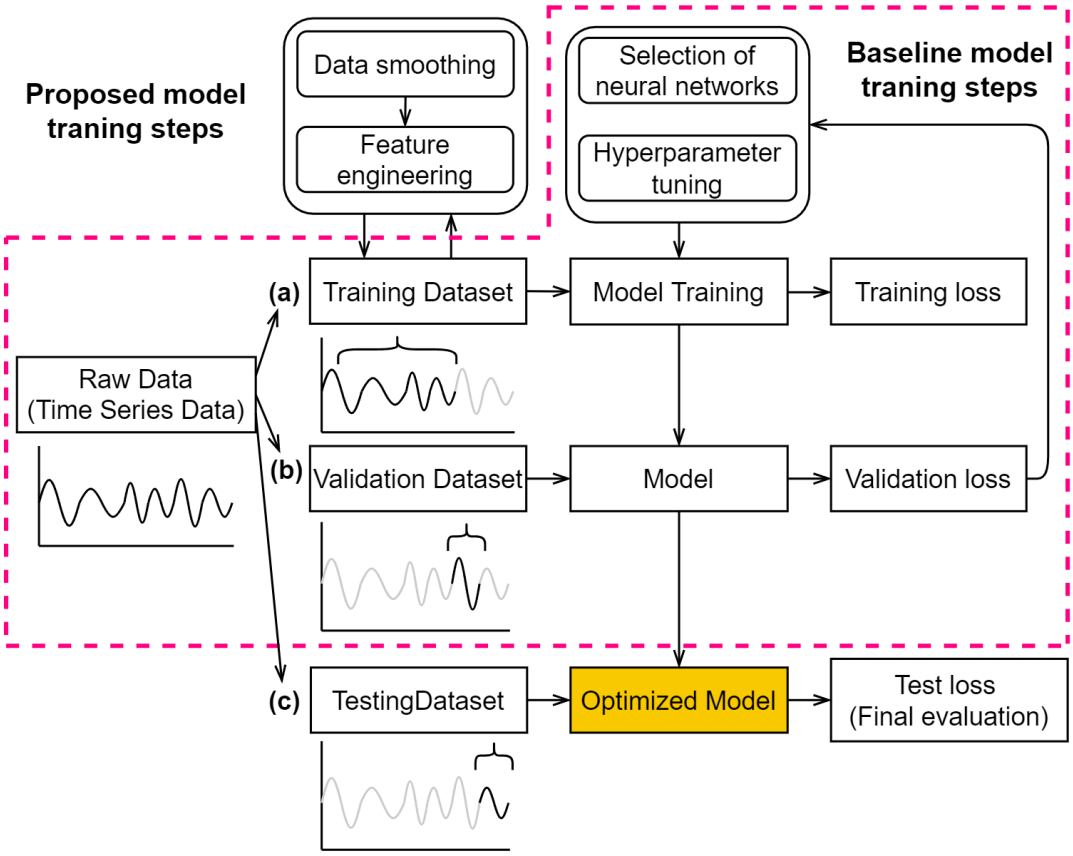


Figure 1.7: Machine learning model training steps.

extremely values higher than 300 Hazen Unit were also replaced by NaN. Same as the data cleaning method used for ammonia dataset, the missing values and NaN were filled up via interpolation.

1.2.3.1 Data smoothing with Savitzky-Golay and EWMA filter

Data smoothing was performed on both ammonia and colour datasets using the same method. One of the effective ways to remove the noise from the dataset is to apply data smoothing filters. Two filteres were applied in this study, Savitzky-Golay (SG) and Exponentially Weighted Moving Average (EWMA) filters.

A SG filter is a digital filter that can be applied to a set of digital data points for the purpose of smoothing the data without distorting the data tendency. This is achieved via convolution, by fitting successive sub-sets of adjacent data points with a low-degree polynomial by the method of linear least squares (Wikipedia, 2022b). The illustration is shown in Fig. 1.8a and the procedures of how data points are smoothed is presented in the following steps:

1) Extract short-time window (i.e., blue dots in Fig.1.8a)

2) Determine polynomial degree (e.g., different polynomial degree is compared in Fig. 1.8a).

3) Find the smoothed data point (i.e., at center of the window).

4) Repeat for shifted window (e.g., similar to moving average).

The equation to decribed the smoothed value of $\boldsymbol{Y_j}$ can be expressed in Eq. 1.2.2:

$$Y_j = (C \otimes y)_j = \sum_{i=\frac{1-m}{2}}^{\frac{m-1}{2}} C_i y_{j+i}, \; \frac{m+1}{2} \leq j \leq n - \frac{m-1}{2} \tag{1.2.2}$$

where Y_j corresponds to the j^{th} smoothed data point, m to the window size (i.e., numer of data points intended to smooth out) and C_i to the convolution coefficients (i.e., determined by Savitzky and Golay (1964)).

Exponentially weighted moving average (EWMA), also known as auto-regressive (AR) filtering, is a technique that filters measurements. An EWMA filter smoothes a measured data point by exponentially averaging that particular point with all previous measurements. The EWMA equation can be expressed in Eq. 1.2.3:

$$\begin{aligned} \alpha &= \frac{2}{span + 1} \\ y_0 &= x_0 \\ y_t &= (1 - \alpha) y_{t-1} + \alpha x_t \end{aligned} \tag{1.2.3}$$

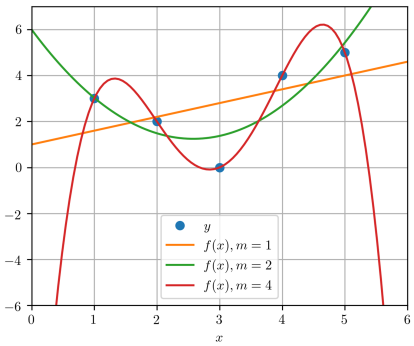
where α corresponds to the decay paratmeter, x_t to the value at a time period, y_t to the value of the EWMA at any time period t, span to the window size.

Both SG and EWMA filters are required to select the hyperparamters, the selected values are presented in Table. 1.1.

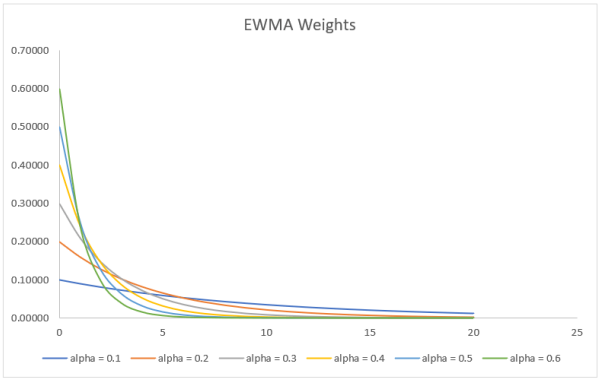
Fig. 1.9 shows the influences of different windows sizes on SG filters as in Fig. 1.9a and on EWMA filters as in Fig. **??**.

1.2.3.2 Outlier Removal

Despite the extreme values in the ammonia raw dataset were removed based on simple conditions (i.e., concentration higher than 7.0 mg/L), the ammonia sensor can still capture



(a) SG filter with different polynomial degree (Taal, 2017).



(b) Examples of weights with exponential decay at varied alpha values (CFI, 2022).

Figure 1.8: Illustration of the influence of different polynomial degrees in the fitting of SG filter and the weigth decay with varied alpha values in EWMA filter.

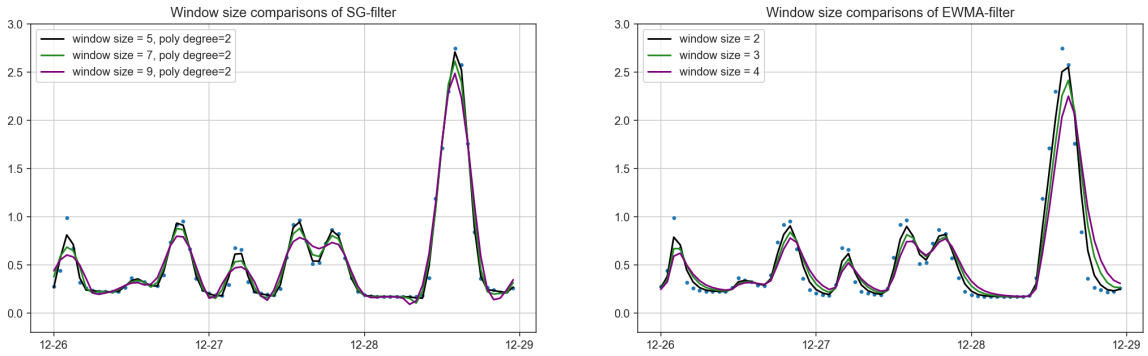
Table 1.1: The selected hyperparameters for SG and EWMA filters.

Group Name	Window size	Polynomial degree
SG-5	5	2
SG-7	7	2
SG-9	9	2
EWMA-2	2	-
EWMA-3	3	-
EWMA-4	4	-

unideal data points collectively. In the outlier removal process, we intended to identify the collective faults of ammonia data in the unit of an entire day. To determine whether the ammonia data on a specific day shows collective fault, two abnormal conditions are defined:

1) NH_3-N fluctuation ≤ 0.1 (i.e., lower than the sensor resolution).
2) No diurnal fluctuation (i.e., Fluctuation = peak value – bottom line value).

To automiatically realize the identification of noraml or abnormal signals, peak analysis was performed on the daily ammonia data. The analysis takes a one-dimension array (i.e., the data form of ammonia in a day) and finds all local maximum values by simple comparison of neighboring values. This function will also provide information such as width and prominence, as in Fig. 1.10 to help us identify whether the diurnal fluctuation is existed.



(a) (b)

Figure 1.9: Illustration of the influence of different polynomial degrees in the fitting of SG filter and the weigth decay with varied alpha values in EWMA filter.

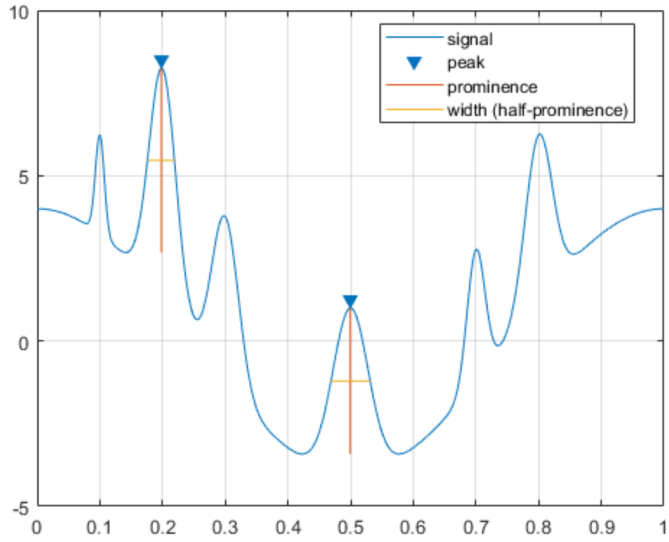


Figure 1.10: Illustration of peak analysis. Four important elements are automatic cacu-lated by the function (MathWorks, 2022).

1.2.3.3 Feature Engineering

To create addition features from the raw datasets, we have carefully observed and analyzed the SWHEPP influent. We discovered that the SWHEPP influent is consisted of treated landfill effluent from NENT landfill leachate site and municipal wastewater, as shown in Fig. 1.11. We observed that with higher blending ratio, which is calculated from the daily volume of treated leachate effluent divided by the daily inflow volume of SHWEPP, the colour level is also higher, as shown in Fig 1.13a. With the the Pearson coefficient of 0.68, the increased volume of treated leachate effluent in public sewege system is proportional to the increase of the colour levels in the SHWEPP influent, while the ammonia concentration is mostly from the municipal wastewater. During the mixing

of both type of the wastewater as in Fig. 1.12a, pollutants contribute to colour levels will be diluted by the municipal wastewater, same as the opposite for the dilution of the ammonia concentration. In Fig. 1.13b, we can observe the time when the lowest colour level of the day occured is close to when the highest of ammonia concentration was observed. The changes of colour levels and ammonia concentration are interative, thus, in feature engineering, colour level data was selected for training ammonia forecasting model; ammonia data was selected for training colour forecasting model, as shown in Fig. 1.17.

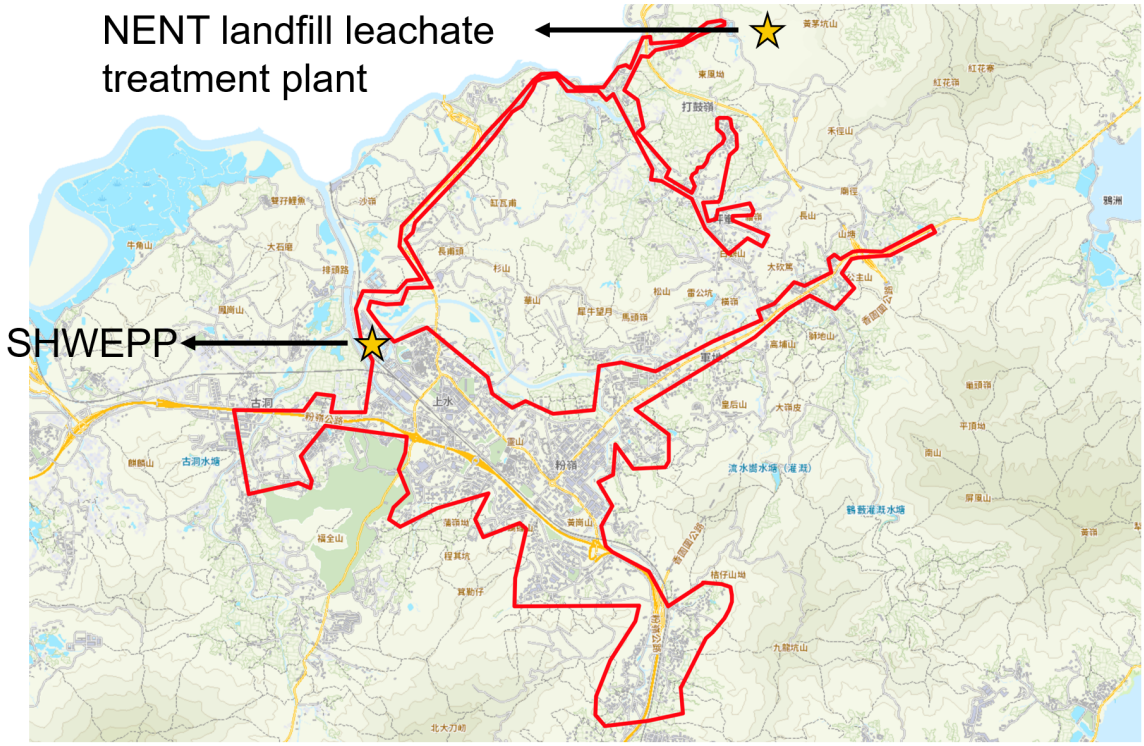
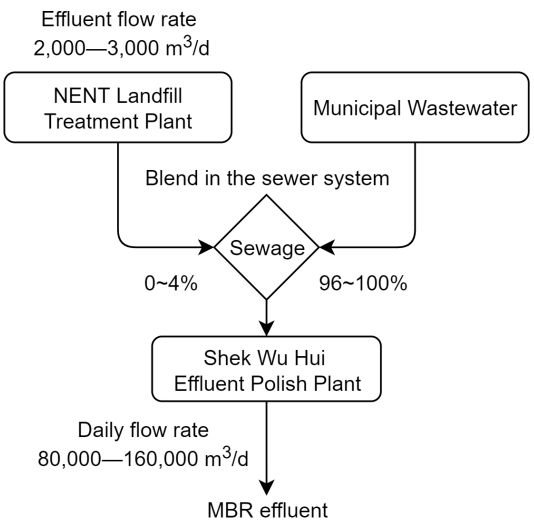
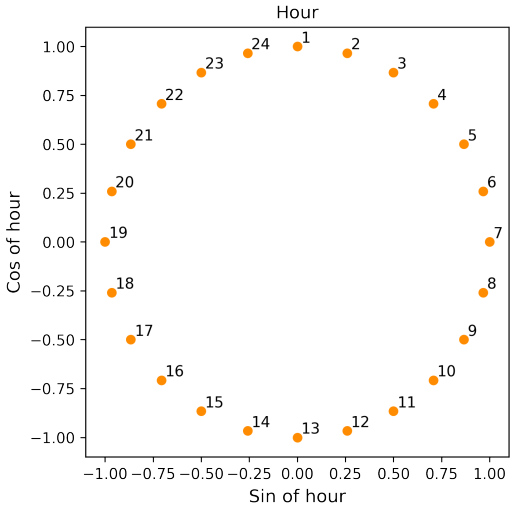


Figure 1.11: Sewer system coverage of SHWEPP. The covered areas (i.e., area circled in red boundary) include Fanling/Sheung-Shui new towns and NENT landfill leachate treament plant.

The new features are inspired from the research work of Abu-Bakar et al. (2021). The author pointed out the four types of hourly household water consumption patterns as in Fig. 1.14, which correlates the specific time of the day to the volume of the water consumed in households. In other words, as fresh water is consumed, wastewater is generated at the same time, the wastewater then enters the public sewage system and result in the increase of ammonia concentration. As shown in Fig. 1.15, the peak analysis tool helped us to identify the peak hour of the ammonia concentration, which occured at around 13:00 to 14:00 o'clock at noon, and 20:00 to 21:00 o'clock at evening. Thus, it is convinced that



(a) Flowchart showing the blending of treated leachate effluent with municipal wastewater.



(b) Positional encoding of hour components.

Figure 1.12: Analysis of influent quality composition and the illustration of the positional encoding.

time features will be able to help the machine learning models to better correlate and predict the change of ammonia concentration in the wastewater.

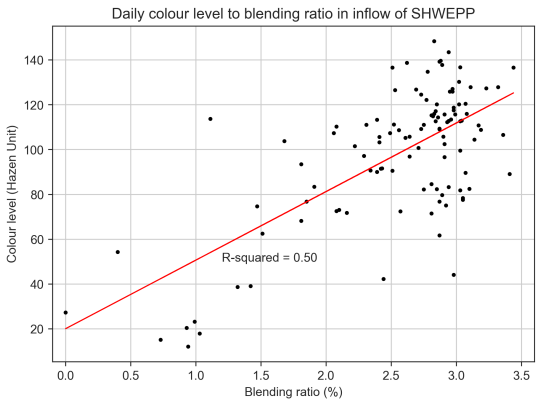
Time feature is realized through a technique called positional encoding (POS). The positioanl encoded feature was achieved as the following steps:

1) The timestamp are represented as three elements—hour, day and month.
2) Each element will bed decomposed into sine and cosine components.
3) Last step is applied to hours and days to make all elements represented cyclically.

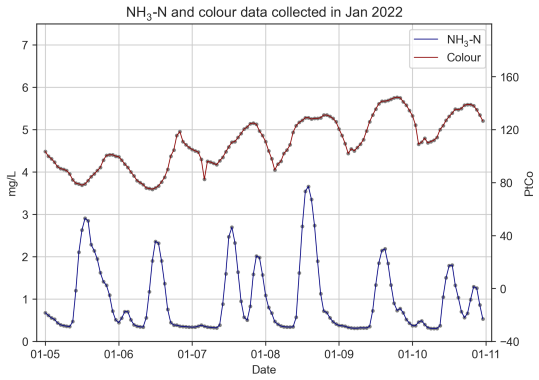
Due to the size of the datasets used in this study for training ammonia and colour forecasting model is 31 days, only hour element was transformed into sine and cosine components as in Fig. 1.12b.

1.2.4 Data transformation

Before the pre-processed data is fed into the models for training, we need to split the data into three clusters, which are training (60%), validation (20%), and testing dataset (20%). Among each training dataset, the data will be further split into input variables $\boldsymbol{X}$ and output variable $\boldsymbol{Y}$ (i.e., training X/training Y, testing X/testing Y). During the



(a) Coefficienct between blending ratio and colour levles.



(b) Trend comparison of ammonia concentration and colour levels.

Figure 1.13: Observatios of ammonia concentration and colour levels in SHWEPP influent.

training process, machine learning algorithms will learn a target function $\boldsymbol{f}$ to best map $\boldsymbol{X}$ to $\boldsymbol{Y}$.

A taining dataset is a set of examples (e.g., historical data) for models to learn the hidden trends and information in the data, shown in (a) in Fig. 1.7, and the training loss is calculated by taking the sum of loss for each example in the training dataset after each epoch. Since it is impossible to have the optimized hyperparameters in the first try of the training, a validation dataset as in (b) in Fig. 1.7 is used to assess the model performance until we obtain the optimzed settings. The validation loss plays an important role during the model training, the adjustments of the hyperparameters will directly reflect on the change of the validation loss, the lower the values, the better the model performance is. As the optimzed model is obtained, testing dataset is used to evaluate the performance of the forecasting model, as shown in (c) in Fig. 1.7. To the forecasting Models, testing dataset has never been seen by the models. If the model tuning process was performed on the testing dataset, the model performance would be a biased result since the hyperparameters are revised in favor to the evalution of the testing dataset.

In Fig. 1.7, the hyperparameters will remained the same once the optimzed values are found, thus generating a baseline model perfromance of a specific machine learning model. The baseline results will be further compared with the results from the model trained by the proposed model trianing steps, which include datasets that have been performed data smoothing and feature engineering techniques.

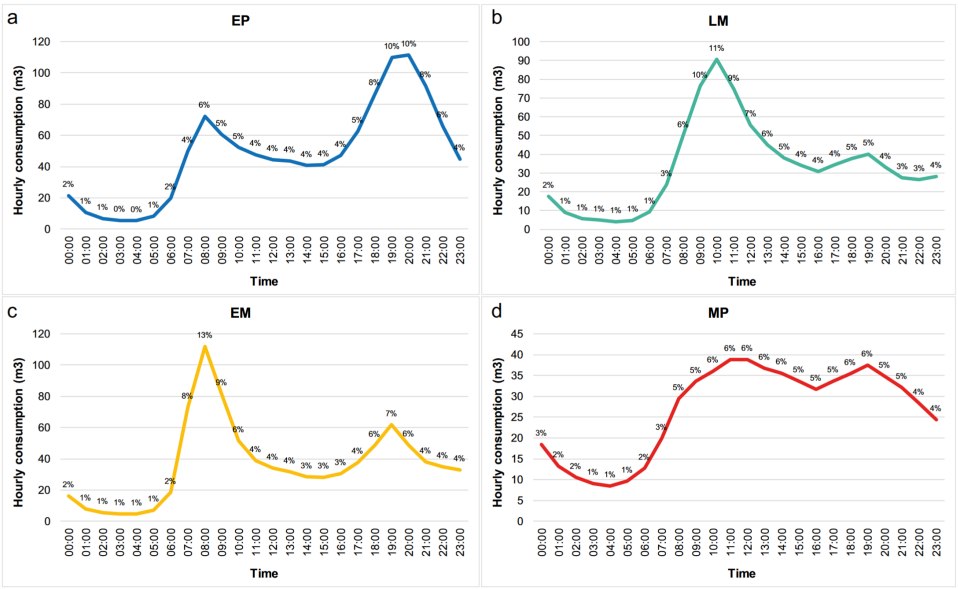


Figure 1.14: Hourly water consumption patterns in households (Abu-Bakar et al., 2021). (a) Cumulative pattern and percentage of hourly consumption for households in the "Evening Peak (EP)" cluster (b) Cumulative pattern and percentage of hourly consumption for households in the "Late Morning Peak Peak (LM)" cluster. (c) Cumulative pattern and percentage of hourly consumption for households in the "Early Morning Peak (EM)" cluster. (d) Cummulative pattern and percentage of hourly consumption for households in the "Multiple Peak (MP)" cluster. Consumption is in (m3).

1.2.5 Feature selection

Fig. 1.17 illustrates which features are selected during the model training processes. In baseline model trianing steps, for both ammonia and colour forecasting model, only one feature is used for training for each model, which is ammonia and colour data, respectively. The model trained by a single feature, followed the baseline model training steps, will generate baseline models. The results from the final evaluation will be defined as the baseline model performance, which will be compared with the model evaluated results from proposed model training steps. Once the baseline model performance is obtained, more features will be input to the model training processes in the order of 2 inputs, 3 inputs, and 4 inputs.

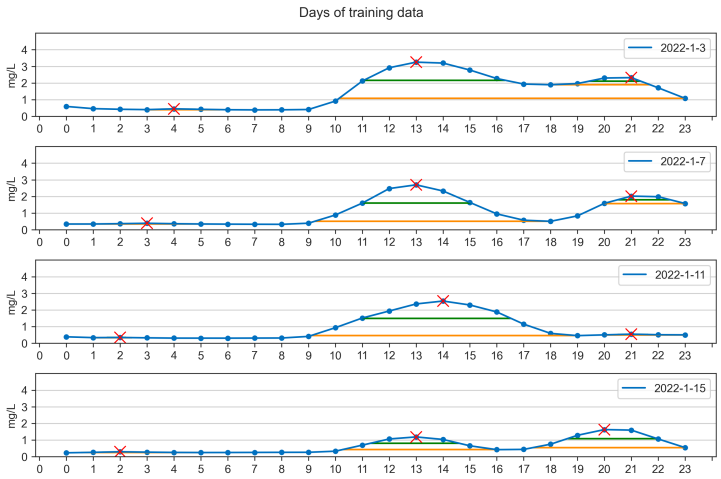


Figure 1.15: The daily patterns of ammonia concentration on 3, 7, 11, 15 Janurary 2022.

1.3 Machine learning models

1.3.1 Random Forest

The machine learning model used in this study (i.e., not deep learning models) is random forest (RF). It is an ensemble method which the final output is obtained by averaging the results from multiple tree learners (Wang et al., 2021), as shown in Fig. 1.18a. The training algorithm applies the general technique of bootstrap aggregating, also known as bagging, to tree learners. Given a training set $X = x_1, ..., x_n$ with targets $Y = y_1, ..., y_n$, bagging repeatedly (B times) selects a random sample with replacement (i.e., not putting the samples back to the population) of the training set and fits trees to these samples (Wikipedia, 2022a), RF generate an output through the following steps:

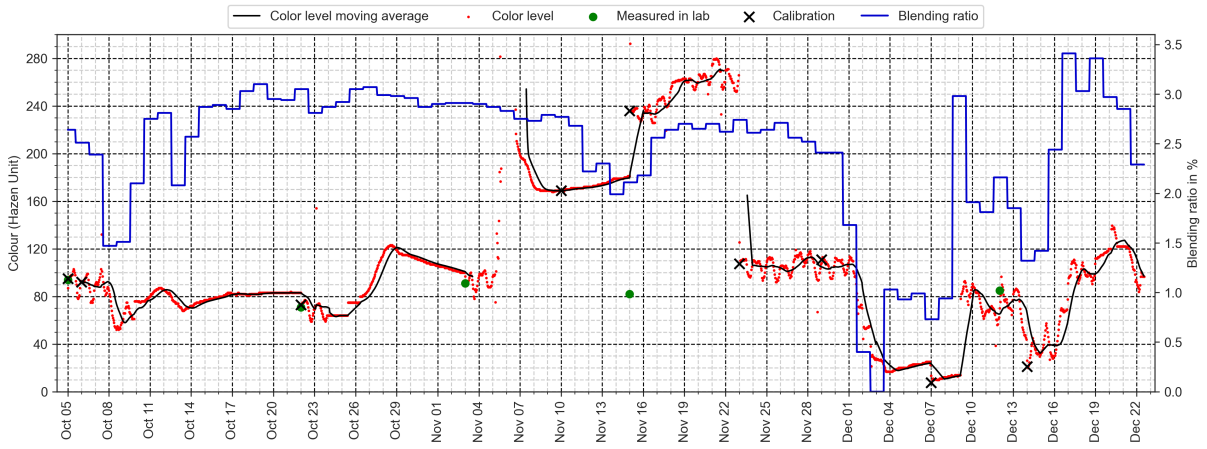
For $b = 1, ..., B$:

1) Sample (with replacement) n training examples from X, Y, call these X_b, Y_b.

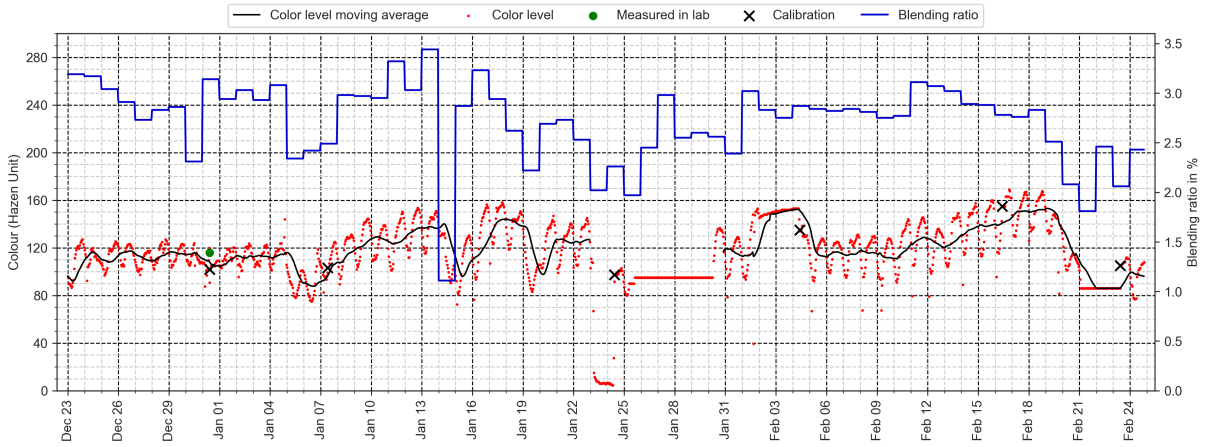
2) Train a regression tree f_b on X_b, Y_b.

3) Predict unseen samples $x^{'}$ by averaging the predicdtions from all the regression tree learners on $x^{'}$ as in Eq. 1.3.1:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x^{'}) \tag{1.3.1}$$



(a) Data collected from 5 October 2021 to 22 December 2021.



(b) Data collected from 23 December 2021 to 24 February 2022.

Figure 1.16: Monitored colour level in MBR effluent and the change of blending ratio (v/v) of treated leachate effluent to municipal wastewater in the inflow of SWHEPP during December 2021–January 2022. Date of manually calibration and colour level measured in laboratory is also provided as black cross and green dot. The moving average of colour level is calculated by averaging the colour level in the past 24 hours. Note: The colour levels analysed by the on-line colour monitoring system were compared to the manually measured data obtained from the laboratory, which showed errors of 2.08%, 4.05%, 1.11%, 65.25%, 4.94% and 11.0% in the TSE samples collected 5 Oct, 22 Oct, 3 Nov, 15 Nov, 12 Dec, and 31 Dec 2021, respectively.

1.3.2 Deep Neural Networks

Artificial Neural Network (ANN) is a very broad term that encompasses any form of Deep Learning model. A typical ANN consists with input, hidden and output layers, and each layer comprises multiple neurons (i.e., nodes). The connected neurons are to simulate the human brain by process and transmit input signals to the next nodes (Mohseni-Dargah et al., 2022) . What sets apart from a ANN model to a DNN model is that the former

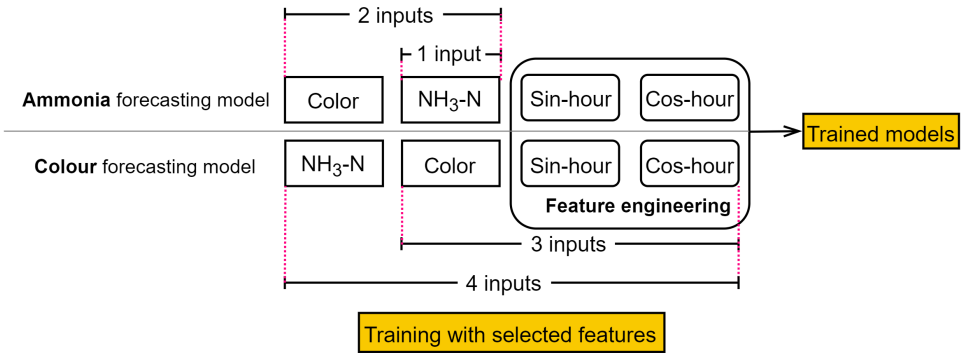
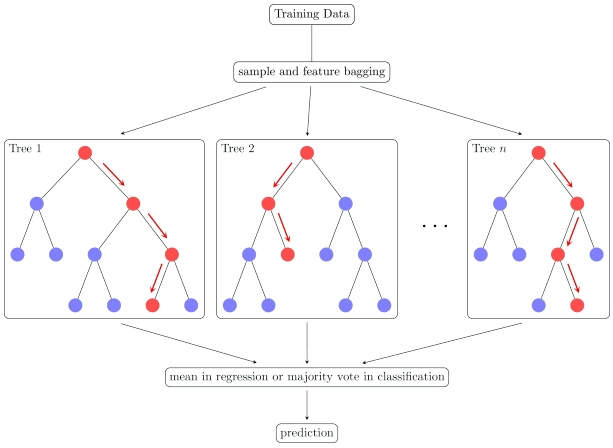
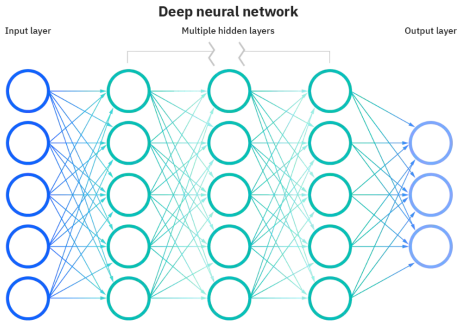


Figure 1.17: Illustration of feature selections for model training.

contains only one hidden layer while the latter has more than one, as shown in Fig. 1.18b. The DNN models are nonlinear, which finds the correct mathematical manipulation to turn the input into the output (Bangaloreai, 2018).



(a) Random Forest (RF) (Riebesell, 2022).



(b) Deep Neural Network (DNN) (IBM, 2022).

Figure 1.18: Illustration of RF and DNN model structure.

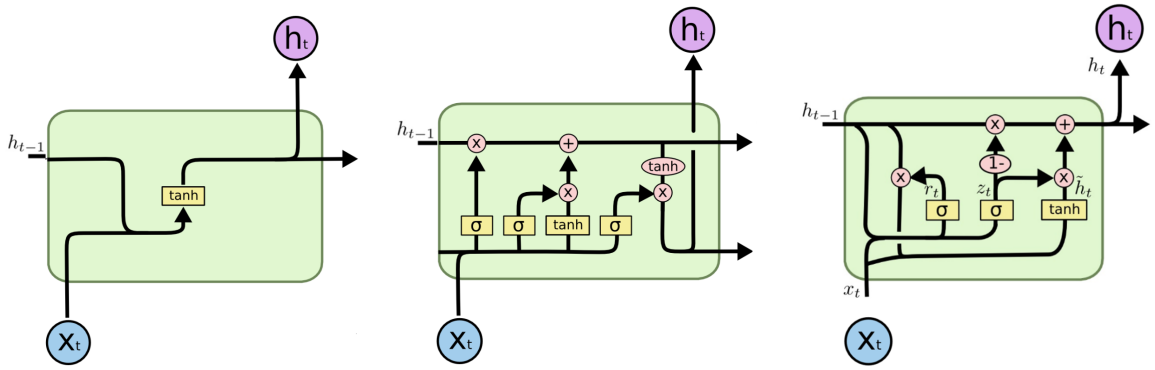
1.3.3 Recurrent Neural Network

A recurrent neural network (RNN) is a type of Artificial Neural Network which designed to work with sequence data. For instance, sequence data are time series, DNA, language, speech and sequences of user actions data, etc. The ammonia concentration and colour level data are time series data, which is a series of data points listed in minute orders (Donges, 2021). A distinguished characteristic of RNN is that they share parameters across each layer of the network by allowing information to be passed from last step of the network to the next. Unlike RNN, feedforward networks like DNN have different weights across each node. The reuse of previous information for making decision

on RNN makes it capable of "learning" from the previous inputs. The realization of the memorizing function is through a memory unit called hidden state (i.e., a vector contains weights) in RNN architecture, which enable RNN to presist data, thus capture short term dependencies. The RNN architecture is presented in Fig. 1.19a. The general formulation of a RNN is expressed in Eq. 1.3.2 (Mamandipoor et al., 2020):

$$h_t = \sigma(W^h h_{t-1} + W^x x_t + b) \tag{1.3.2}$$

where x_t is the current input, h_t is the current hidden state (output), h_{t-1} is the previous output, W^x is the weights of the hidden state, W^h is the weight of the input, b is the bias, σ is the sigmoid activation function.



(a) Recurrent Neural Network (RNN).

(b) Long Short-term Memory (LSTM).

(c) Gate Recurrent Unit (GRU).

Figure 1.19: Variant architectures of Recurrent Neural Networks (adapted from Olah (2015)). x_t corresponds to the current input, h_{t-1} to the last hidden state (output), h_t to the current output, tanh is the tangent activation function, σ is the sigmoid activation function,$\times$ is the vector pointwise multiplication, $+$ is the vector pointwise addition.

1.3.4 Long Short-term Memory

Long Short-term Memory (LSTM) is a deep recurrent neural networks (RNN), an advanced and improved version of RNN. The advent of LSTM is to solve problems requiring learning long-term temporal dependencies which cannot be learned by RNN due to the simple model architecture. The fundmental of LSTM network is built on memory blocks called "cells", which are responsible for tranferring and receiving the states (i.e., vectors) recording the information from the previous cells. In a cell block, there are input gate, forget gate and the output gate. The function of these three gates is to control

the movement of the information into and out of the cell via the sigmoid function. The inputs of the cell will first go through a forget gate (f_t) as Eq. 1.3.3a, where the function will mutiply each element in the input states by values ranging from 0 to 1 to realize the effect of "forget". Next, a input gate (i_t) as in Eq. 1.3.3b will decide whether the new information should be updated or ignored by sigmoid function (i.e., 0 or 1), followed by a tangent function giving weight of importance (i.e., -1 to 1) to the values which passed by as in Eq. 1.3.3c. New memory then is appended to the previous memory C_{t-1} resulting a new C_t. Lastly, output values (h_t) is obtained based on output cell state (O_t) as in Eq. 1.3.3e and Eq. 1.3.3f (Le et al., 2019). The equations for LSTM structure are shown in Eq. 1.3.3:

$$f_t = \sigma(W_f[h_{t-1}, X_t] + b_f) \tag{1.3.3a}$$

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \tag{1.3.3b}$$

$$\tilde{C}_t = tanh(W_n[h_{t-1}, X_t] + b_n) \tag{1.3.3c}$$

$$C_t = C_{t-1} f_t + \tilde{C}_t i_t \tag{1.3.3d}$$

$$O_t = \sigma(W_o[h_{t-1}, X_t] + b_o) \tag{1.3.3e}$$

$$h_t = O_t tanh(C_t) \tag{1.3.3f}$$

where f_t corresponds to the forget gate, i_t to the input gate, $\tilde{C}_t$ to the candidate cell state, C_t to the current cell state, O_t to the output cell state, h_t to the output values, σ to the sigmoid function, X_t to the current input, tanh to the tangent function, W and b are the weight matrices and bias of the corresponding output gate, respectively.

1.3.5 Gate Recurrent Unit

Gated Recurrent Unit (GRU) model is a variant of LSTM model, by combining the forget gate and input gate into an update gate as in Fig. 1.19c, GRU has less parameters compared to LSTM. The advantage of GRU over LSTM is less computing power required while maintaining a similar model performance compared to LSTM. The inputs of GRU model first enter the update gate (z_t) as in Eq. 1.3.4a, where the function will help the model to determine how much of the past information needs to be passed along to the future via sigmoid functions. Followed by the reset gate (r_t) as in Eq. 1.3.4b, which is used to decide how much of the past information to forget. Althogh Eq. 1.3.4a and Eq. 1.3.4b

have the same inputs of X_t and h_{t-1}, the usages of the gates are different. The outputs of reset gate will be used to determine the candidate hidden state (]$tildeh$) as in Eq. 1.3.4c, where the tangent fucntion will determine the importance of current input (X_t), reset gate output, and previous hidden state (h_t). At the last step, the output values (h_t) is calculated from the candidate hidden state ($\tilde{h_t}$), previous hidden state (h_{t-1}), and the outputs of update gate as in Eq. 1.3.4d. The equations of GRU structures are presented in Eq. 1.3.4 (Cheng et al., 2020):

$$z_t = \sigma(X_t W_{xz} + h_{t-1} W_{hz} + b_z) \tag{1.3.4a}$$

$$r_t = \sigma(X_t W_{xr} + h_{t-1} W_{hr} + b_r) \tag{1.3.4b}$$

$$\tilde{h_t} = tanh(X_t W_{xh} + (r_t \circ h_{t-1}) W_{hh} + b_h) \tag{1.3.4c}$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h_t} \tag{1.3.4d}$$

where z_t corresponds to the update gate, r_t to the reset gate, $\tilde{h_t}$ to the candidate hidden state, h_t to the output values, σ to the sigmoid function, tanh to the tangent function, X_t to the current input, W and the b are the weight matrices and bias of the corresponding output gate, respectively.

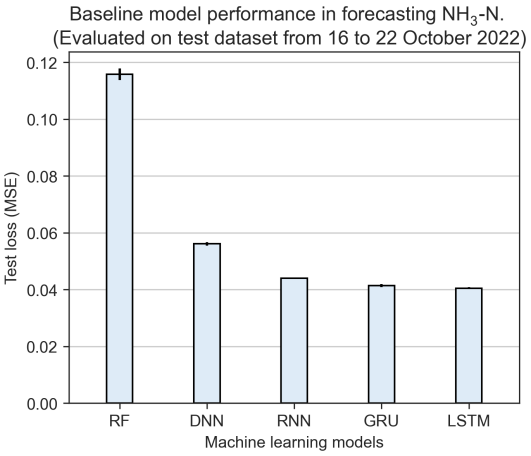
CHAPTER 2

RESULTS AND DISCUSSION

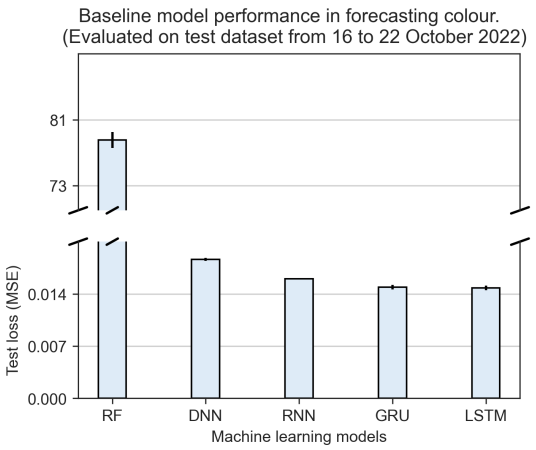
2.1 Baseline performance of the forecasting models

We first review the ammonia and colour forecasting models trained by single-featured datasets. The performance of RF models in Fig. 2.1a and Fig. 2.1b showed poorer performance compared to the other four deep learning models, while the LSTM models showed the lowest values of test loss.

The significant higher test loss of RF models compared to other models can be visualized by plotting the forecasted values with the ground truths (i.e., observed values). In Fig. 2.2, one-step-ahead forecast horizon of ammonia concentration and colour level is plotted by RF as in Fig. 2.2a and Fig. 2.2c and LSTM models as in Fig. 2.2b and Fig. 2.2d. It's easier to observe that the RF models are less capable of predicting the water quality paratmers.

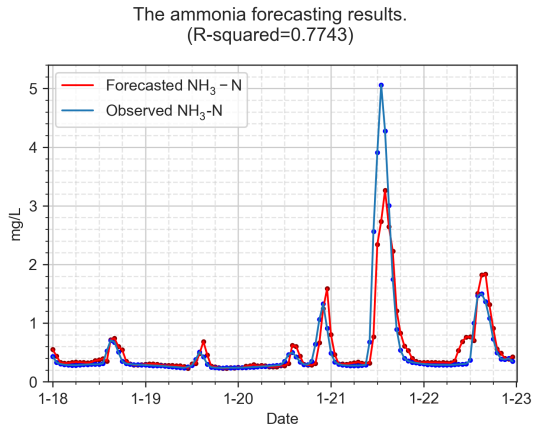


(a) Test loss values from five ammonia forecasting models.

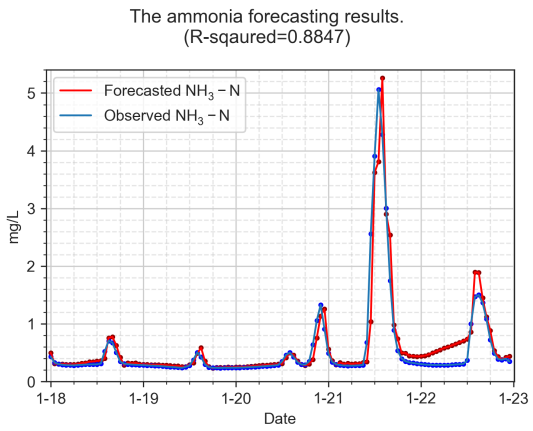


(b) Test loss values from five colour forecasting models.

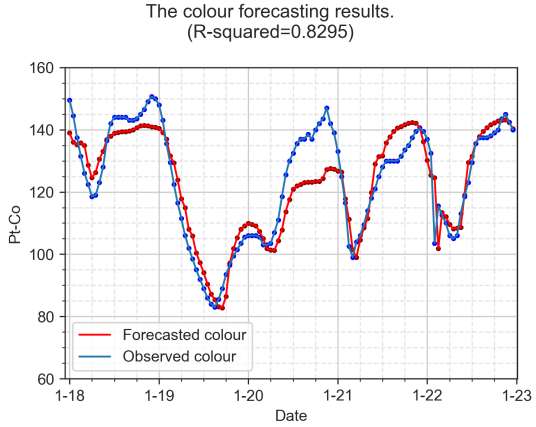
Figure 2.1: Baseline performance of ammonia and colour forecasting models.



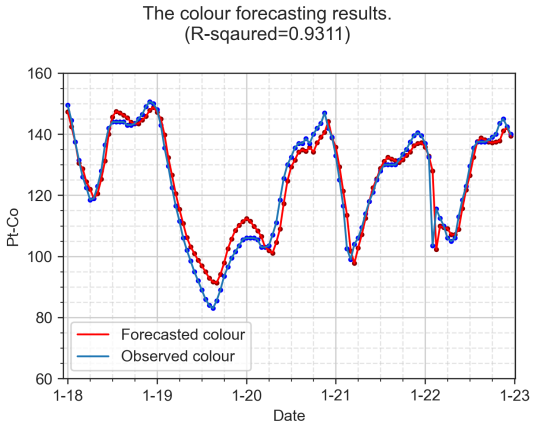
(a) Baseline RF model forecasting ammonia concentration.



(b) Baseline LSTM model forecasting ammonia concentration.



(c) Baseline RF model forecasting colour levels.



(d) Baseline LSTM model forecasting colour levels.

Figure 2.2: Visulization of the model forecasting results.

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

2.2.1 Ammonia forecasting models

With the baseline model performance is known, we investigated the influence of pre-processed datasets on the ammonia forecasting models. In Table. 2.1, we listed all the test loss of models trained with each proposed pre-processed methods. The models trained by SG filters at different window size are denoted as model-sg5, model-sg7, model-sg9; the naming rule applies the same to EWMA filters, and for the method of outlier removal for ammonia data is denoted as model-or.

We found that SG filters improved most in the quality of the raw dataset, as the

top lowest test loss values are from GRU-sg7 and GRU-sg9 (i.e., at the window size of 7 and 9, respectively), followed by LSTM-ew3. However, the improvements of model performance resulted from the use of data smoothing methods are not consistant across different models. In other words, the best training datasets for GRU, LSTM, and DNN models are different.

Table 2.1: Baseline performance of ammonia forecasting model, evaluated on test dataset from **16 to 22 Janurary 2022**. Loss values are calculated by MSE.

Model-Dataset	Test loss	Valid loss	Model-Dataset	Test loss	Valid loss
GRU-sg7	0.0383	1.2508	RNN-or	0.0432	1.6345
GRU-sg5	0.0385	1.2644	RNN-ew3	0.0434	1.6041
LSTM-ew3	0.0388	1.0796	RNN-obs	0.0440	1.6734
LSTM-sg5	0.0388	1.2346	RNN-sg9	0.0442	1.7046
LSTM-sg7	0.0388	1.1804	DNN-obs	0.0561	3.2383
GRU-ew2	0.0389	1.1891	DNN-sg5	0.0562	3.2170
GRU-ew4	0.0391	1.2390	DNN-ew2	0.0563	3.1677
GRU-ew3	0.0392	1.2199	DNN-ew3	0.0569	3.2317
LSTM-ew2	0.0392	1.0969	DNN-sg7	0.0570	3.2014
LSTM-ew4	0.0395	1.1219	DNN-ew4	0.0571	3.2188
GRU-sg9	0.0396	1.3097	DNN-or	0.0572	3.1972
LSTM-or	0.0398	1.2612	DNN-sg9	0.0574	3.2484
LSTM-obs	0.0405	1.3993	RF-obs	0.1158	-
GRU-or	0.0405	1.2366	RF-sg9	0.1196	-
LSTM-sg9	0.0410	1.3076	RF-ew2	0.1286	-
GRU-obs	0.0414	1.3638	RF-or	0.1294	-
RNN-sg5	0.0415	1.5088	RF-sg5	0.1298	-
RNN-ew2	0.0421	1.5425	RF-ew3	0.1313	-
RNN-sg7	0.0423	1.6267	RF-sg7	0.1409	-
RNN-ew4	0.0432	1.5992	RF-ew4	0.1441	-

Empirically, when different models are evaluated by the same testing dataset, the best Model-Dataset combination shold have both the lowest values of test and validation loss. For instance, GRU-sg7 model in forecasting ammonia has the lowest test loss of 0.0383, yet the validation loss of 1.2508 only ranks the tenth from the smallest validation loss vlaues. The models with top three lowest values of the validation loss are LSTM-ew3, LSTM-ew2 and LSTM-ew4. This finding points to the potential of the heterogeneity between the trianing and testing datasets. This hypothesis was the explanation with the highest likelihood when no overfitting was observed in the training datasets. Further tests were carried out using testing dataset from October to examine how the Model-Dataset ranks of test and validation loss values will change into. To the best of my understanding,

the comparisons between testing and validation loss are not discussed on the currently available research papers in modelling of wastewater treatment industry.

As shown in Table. 2.2, the top three ranks of Model-Dataset in the lowest validation loss is the same to the top three ranks in the test loss values. This is in good agreement with how the heterogeneity of the datasets can impact on the model performance. The evluations of the ammonia forecasting models in October 2021 showed a complete different outcomes compared to the one in January 2022. Surpirsingly, the top three ranks of Model-Dataset in the lowest validation loss are the same of the lowest test loss, which are 0.0158 from LSTM-ew3, 0.0161 from LSTM-ew2, and 0.0163 from LSTM-ew4. Instead of GRU, LSTM becomes the best model for training ammonia forecasting model. The most remarkable results in Table. 2.2 is that EWMA filter seems to be the most ideal pre-processing methods for traning deep learning models as LSTM-ew3, GRU-ew3, RNN-ew4 and DNN-ew3 models showed the best model performance in test loss compared to the same models trained by other data pre-processing methods.

Table 2.2: Baseline performance of ammonia forecasting model, evaluated on test dataset from **10 to 16 October 2021**. Loss values are calculated by MSE.

Model-Dataset	Test loss	Valid loss	Model-Dataset	Test loss	Valid loss
LSTM-ew3	0.0158	1.0796	RNN-or	0.0197	1.6345
LSTM-ew2	0.0161	1.0969	RNN-sg7	0.0201	1.6267
LSTM-ew4	0.0163	1.1219	RNN-sg9	0.0205	1.7046
LSTM-sg5	0.0166	1.2346	RNN-obs	0.0206	1.6734
GRU-ew3	0.0167	1.2199	DNN-ew3	0.0316	3.2317
GRU-ew4	0.0169	1.2390	DNN-or	0.0316	3.1972
GRU-ew2	0.0170	1.1891	DNN-sg7	0.0316	3.2014
GRU-sg9	0.0174	1.3097	DNN-ew2	0.0318	3.1677
LSTM-obs	0.0175	1.2366	DNN-ew4	0.0319	3.2188
LSTM-or	0.0177	1.2612	DNN-obs	0.0319	3.2383
GRU-sg5	0.0178	1.2644	DNN-sg5	0.0319	3.2170
GRU-sg7	0.0180	1.2508	DNN-sg9	0.0319	3.2484
LSTM-sg7	0.0180	1.1804	RF-sg9	0.1307	-
GRU-or	0.0187	1.3993	RF-sg7	0.1311	-
LSTM-sg9	0.0188	1.3076	RF-sg5	0.1343	-
GRU-obs	0.0189	1.3638	RF-ew2	0.1346	-
RNN-ew4	0.0190	1.5992	RF-ew3	0.1368	-
RNN-ew2	0.0191	1.5425	RF-obs	0.1443	-
RNN-ew3	0.0193	1.6041	RF-ew4	0.1451	-
RNN-sg5	0.0195	1.5088	RF-or	0.1477	-

2.2.2 Colour forecasting models

The best performed colour forecasting models are the LSTM models trained by EWMA filters, which are 0.0136 from LSTM-ew4, 0.0138 from LSTM-ew2 and LSTM-ew3 as shown in Fig. 2.3. Interestingly, LSTM models trained by EWMA also showed the best performance in ammonia forecasting models. The top three ranks of Model-Dataset in the lowest validation loss ranks the 6th, 20th, and 1st from the lowest test loss values. However, we don't have extra testing datasets for re-evaluating the colour forecasting models. Compromises have to be made during the analysis of colour forecasting models.

Table 2.3: Baseline performance of colour forecasting model, evaluated on test dataset from **16 to 22 Janurary 2022**. Loss values are calculated by MSE.

Model-Dataset	Test loss	Valid loss	Model-Dataset	Test loss	Valid loss
LSTM-ew4	0.0136	0.7515	RNN-obs	0.0160	1.0623
LSTM-ew2	0.0138	0.8011	LSTM-sg7	0.0161	0.7439
LSTM-ew3	0.0138	0.7547	LSTM-sg5	0.0168	0.8355
GRU-ew3	0.0140	0.8068	DNN-sg5	0.0180	1.4702
GRU-ew2	0.0142	0.8330	DNN-sg7	0.0180	1.4823
GRU-ew4	0.0143	0.7694	DNN-sg9	0.0180	1.4574
LSTM-sg9	0.0143	0.7137	DNN-ew4	0.0181	1.4632
RNN-ew3	0.0144	0.8492	DNN-ew3	0.0182	1.4716
RNN-ew4	0.0147	0.8476	DNN-ew2	0.0183	1.4946
RNN-sg9	0.0147	0.8363	DNN-obs	0.0186	1.5397
LSTM-obs	0.0148	0.9744	RF-sg9	63.6847	
GRU-obs	0.0149	0.9927	RF-sg7	73.8263	
RNN-ew2	0.0150	0.9083	RF-ew3	75.1974	-
GRU-sg9	0.0151	0.7575	RF-ew4	77.8829	-
RNN-sg5	0.0158	0.8846	RF-obs	78.5296	-
RNN-sg7	0.0158	0.8755	RF-ew2	78.8753	-
GRU-sg7	0.0159	0.7791	RF-sg5	81.0696	-
GRU-sg5	0.0160	0.8080	-	-	-

By comparing the baseline performance and the influences of data pre-processing methods on machine learning models, our findings appear to be well substantiated the use of LSTM models for training ammonia and colour forecasting models due to it's outstanding model performance evaluated by test loss values. Although EWMA filters showed surprising effects on improving the performance of most models, the influence of pre-processing methods are still not consistant across different models and training datasets. Thus, in the testings of the proposed model training processes will include all the pre-processing

methods for model training, and LSTM will be used as the only machine learning model.

2.3 Exploit hidden patterns in MBR effluent water quality to enhance model performance

2.3.1 Ammonia forecasting models

In the section of feature engineering, we have introduced the selection and creation of the extra input features for training forecasting models. In Fig. 2.4, the performance of ammonia forecasting models trained by 2 to 4 input features are compared with the baseline performance to demonstrate how the feature engineered features influenced on the model outputs. Notice that due to colour data are not available from 10 to 16 October 2021, the models in Fig. 2.4 were evaluated on training dataset from 16 to 22 January 2022.

Leaving out the potential influences of heterogeneity between training and testing datasets on comparing the model performance, interesting resutls were still observed. For LSTM models trained by non pre-processed datasets, more input features deteriorate the model performance as LSTM-4-obs has the highest test loss and LSTM-1-obs has the lowest test loss. Based on our understanding to the extra features such as color level and positional encodings, the test loss of the forecasting models should be lower and inversely proportional to the increasing number of input features. The model performance from LSTM-sg7 and LSTM-sg9 fit well with what we hypothesized. The highest test loss values was observed in LSTM-1 models, followed by LSTM-2, LSTM-3 and LSTM-4.

LSTM-2 models show intriguing results. Except for LSTM-obs, all the LSTM-2 models showed lower test loss compared to LSTM-1. This finding lead us to believe that the flucturation of ammonia concentration is highly correlated with the colour level in SHWEPP influent.

Comparing to the baseline model performance using the LSTM-sg7 approach, the test loss values of LSTM-1-sg7, LSTM-2-sg7, LSTM-3-sg7 and LSTM-4-sg7 decreased by 4.2%, 6.4%, 7.9%, and 8.9%, respectively comapred to the baseline model test loss.

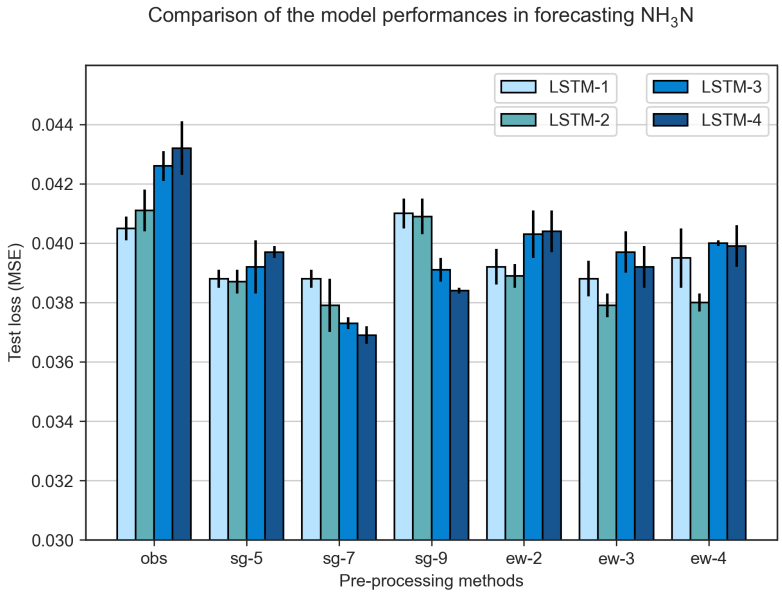


Figure 2.3: cap .

2.3.2 Colour forecasting models

Fig. 2.4 presents how the various pre-processing methods affect the performance of colour forecasting models evaluated on testing dataset from 16 to 22 January 2022. Unexpectedly, what we found in the results of colour forecasting models have many commons to the findings in Fig. 2.3, which are decribed as the followings:

1) In LSTM-obs, models trained with more inputs resulted in poorer model performance, except for LSTM-obs, LSTM-sg9 and LSTM-ew2.
2) In the results of LSTM-sg7, the test values decreased with the increased number of model inputs, which satified the hypothesis we claimed in previous section.
3) The test loss values of LSTM-2 in all the pre-processed datasets are lower than LSTM-1 except for LSTM-obs, LSTM-sg9 and LSTM-ew2.

Looking at the color forecasting model with the lowest test loss values of 0.0121, LSTM-3-sg9 successfully improved the model performance, by lowering test loss by 28.6% compared to the baseline model test loss.

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

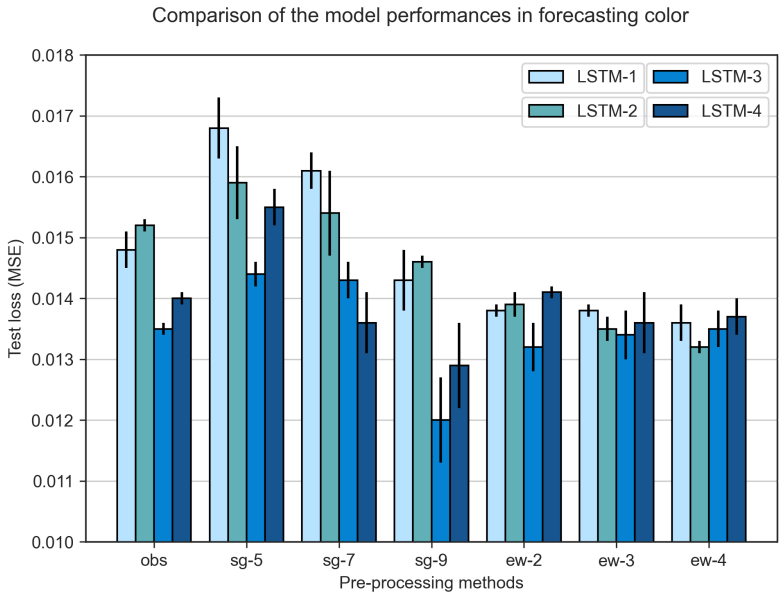


Figure 2.4: cap .

CHAPTER 3

CONCLUSION

The above findings suggest the proposed model training processes are reliable even with the models are trained to forecast different water quality parameters. Although to the current available datasets we cannot explain why only certain datasets follow the hypothesis of

Bibliography

Halidu Abu-Bakar, Leon Williams, and Stephen H. Hallett. Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England. *npj Clean Water*, 4(1):13, December 2021. ISSN 2059-7037. doi: 10.1038/s41545-021-00103-8.

Bangaloreai. Deep neural network (DNN) is an artificial neural network (ANN), March 2018.

CFI. Exponentially Weighted Moving Average (EWMA), January 2022.

Tuoyuan Cheng, Fouzi Harrou, Farid Kadri, Ying Sun, and Torove Leiknes. Forecasting of wastewater treatment plant key features using deep learning-based models: A case study. *IEEE Access*, 8:184475–184485, 2020. doi: 10.1109/ACCESS.2020.3030820.

DeepAI. Loss Function, June 2022.

Niklas Donges. A Guide to RNN: Understanding Recurrent Neural Networks and LSTM Networks, July 2021.

IBM. Neural Networks, June 2022.

Le, Ho, Lee, and Jung. Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. *Water*, 11(7):1387, July 2019. ISSN 2073-4441. doi: 10.3390/w11071387.

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.

MathWorks. Documentation-Findpeaks, June 2022.

Masoud Mohseni-Dargah, Zahra Falahati, Bahareh Dabirmanesh, Parisa Nasrollahi, and Khosro Khajeh. Chapter 12 - Machine learning in surface plasmon resonance for environmental monitoring. In Mohsen Asadnia, Amir Razmjou, and Amin Beheshti, editors,

Artificial Intelligence and Data Science in Environmental Sensing, Cognitive Data Science in Sustainable Computing, pages 269–298. Academic Press, January 2022. ISBN 978-0-323-90508-4. doi: 10.1016/B978-0-323-90508-4.00012-5.

National Center for Biotechnology Information. "PubChem Compound Summary for CID 222, Ammonia" PubChem, June 2022.

Christopher Olah. Understanding LSTM Networks, August 2015.

Janosh Riebesell. Random Forest, June 2022.

Abraham. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, July 1964. ISSN 0003-2700. doi: 10.1021/ac60214a047.

Cees Taal. Smoothing your data with polynomial fitting: A signal processing perspective, April 2017.

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.

Wikipedia. Random forest, June 2022a.

Wikipedia. Savitzky–Golay filter, June 2022b.