University of Bahrain

College of Science

Department of
Postgraduate Programs

Predicting the Nitrification Performance in Muharraq Domestic Wastewater Treatment Plant Using Machine Learning Techniques

Thesis submitted in partial fulfilment of the requirements for the degree of M.Sc. In Big Data Science and Analytics

Prepared by:

Salman Yusuf Hammad

20167572

Supervised by:

Dr. Majeed Safar Jasim

Associate Professor

University of Bahrain

Kingdom of Bahrain

March 2023

University of Bahrain

College of Science

Department of Postgraduate Programs



Predicting the Nitrification Performance in Muharraq Domestic Wastewater Treatment Plant Using Machine Learning Techniques

Thesis submitted in partial fulfilment of the requirements for the degree of M.Sc. In Big Data Science and Analytics

Prepared by:

Salman Yusuf Hammad

20167572

Supervised by:

Dr. Majeed Safar Jasim

Associate Professor

University of Bahrain

Kingdom of Bahrain

March 2023

Defense Committee's Approval

University of Bahrain
Deanship of Graduate Studies
And scientific Research

جامعة البحرين
عمادة الدراسات العليا والبحث العلمي

Thesis Discussion Committee Decision

Student Name: Salman Yusuf Hammad	Student ID: 20167572
Department: Choose an item.	College: **Science**
Program: **M.Sc. in Big Data Science and Analytics**	
Thesis Title: Predicting the Nitrification Performance in Muharraq Domestic Wastewater Treatment Plant Using Machine Learning Techniques	
Supervisor Name: **Dr. Majeed Safar**	Student GPA: **3.95**

The Discussion Committee met on 24-May-23 In S41-1002. Upon discussing the student whose details are stated above in public, the discussion committee recommends the following:

The student is awarded a grade Choose an item. in the thesis subject.

☐ Without making any changes to the thesis.

☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

Name		Signature
1. **Dr. Majeed Safar**	(Advisor)	. Majeed
2. **Prof. Nabil Hewahi**	(Internal Examiner)	. Nabil Hewahi
3. **Dr. Saeed Sharif**	(External Examiner)	. SaeedSharif

Approved at the University Council 6th meeting, 2013
In accordance with Decision No. 736/2013 dated 14/3/2013
Modified by University Council Decision No. 2043/2013 dated 20/11/2013.

Abstract

Machine Learning (ML) is rapidly becoming the state-of-the-art methodology for modelling chemical and biological processes. A Wastewater Treatment Plant (WWTP) removes contaminants from sewage, producing effluent that satisfies environmental standards prior to discharging into water bodies. The purpose of the nitrification step in the WWTP is to remove nitrogen using nitrifying bacteria, as it can be harmful to aquatic organisms in the released bodies of water. This study utilizes K Nearest Neighbour (KNN), Multilayer Perceptrons (MLPs), Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting Machine (GBM) as supervised ML models to predict nitrification performance in treated sewage effluent in Muharraq WWTP in Bahrain. The developed models incorporates ten influent variables to predict the concentration of effluent Ammonium Nitrogen (NH_4-N), Nitrite Nitrogen (NO_2-N), Total Kjeldahl Nitrogen (TKN), and Total Nitrogen (TN). The evaluation metrics used to assess the models performance are: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). The results showed that SVR model was the most effective in predicting NH_4-N with MAE value of 0.2759, while the RF model outperformed others in predicting NO_2-N, TKN, and TN with MAE values of 0.0878, 0.3812, and 0.4583, respectively. The Steady State Design Temperature (SSDT) is identified as the most significant variable influencing, the effluent NH_4-N and TN. This work highlights importance of using ML models to evaluate WWTP performance in order to assist meeting Bahrain's irrigation standards and protecting the local ecosystem.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Majeed Safar for his invaluable guidance and support to successfully complete my thesis. I would also like to express my gratitude to Dr. Gulnur Coskuner for her guidance and encouragement during this research. Additionally, they were always willing to offer their time and knowledge, answering my questions and providing valuable insights. I also would like to thank Dr. Sawsan Hilal for facilitating the required procedures and for all the assistance I received throughout the program.

I would like to express my sincere appreciation to Tatweer Petroleum Company, represented by Abdulla Althawadi, Manager-Field Operations, and Khaled Hassan, Head of Process and Technical Services, for their invaluable support. Their constant encouragement and willingness to make my training plan flexible during my academic journey have been instrumental in my success.

I am extremely grateful to Mr. Peter Naylor, General Manager, and Mr. Binu Mathew, Head of Technical Services at Muharraq WWTP, for providing the data and giving me a tour to understand the plant's process, as well as helping me whenever I faced a problem in the project.

Finally, I would like to express my sincere gratitude to my family and friends for their unwavering support throughout my entire master's thesis preparation. I am forever grateful for their assistance. Thank you.

Table of Contents

Defense Committee's Approval ..... A

Abstract ..... B

Acknowledgements C

Table of Contents ..... D

List of Tables ..... G

List of Figures ..... H

List of Abbreviations ..... J

Chapter One Introduction to the Study 2

1.1. Overview 2

1.2. Research Problem 3

1.3. Research Questions 4

1.4. Research Aim and Objectives 4

1.5. Research Significance 5

1.6. Thesis Structure 6

Chapter Two Literature Review 8

2.1. Background Machine Learning Methadologies 8

2.1.1. K-Nearest Neighbors (KNN) 8

2.1.2. Multilayer Perceptrons (MLPs) 10

2.1.3. Support Vector Regression (SVR) 12

2.1.4. Random Forest (RF) 15

2.1.5. Gradient Boosting Machine (GBM) 17

2.1.6. Summary of Different ML Regression Techniques 20

2.1.7. Evaluation Metrics 21

2.2. Artificial Intelligence in Environmental Studies 24

2.3. Nitrification in Wastewater Treatment Plants 25

2.3.1. Nitrifying Bacteria 25

2.3.2. Factors Affecting Nitrification 26
2.4. Local Studies about WWTP 29
2.5. Predicting Wastewater Treatment Parameters Using AI and ML 31
2.5.1. Total Nitrogen (TN) 32
2.5.2. Total Kjeldahl Nitrogen (TKN) 34
2.5.3. Ammonium Nitrogen (NH_4-N) 35
2.6. Research Gaps 41
Chapter Three Methods and Procedures 43
3.1. Research Methodology 43
3.2. Exploratory Data Analysis 44
3.3. Data Pre-Processing 44
3.4. Data Description 48
Chapter Four Results and Discussion 50
4.1. Results 50
4.1.1. Variable Selection 50
4.1.2. Descriptive Statistics 52
4.1.3. Statistical Analysis 53
4.1.4. KNN Model Performance 59
4.1.5. MLPs Model Performance 61
4.1.6. SVR Model Performance 66
4.1.7. RF Model Performance 72
4.1.8. GBM Model Performance 73
4.1.9. Model Comparison 79
4.1.10. Impact of Outlier Removal 81
4.2. Discussion 82
Chapter Five Conclusion and Future Work 88
5.1. Conclusion 88

5.1.1. Research Limitations and Weaknesses 89
5.1.2. Practical Recommendations 90
5.2. Future Work 91
References 92
Appendices 103
Appendix (A): Sample of Influent and Effluent Data of Muharraq WWTP 104
Appendix (B): Boxplots of Influent Parameters 106
Appendix (C): Statistical Analysis 108
Appendix (D): Predicted versus Actual values of the GBM Model 111
الملخص 113

List of Tables

Table (1): Overview of Various Machine Learning Regression Techniques 21
Table (2): Various AI Models and ML Algorithms Used to Predict Influent and Effluent Parameters of WWTP in Similar Studies 37
Table (3): The TSE Range of the Predictors with the Cause of Outliers 47
Table (4): Wastewater Parameters and Definitions 48
Table (5): GVIF before and after Variable Selection 52
Table (6): Statistical Summary of Laboratory Measurements of Influent and Effluent Parameters 52
Table (7): The Effluent Parameters that are out of the TSE Range 59
Table (8): Training and Testing Results of KNN using Optimal Hyperparameters 60
Table (9): Training and Testing Results of MLPs using Optimal Hyperparameters ... 66
Table (10): Comparing SVR Kernels using their Optimal Hyperparameters 71
Table (11): Training and Testing Results of SVR using Optimal Hyperparameters ... 71
Table (12): Training and Testing Results of RF using Optimal Hyperparameters 73
Table (13): Training and Testing Results of GBM using Optimal Hyperparameters .. 76
Table (14): Best Model Results with Optimal Hyperparameters before and after Outliers Removal 82
Table (15): Sample of Influent Laboratory Measurements Data of Muharraq WWTP 104
Table (16): Sample of Effluent Laboratory Measurements Data of Muharraq WWTP 105

List of Figures

Fig. 1: Typical Process Flow Sheet of Biological Nitrogen Removal 3
Fig. 2: Hidden Layers and Neurons of MLPs 11
Fig. 3: SVR Structure 12
Fig. 4: RF Structure 16
Fig. 5: Schematic of GBM Principle 18
Fig. 6: The Number of Publications Classified by their Specific Environmental Areas of Application 24
Fig. 7: The Statistics of Various AI Methods Utilized in the Reviewed Publications 32
Fig. 8: Research Framework 44
Fig. 9: Schematic of Ten Cross-Fold Validation 45
Fig. 10: Boxplots of IF and SS before Dealing with the Outliers. 46
Fig. 11: The Collinearity between the Predictors 51
Fig. 12: Monthly, Seasonal, and Yearly Behavior of Influent and Effluent Temperature 54
Fig. 13: Monthly, Seasonal, and Yearly Behavior of Influent Parameter 56
Fig. 14: Monthly, Seasonal, and Yearly Behavior of Effluent Parameter 58
Fig. 15: The Effect of Different K Values on the Model performance 60
Fig. 16: The Effect of LR on MLPs Performance 62
Fig. 17: The Effect of Momentums on MLPs Performance 62
Fig. 18: The Impact of Adding Hidden Layers on MLP Performance 64
Fig. 19: Effect of Number of Epochs on MLP Convergence 65
Fig. 20: How the Performance of SVR is Impacted by Varying Exponent Values 67
Fig. 21: The Effect of the Different Gamma Values on SVR Performance 69
Fig. 22: The MAE Result of PUK Kernal 70
Fig. 23: The Effect of the Tree Depth on RF Performance 72
Fig. 24: The Effect of Distribution Function on MAE 74
Fig. 25: The Effect of Shrinkage, n.minobsinnode, and n.trees on GBM Performance 75
Fig. 26: The Influence of influent Variables on Each Response 78
Fig. 27: The Evaluation Metrics of the Best Result out of Each Method 80
Fig. 28: Boxplots of IF and SS before and after Dealing with the Outliers 81

Fig. 29: The Boxplots of Different Variables that their Outliers below the TSE Range 106
Fig. 30: The Boxplots of NH_3-N and Alkalinity before and after Removing the Outliers 106
Fig. 31: The Boxplots of COB before and after Removing the Outliers 107
Fig. 32: Monthly, Seasonal, and Yearly Behavior of Influent Variables 108
Fig. 33: Monthly, Seasonal, and Yearly Behavior of Effluent Variables 109
Fig. 34: Monthly, Seasonal, and Yearly Behavior of Removal Efficiency of Different Variables 110
Fig. 35: Predicted vs True Values of the Test Data of NH_4-N 111
Fig. 36: Predicted vs True Values of the Test Data of NO_2-N 111
Fig. 37: Predicted vs True Values of the Test Data of TKN 112
Fig. 38: Predicted vs True Values of the Test Data of TN 112

List of Abbreviations

Abbreviation	Meaning
ACC	Absolute Correlation Coefficient
AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANNs	Artificial Neural Networks
AOA	Ammonia-Oxidizing Archaea
AOB	Ammonia Oxidizing Bacteria
BASR	Biofilm Airlift Suspension Reactor
BOD	Biochemical Oxygen Demand
COB	The Ratio between Chemical Oxygen Demand over Biochemical Oxygen Demand
COD	Chemical Oxygen Demand
DDT	Dynamic Design Temperature
DO	Dissolved Oxygen
FNN	Feed-forward Neural Network
GA-ANN	Genetic Algorithm-Artificial Neural Network
GBM	Gradient Boosting Machine
GTB	Gradient Tree Boosting
GVIF	Generalised Variance Inflation Factor
IBK	Instance-Based Learning with Parameter K
IF	Influent Flow
IQR	Inter Quartile Range
KNN	K Nearest Neighbors
LR	Learning Rate
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCM	Million Cubic Meter
ML	Machine Learning
MLPs	Multilayer Perceptrons
n.minobsinnode	Minimum Number of Observations
n.trees	Number of Trees
NH_3-N	Ammonia Nitrogen
NH_4-N	Ammonium Nitrogen
NO_2-N	Nitrite Nitrogen

NO_3-N	Nitrate Nitrogen
NOB	Nitrite Oxidizing Bacteria
OG	Oil and Grease
PUK	Pearson Universal Kernel
R^2	Coefficient of Determination
RBF	Radial Basis Function
RF	Random Forest
RSME	Root Mean Squared Error
SARIMAX	Seasonal Autoregressive Integrated Moving Average
SIDS	Small Island Developing State
SMO	Sequential Minimal Optimization
SRT	Solids Retention Time
SS	Suspended Solid
SSDT	Steady State Design Temperature
SVR	Support Vector Regression
TDS	Total Dissolved Oxygen
TKN	Total Kjeldahl Nitrogen
TN	Total Nitrogen
TP	Total Phosphorus
TSE	Treated Sewage Effluent
TSS	Total Suspended Solid
WWT	Wastewater Treatment
WWTP	Wastewater Treatment Plant

Chapter One

Introduction to the Study

1.1. Overview

1.2. Research Problem

1.3. Research Questions

1.4. Research Objectives

1.5. Research Significance

1.6 Thesis Structure

Chapter One
Introduction to the Study

1.1. Overview

The Kingdom of Bahrain is categorised as an arid country with a desert climate characterised by hot summers and mild winters. The water resources in Bahrain consist of conventional resources, such as groundwater and non-conventional resources, primarily desalinated water and treated wastewater. Unfortunately, due to overuse and population growth, the groundwater reserves in Bahrain are gradually depleting. Consequently, treated water has become increasingly crucial for the region's water supply (Al-Noaimi, 2005).

Treated water is essential for human health, environmental sustainability, and economic development. Wastewater Treatment (WWT) facilities remove harmful bacteria, viruses, and contaminants from water to prevent the spread of waterborne diseases. Before releasing into the environment helps to minimize negative impacts on natural resources, such as aquatic ecosystems and wildlife. Treating wastewater can increase water availability for industries such as agriculture, manufacturing and energy production (Maier et al., 2004).

WWT procedures are known for their dynamic nature, given the varying influent wastewater composition, concentrations, and flowrates. Thus, there arises a requirement for mathematical models and computer simulations to predict and regulate the treatment processes. With the advancement of computer power, ML models have become a viable option for simulating WWT processes, owing to their accuracy and ease

of use. These models can potentially aid in optimizing treatment efficiency, identifying and resolving operational issues, and improving overall plant performance (Bekkari & Zeddouri, 2019).

1.2. Research Problem

Nitrification plays a crucial role in WWT plants by removing ammonia, which can harm aquatic life. Nitrification is a biological process that occurs in two steps, where ammonia is converted to nitrate. Initially, ammonia-oxidizing bacteria (AOB) are responsible for converting ammonia to nitrite, which is then converted to nitrate by nitrite-oxidizing bacteria (NOB) in the second step. As shown in Fig. 1, this two-step process is critical for removing nitrogen entirely from wastewater, which usually contains nitrogen in reduced forms such as organic nitrogen and ammonia (Coskuner & Jassim, 2008).

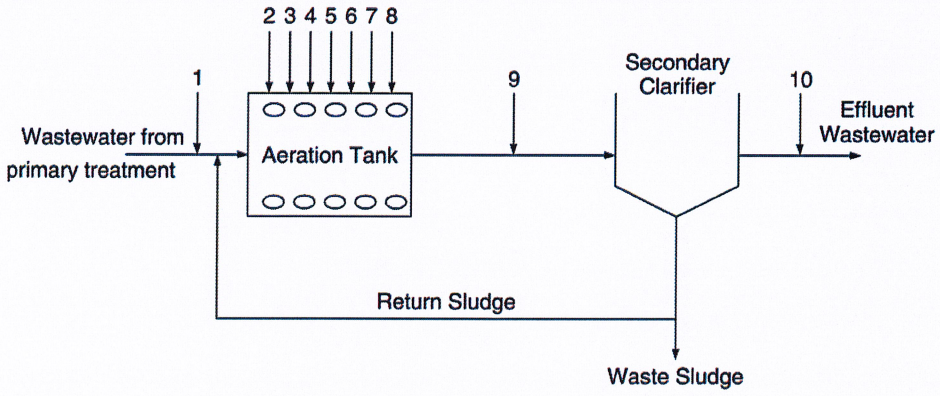


Fig. 1: Typical Process Flow Sheet of Biological Nitrogen Removal (Coskuner & Jassim, 2008)

Maintaining nitrification in municipal and industrial WWTP can be challenging due to the slower growth rate and lower competitiveness for oxygen of nitrifying bacteria compared to aerobic heterotrophs. These bacteria are also sensitive to reactor conditions such as temperature, pH, and Dissolved Oxygen (DO) concentration. Therefore, the application of ML algorithms to predict nitrification performance in a WWTP leads to a

reduction in the number of sample analyses, lowers operational costs, and improves time management (Cho et al., 2014).

1.3. Research Questions

This study will address the following research questions:

- What are the benefits of predicting nitrification performance in Muharraq WWTP?
- What has been the behaviour of influent wastewater and effluent treatment water data at Muharraq WWTP for the past four years?
- Which of the following methods (KNN, MLPs, SVR, RF, and GBM) will provide the most accurate prediction of nitrification performance?
- What are the influent parameters that affect the nitrification performance of WWTP, and to what degree do they influence it?

1.4. Research Aim and Objectives

This study aims to develop models that predict nitrification performance in treated sewage effluent for Muharraq WWTP using ML techniques. The primary objectives of this investigation are as follows:

- Statistically analyse the behaviour of influent wastewater characteristics and effluent treatment water data for the last four years.
- Predict the nitrification performance by predicting its effluent parameters (NH_4-N, NO_2-N, TKN, and TN) using ML methods like (KNN, MLPs, SVR, RF, and GBM).

- Identify the most appropriate variables which might be effective on the nitrification performance of WWTP.
- Find the optimal ML models by comparing the predicted values with the actual values of each method using four evaluated metrics (MAE, MAPE, RMSE, and R^2).

1.5. Research Significance

The significance of predicting the nitrification performance is substantial, as it has implications not only for the company but also for governments and society. It can be summarized as follows:

- This study utilized local data from the Muharraq Domestic WWTP. Conducting nitrification performance studies locally is important because each location has unique characteristics that can affect the treatment process and the quality of the effluent produced. Furthermore, every country has its own set of regulations and standards that must be considered.
- Compliance with regulations is essential for WWTPs, as local regulations often mandate that certain discharge limits for effluent parameters such as (NH_4-N, NO_2-N, TKN, and TN) must be met. Predicting these effluent parameters can help ensure compliance and prevent penalties.
- Predicting nitrification performance can lead to cost-effective treatment by reducing the need for excessive chemical usage and laboratory testing. This reduction in sample testing can also result in decreased workforce requirements, thereby reducing overall operational costs.
- The nitrification process is sensitive to various factors, including temperature, pH, Influent Flow (IF), and others.

Therefore, local studies of nitrification can help optimise these factors and improve treatment efficiency.

- The discharge of nitrogen compounds from WWTP can significantly impact the local environment, including water quality and aquatic life. Optimizing the nitrification process can help minimize the impact of nitrogen compounds on the local environment and protect public health.

1.6. Thesis Structure

This research is organized into five chapters. Apart from chapter one that provides introduction, chapter two includes a background and literature review. The third chapter outlines the methods and procedures implemented in this study. The fourth chapter presents the results found and discusses them. Finally, the fifth chapter includes the conclusions, limitations, and recommendations for future research.

Chapter Two

Literature Review

2.1. Background Machine Learning Methadologies

2.2. Artificial Intelligence in Environmental Studies

2.3. Nitrification in Wastewater Treatment Plants

2.4. Local Studies about WWTP

2.5. Predicting Wastewater Treatment Parameters Using AI and ML

2.6. Research Gaps

Chapter Two
Literature Review

2.1. Background Machine Learning Methadologies

The background section will provide a comprehensive understanding of different traditional ML methods that are commonly used in regression tasks. Regression is a supervised learning technique that aims to predict a continuous numerical output variable based on a set of input features.

2.1.1. K-Nearest Neighbors (KNN)

KNN is a popular machine learning algorithm used for classification and regression problems. Instance-Based learning with parameter K (IBk) is a type of instance-based learning algorithm that is closely related to the KNN algorithm. In IBk, KNN is used to make predictions, similar to KNN. However, the main difference between KNN and IBk is that the value of K in IBk is determined through cross-validation (Aha et al., 1991).

The methodology of KNN involves selecting the K value to a given input data point and then making a prediction based on the majority classes in classification or the mean in regression, as shown in equation (1). In addition, the distance metric is also important for calculating the distance between the input and each data point. Euclidean distance is the most commonly used distance metric, as seen in equation (2). Still, other metrics, such as Manhattan distance, Minkowski distance, Chebyshev distance etc., can also be used (Ontañón, 2020).

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} \hat{y}_i \quad (1)$$

$$D = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \quad (2)$$

Where ($\hat{y}$) is the predicted value of the target variable, (k) is the number of nearest neighbours used to predict the target variable, ($\hat{y}_i$) the target variable value of the ith nearest neighbour. (x_i, y_i) is the coordinates of the first point and (x_{i+1}, y_{i+1}) is the coordinates of the second point.

KNN has many advantages, including its simplicity, efficiency, and ability to handle categorical and continuous data. Additionally, the algorithm is non-parametric and does not make any assumptions about the underlying distribution of the data. Furthermore, the value of K is determined through cross-validation, which helps optimise the model's performance. However, KNN does have some limitations that should be taken into account. The algorithm can be sensitive to the choice of the similarity measure and weighting scheme, which can affect its performance.

Additionally, KNN can be computationally expensive for large datasets, as it requires the computation of the similarity measure for each training instance. The algorithm can also be sensitive to outliers based on a local search for nearest neighbours. Finally, the performance of KNN can be affected by choice of K value, which may need to be adjusted based on the dataset's characteristics (Alpaydın, 2010).

2.1.2. Multilayer Perceptrons (MLPs)

MLPs are kinds of feed-forward neural networks that are commonly used for regression and classification tasks. One of the key components of an MLPs is the backpropagation algorithm, which is utilised to update the network's weight and biases during training. Various parameters affect MLPs, including the LR, momentum, hidden layers, neurons, and the number of epochs (Yousefi et al., 2012).

The LR is a hyperparameter that determines the dimensions of the steps taken during gradient descent. A lower LR might result in slow convergence, although a higher LR may cause the network to overshoot the optimal solution. Generally, the LR should be set to small values, such as 0.1, and gradually increased if the network is not converging. Another important hyperparameter used in conjunction with the LR is momentum. It helps to smooth the gradients and prevent the network from getting stuck in local minima. The momentum term is typically set to a value between 0 and 1, with a common default value of 0.2 (Zhang et al., 2015).

The hidden layers and the neurons therein, shown in Fig. 2, are also essential factors that affect the performance of MLPs. In general, deeper networks with more neurons tend to perform better, but they are more computationally expensive and may be prone to overfitting. Therefore, a common rule of thumb is to start with one or two hidden layers and then increase the layers and neurons as needed (Rachmatullah et al., 2021).

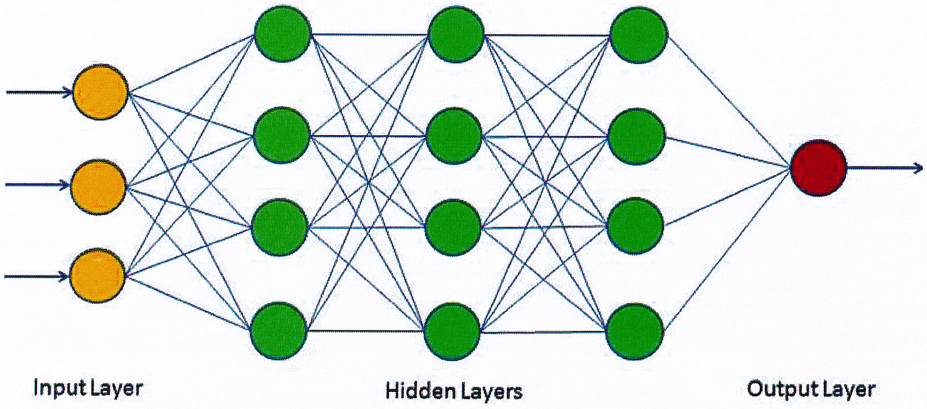


Fig. 2: Hidden Layers and Neurons of MLPs (Baek & Choi, 2020)

Another crucial hyperparameter is the number of epochs. An epoch is the number of times the network is trained on the entire dataset. The higher the number of epochs, the better the model will perform. However, there is a trade-off between performance and computational cost, as the higher the number of epochs, the longer the model will take to train, which may lead to overfitting (Cheng et al., 2020).

The activation function used in MLPs is also an important consideration. Using the sigmoid function for hidden layers and unthresholded linear units for output is common for regression tasks. Sigmoid functions introduce non-linearity into the neural network, allowing the network to model more complex relationships between inputs and outputs. Unthresholded linear units simply output the weight of their inputs without applying the activation function. This tool is helpful for regression tasks, as the output is not limited. In addition, sigmoid units may work in tandem with unthresholded linear units in the output layer. This configuration can work well if the output variable is continuous and there is a non-linear relationship between the input feature and the output variable (Bouckaert et al., 2013).

MLPs are commonly used for both regression and classification problems. One of the main advantages of MLPs is their ability to model non-linear relationships between input and output, making them useful

for an extensive range of applications. Additionally, MLPs are easy to train, thanks to the backpropagation algorithm. However, MLPs also have some disadvantages. One of the main disadvantages is that they are prone to overfitting, especially when the hidden layers and neurons are too numerous. Additionally, MLPs can be computationally expensive, particularly for deep networks. Furthermore, MLPs are unsuited for problems involving many categorical inputs, such as image classification (Park & Lek, 2016).

2.1.3. Support Vector Regression (SVR)

As shown in Fig. 3, the main idea behind SVR is finding the optimal hyperplane that maximizes the margin between data and hyperplane while still allowing some data to be within the margin (S.K. Shevade, 1999). SVR is a powerful tool for modelling non-linear relationships, and it is beneficial in cases where the number of features outweighs the number of samples. The SVR algorithm is implemented using the Sequential Minimal Optimization (SMO) algorithm, which is a fast and efficient method for solving large-scale optimization problems (Chang, 2011).

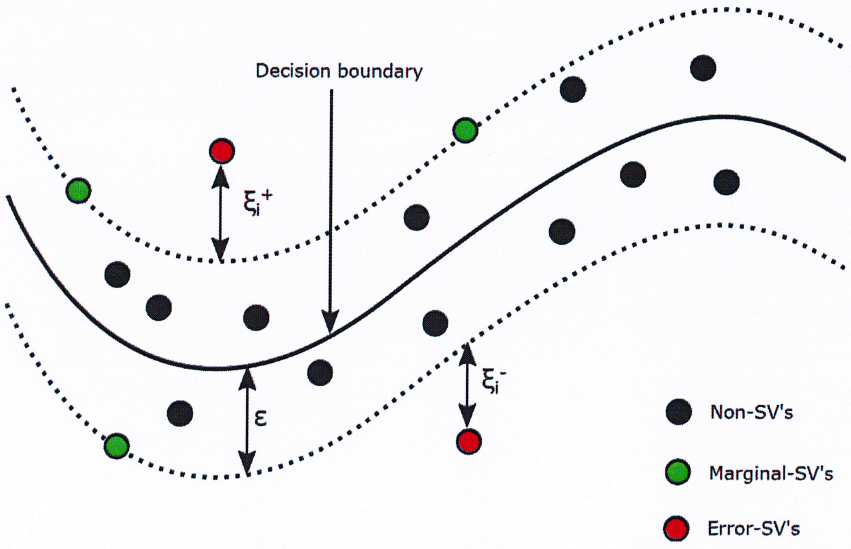


Fig. 3: SVR Structure (SINGH et al., 2020)

With SVR, the problem is defined as a convex optimization problem by creating a cost function and constraints. The cost function is the sum of the squared error between the predicted and actual output, with a regularization term. The regularization term is used to prevent the overfitting of the model. Constraints are used to ensure that the data is correctly classified. The cost function and constraints can be represented in equation (3):

$$\begin{aligned} &\text{Minimize,} && (1/2)\,||\omega||^2 + C * sum(\xi_i^+ - \xi_i^-) \\ &\text{subject to,} && y_i(\omega.T * x_i + b) \geq 1 - x_i \end{aligned} \tag{3}$$

Where (ω) is the weight vector, (x_i) is the input feature vector, (y_i) is the actual output, (b) is the bias term, (ξ_i^+, ξ_i^-) slack variables, and (C) is the regularization parameter (SINGH et al., 2020).

A kernel function must be chosen, and the kernel functions are mathematical transformations that map the input data into a high-dimensional feature space where it becomes linear. Different types of kernel functions can be used in SVR, such as the polynomial kernel, PUK kernel, and RBF kernel. The algorithm is initialized by choosing a set of support vectors and corresponding Lagrange multipliers.

2.1.3.1. Polynomial Kernel

The polynomial kernel is a non-linear kernel that can be used to model polynomial relationships between the input features and the output variable. It is defined as:

$$K(x_i, x_j) = (x_i.T * x_j + c)^d \tag{4}$$

Where (x_i) and (x_j) are input feature vectors, (c) is a constant, and (d) is the degree of the polynomial. The polynomial kernel has the advantage of being able to model complex relationships between the input features and the output variable. However, a disadvantage of the

polynomial kernel is that it can be sensitive to the choice of the degree of the polynomial and can be computationally expensive (Smola, 2004).

2.1.3.2. Radial Basis Kernel

The RBF kernel is a non-linear kernel that can be used to model non-linear relationships between the input features and the output variable. The RBF kernel is defined as:

$$K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2}, \gamma > 0 \quad (5)$$

Where (γ) is the width of the RBF kernel. The RBF kernel has the advantage of being able to model complex relationships between the input features and the output variable, and it is not sensitive to the choice of parameters. However, the disadvantage of the RBF kernel is that it can be computationally expensive, and the choice of width parameter can be crucial for the model's performance (Vapnik, 1998).

2.1.3.3. Pearson Univesal Kernel (PUK)

The PUK kernel is the product of the polynomial and RBF kernels. The PUK kernel is defined as:

$$K(x_i, x_j) = \frac{1}{\left[1 + \left(\frac{2\sqrt{||x_i - x_j||^2}\sqrt{2^{(1/\omega)} - 1}}{\sigma}\right)\right]^{\omega}} \quad (6)$$

Where (ω) is the weight vector, and (σ) is the width of the RBF kernel. The PUK kernel has the advantage of being able to model both linear and non-linear relationships between input features and output variables. However, a disadvantage of the PUK kernel is that it can be sensitive to the choice of the parameters of the polynomial and RBF kernels (B. Uestuen, 2006).

2.1.4. Random Forest (RF)

RF regression is a popular ML algorithm used for regression problems where the goal is to predict a continuous target variable. Leo Breiman introduced the algorithm in 2001, an extension of the RF algorithm for classification problems. The key idea behind RF regression is to create an ensemble of decision trees to make a prediction, where each tree is trained on a different subset of the data and a different subset of the features (Segal, 2004).

RF for regression is created by developing trees based on a random vector (Θ). The tree predictor $h(X, \Theta_k)$, generates numerical outputs instead of class labels. The training data is assumed to be obtained independently from the distribution of the random vector (Y), (X). The mean-squared generalization error for any numerical predictor $h(X)$ calculated as:

$$E_{x,y}\,(Y - h(X))^2 \tag{7}$$

The RF predictor is formed by averaging k trees. Just like in the classification scenario, it holds that as the number of trees in the forest increases towards infinity, the mean-squared generalization error:

$$\mathrm{E}_{x,y}(Y - av_k h(X, \Theta_k))^2 \rightarrow E_{x,y}(Y - E_h(X, \Theta))^2 \tag{8}$$

The RF regression algorithm works by first selecting a random sample of the data to train each decision tree. This random sample, called a bootstrap sample ensures each decision tree has a different training data set. This process helps to reduce the variance of the model and increases the diversity of the trees in the forest. Once the bootstrap samples have been created, the RF regression algorithm selects a random subset of features to use when splitting each node in the decision tree (Liaw &

Wiener, 2002). This process helps to reduce the correlation between the trees in the forest and further increases the diversity of the trees.

Each decision tree in the forest makes its own prediction when making a prediction, and the RF regression algorithm averages these predictions, as illustrated in Fig. 4. The result of this process is that the RF regression algorithm can capture non-linear relationships between the features and the target variable. In addition, it can handle noisy and missing data better than a single decision tree. The final prediction made by the RF regression model is the average of the predictions made by each decision tree in the forest and is calculated using the following equation:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \tag{9}$$

where $(\hat{y})$ is the predicted value, (N) is the number of trees in the forest, and $(\hat{y}_i)$ is the prediction made by the decision tree in the forest (Cutler et al., 2018).

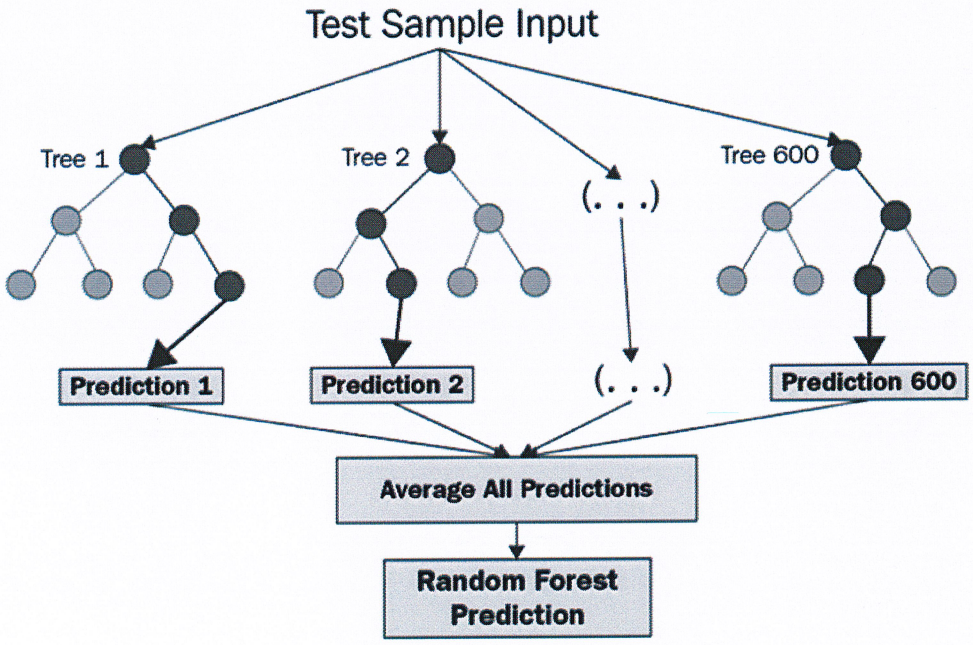


Fig. 4: RF Structure (Wang et al., 2021)

On the one hand, the RF algorithm is a highly accurate and robust method. It can handle simple and complex non-linear relationships. It's user-friendly and does not need much pre-processing or feature selection. Additionally, it's interpretable, transparent, and its decisions can be easily explained. However, this algorithm has some disadvantages, including computational expense, particularly with large numbers of trees. This can make the algorithm slow and unsuitable for large datasets. Furthermore, overfitting or underfitting may occur with too many or too few trees. Therefore, it may not always be the best choice, such as for highly linear relationships, which are better suited to linear regression models or highly skewed data (Chen et al., 2004).

2.1.5. Gradient Boosting Machine (GBM)

GBM is a supervised learning algorithm that can be used for regression and classification problems. The GBM algorithm is based on the idea of boosting, which involves combining weak learners to create a strong learner, as explained in Fig. 5. The weak learners in GBM are decision trees, and the algorithm iteratively adds decision trees to the model until the desired level of accuracy is achieved (Friedman, 2001). The final prediction of GBM is calculated by:

$$G(x) = sign\left(\sum_{m=1}^{M} \propto_m G_m(x)\right) \tag{10}$$

where $(G(x))$ is the final prediction for a given (x), $(\propto_m)$ are the weight assigned to each individual model $(G_m(x))$.

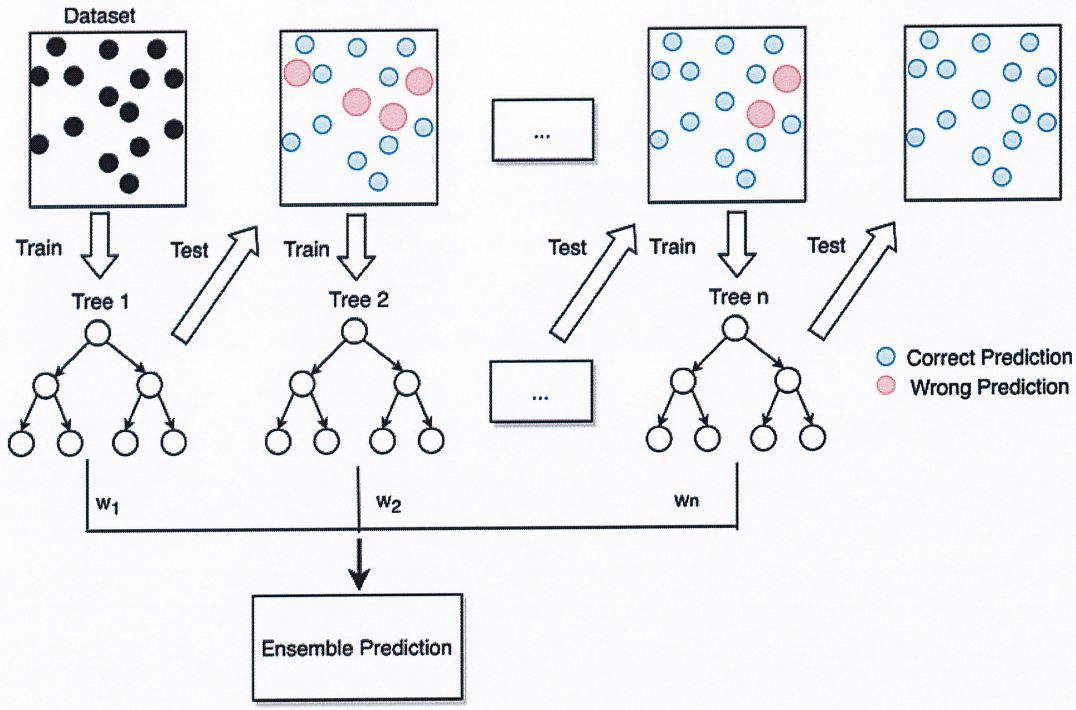


Fig. 5: Schematic of GBM Principle (Zhang et al., 2021)

The GBM algorithm has several key parameters that control the model's behaviour. One of the most critical parameters is distribution. The distribution parameter determines the type of loss function used to train the decision trees. The Gaussian, Laplace, and t-distribution are all commonly used loss functions in GBM. The Gaussian distribution is used for regression problems, while the Laplace and t-distribution are usually used for classification problems (Chen & Guestrin, 2016).

Another critical parameter in GBM is the shrinkage. Shrinkage controls the LR of the algorithm, which determines how quickly the algorithm converges to a solution. The LR is a value between 0 and 1, and a smaller value indicates that the algorithm will converge more slowly but probably will be more accurate. The following equation shows the effect of shrinkage on the LR:

$$LR = \frac{1}{2} * (1 - shrinkage) \tag{11}$$

The number of trees in the model is controlled by the n.trees parameter. This parameter determines the number of decision trees added to the model. The more trees added, the more accurate the model will be, but the computational cost will also increase.

The n.minobsinnode parameter controls the minimum number of observations that are required in a leaf node of a decision tree. This parameter prevents overfitting by ensuring that the decision trees in the model are not too complex. The following equation shows the effect of the minimum number of observations on the complexity of the decision tree:

$$Complexity = \frac{\text{leaf nodes}}{\text{n. minobsinnode}} \tag{12}$$

In addition to these parameters, GBM allows for using different decision trees, such as shallow and deep decision trees, which can further improve the model's performance (Hastie et al., 2009).

There is another method discussed by Natekin and Knoll (2008), known as the "Variable Influence" method. It is based on the idea of measuring the influence of a variable on the log-likelihood of the model. It is a way to analyse the effects of the variables captured by the decision tree ensemble. The following equations show the variable influence method:

$$Influence_j(T) = \sum_{i=1}^{L-1} I_i^2 1(S_i = j) \tag{13}$$

$$Influence_j = \frac{1}{M} \sum_{i=1}^{M} Influence_j(T_i) \tag{14}$$

The variable influence method defines the influence of a variable (j) in a single tree (T) as the number of times that variable is selected for splitting, weighted by the squared improvement of the model as a result of that split. (L) represents the number of splits within each tree. (S_i) represents the current splitting variable, (I_i^2) shows the empirical squared improvement. This measure is then averaged over all trees in the ensemble to obtain the overall influence of the variable. The resulting influences are standardised to add up to 100%, allowing for both forward and backward feature selection procedures.

2.1.6. Summary of Different ML Regression Techniques

This section aims to provide a comprehensive understanding of various traditional ML methods commonly employed in regression tasks, as depicted in Table (1). Regression is a supervised learning technique that focuses on predicting a continuous numerical output variable using a given set of input features.

Table (1): Overview of Various Machine Learning Regression Techniques

ML Method	Summary
KNN	KNN is a non-parametric algorithm that predicts the output value based on the average or majority vote of its k nearest neighbors in the feature space. It is simple to understand and implement, but its performance can be sensitive to the choice of k.
MLPs	MLPs are neural networks used for regression. They consist of interconnected layers of neurons and learn to predict continuous numerical outputs. MLPs capture complex patterns, but require large training data and hyperparameter tuning to prevent overfitting.
SVR	SVR is an extension of support vector machines for regression tasks. It aims to find a hyperplane that maximizes the margin around the training data, while allowing for a specified level of error. SVR can handle non-linear relationships by using kernel functions.
RF	Random forest is an ensemble method that combines multiple decision trees to make predictions. It uses a combination of bagging and random feature selection to reduce overfitting and improve generalization.
GBM	Gradient boosting is an ensemble method that combines weak learners in a sequential manner. It fits each subsequent model to the residual errors of the previous model, gradually reducing the overall prediction error.
Decision Tree	Decision trees are versatile models that recursively split the input space based on the features' values. Each split creates branches, leading to a final prediction at the leaves of the tree. Decision trees can handle both numerical and categorical input features.
Ridge Regression	Ridge regression is a regularized version of linear regression that introduces a penalty term to control the complexity of the model. It can help mitigate overfitting by shrinking the regression coefficients.
Lasso Regression	Lasso regression is another regularized linear regression method that uses the L1 norm penalty to encourage sparsity in the coefficient values. It can automatically select relevant features and perform feature selection.

2.1.7. Evaluation Metrics

Evaluation of a model is crucial when making predictions as it indicates its performance. Numerous evaluation metrics are available, but only a few are suitable for regression analysis, including MAE, RMSE, MAPE, and R^2 (Chatfield, 1995).

2.1.7.1. Mean Absolute Error (MAE)

The evaluation metric MAE measures how far the model's predictions are from the actual values on average. It is calculated as the average of the absolute differences between the actual values (X_a) and the estimated values (Y_a) for a set of (N) measurements, as shown in the equation below. MAE measures average error size in predictions, disregarding their direction. MAE is valuable when the error magnitude is more significant than its direction, as in the case of predicting a continuous value. A lower MAE value indicates a better performance of the model's predictions and vice versa (Willmott & Matsuura, 2005).

$$MAE = \frac{1}{N}\sum_{a}^{N}|X_a - Y_a| \tag{15}$$

2.1.7.2. Root Mean Squared Error (RMSE)

Another metric to assess the performance of a model's predictions is the RMSE. It is calculated by finding the square root of the average of the squared differences between the actual and estimated values for a set of measurements as defined in equation (16). This measures how far the model's predictions are from the actual values on average, considering the magnitude of the differences. RMSE is more sensitive to large errors than MAE, and it can be more useful when there are no significant outliers in the data. It is also easier to interpret since it is expressed in the same units as the target variable. A lower RMSE value indicates a better model's predictions and a higher value indicates a bad prediction (Chai & Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{N}\sum_{a}^{N}(X_a - Y_a)^2} \tag{16}$$

2.1.7.3. Mean Absolute Percentage Error (MAPE)

MAPE measures how far the model's predictions are from the actual values on average as a percentage of the actual values, as seen in equation (17). So, it provides a simple and interpretable measure of the difference between actual and predicted values, which makes it a valuable metric in various fields. A lower MAPE value indicates a better accuracy of the model's predictions, while a higher value indicates lower accuracy (Hyndman & Koehler, 2006).

$$MAPE = \frac{1}{N}\sum_{a}^{N}\frac{|X_a - Y_a|}{X_a} \times 100 \quad (17)$$

2.1.7.4. Coefficient of Determination (R^2)

The evaluation metric (R^2) measures the goodness of fit of a model's predictions as provided in equation (18). The R^2 value ranges from 0 to 1, with 1 indicating a perfect fit and 0 indicating no correlation between the actual and estimated values. However, it is important to note that a high value of R^2 does not necessarily mean that the model is a good fit for the data. A high R^2 may indicate that the model is overfitting the data, capturing noise or random variation rather than the underlying relationship between the predictor and outcome variables (Chicco et al., 2021).

$$R^2 = \left(\frac{n(\sum_a^N X_a Y_a) - (\sum_a^N X_a)(\sum_a^N Y_a)}{\sqrt{[n\sum_a^N X_a^2 - (\sum_a^N X_a)^2][n\sum_a^N Y_a^2 - (\sum_a^N Y_a)^2]}}\right)^2 \quad (18)$$

2.2. Artificial Intelligence in Environmental Studies

According to ISI Web of Science, it is widely acknowledged that Artificial Intelligence (AI) technologies are extensively used to address issues related to wastewater, air pollution, and solid waste management and have emerged as key tools in the fields of modelling, optimization, prediction, and control. Fig. 6 shows the progression of the number of publications related to these areas over the period 2009-2018, highlighting a significant upsurge in research related to environmental issues. Moreover, due to the complexities of WWT, researchers paid more attention.

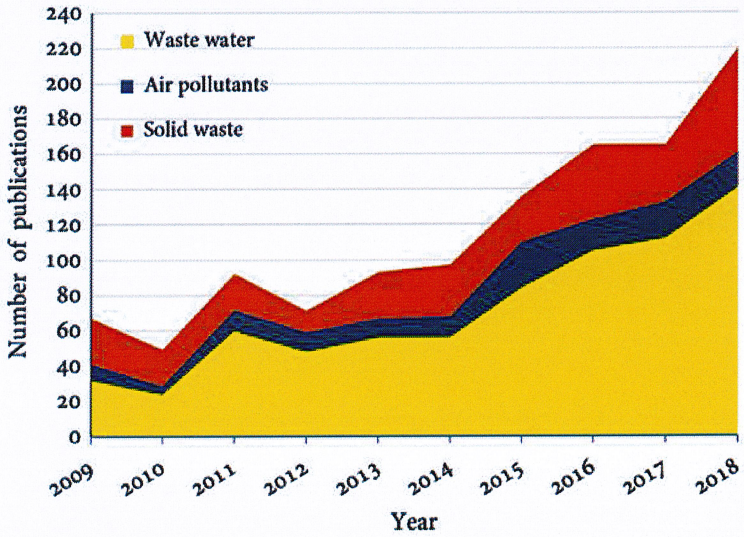


Fig. 6: The Number of Publications Classified by their Specific Environmental Areas of Application (Ye et al., 2020)

Initially, when AI technologies were applied to WWT processes, the primary focus was on eliminating pollutants such as heavy metals, dyes, persistent organic pollutants, nutrients, and other pollutants. Additionally, there were efforts to use AI for the modelling adsorption of various dyes from solutions. However, over time, the primary focus has shifted to advanced control of the wastewater disposal process and the development of soft-sensors to measure hard-to-measure parameters (Ye et al., 2020).

The following sections explores litrature related to nitrification in WWTP, including research on nitrifying bacteria and various factors that impact nitrification (such as temperature, dissolved oxygen, and pH), as well as methods for optimizing these factors to achieve maximum nitrification. The local studies on wastewater and highlights the research that leverages ML and AI to predict influent or effluent parameters for WWTP. Finally, the chapter tries to identify the gaps in the reviewed literature.

2.3. Nitrification in Wastewater Treatment Plants

Due to the fact that Ammonia Nitrogen (NH_3-N) requires oxygen and can be harmful to aquatic organisms in bodies of water where it is released, therefore it is necessary to effectively eliminate it from a variety of wastewater sources (Okabe et al., 2014). This section will present various investigative studies regarding nitrification.

2.3.1. Nitrifying Bacteria

Nitrifying bacteria are essential to the nitrogen cycle as they transform ammonia and nitrite into nitrate, a process known as nitrification (Ward, 2008). This is a significant stage in the biological treatment of wastewater since it eliminates hazardous nitrogen compounds from the water. Two main groups of nitrifying bacteria are involved in this process: AOB and NOB. The AOB convert ammonia to nitrite, while NOB converts nitrite to nitrate. Both groups of bacteria are chemolithotrophic, meaning they obtain energy from the oxidation of inorganic compounds rather than organic compounds. Nitrifying bacteria exist in many environments, including soil, water, and WWTP (Van Kessel et al., 2015).

Daims et al. (2015) studied the Nitrospira bacteria's role in nitrification, a crucial step in wastewater treatment. They found that Nitrospira can fully convert ammonia to nitrate and adapt to varying oxygen levels for efficient nitrification. Gubry-Rangin et al. (2011) studied niche specialization in terrestrial archaeal ammonia oxidizers, vital for the nitrogen cycle. They found distinct ecological preferences, with some prevalent in acidic soils and others in neutral soils. Environmental factors like soil pH and carbon availability strongly influenced their distribution and abundance.

Zhang et al. (2012) compared Ammonia-Oxidizing Archaea (AOA) and AOB's impact on ammonia oxidation in five different acidic soil samples with pH levels lower than 4.5. The findings indicated that AOA were more prevalent than AOB in most soil samples and that AOA were responsible for the majority of ammonia oxidation activity in strongly acidic soils.

2.3.2. Factors Affecting Nitrification

Several studies tried to understand the variables that influence the nitrification process in WWTPs. These studies provided essential knowledge that helped to design reliable biological nitrogen removal processes.

2.3.2.1. Effect of Temperature

A local experimental study investigated the influence of the temperature and DO concentration on nitrification (Adel et al., 2021). The testing of NO_2-N concentration was carried out at different temperatures ranging from 20°C to 45°C with 5°C as a step size while keeping the DO concentration at around 2 mg/L for 30 minutes and recording the result after every 5 minutes. It was observed that NO_2-N concentration

decreased at a faster rate at a temperature of 30°C compared to other temperatures, suggesting that this temperature is the most suitable for the nitrification process. The optimal level of nitrification performance was difficult during the summer months because of the elevated temperatures. The result shows a negligible change in the NO_2-N concentration when varying DO concentration.

Guo et al. (2010) found that temperature had a significant effect on partial nitrification, with higher temperatures promoting higher nitrogen removal rates but also leading to a decrease in the ratio of AOB to NOB.

Gnida et al. (2016) conducted another study at WWTP in Poland to investigate the impact of temperature on the efficiency of nitrification and its potential toxicity because the nitrogen removal in the plant varied between 0 to 98%, unlike the Chemical Oxygen Demand (COD) and Total Phosphorus (TP). During summer months, where the temperature reached 28°C, the nitrification efficiency increased to above 95%. In comparison, it decreases to below 50% in the winter months.

The analysis result negative relation between the NH_4-N in the mixed liquor reactor and its temperature with 0.71 Absolute Correlation Coefficient (ACC).

Finally, Rodriguez-Sanchez et al. (2014) investigated the impact of varying the temperature on the nitrification performance in an intermittently aerated sequencing reactor. The results demonstrated that the peak of nitrification efficiency is at 30°C. However, when the temperature rises over 30°C, the efficiency decreases because of the nitrite accumulation.

2.3.2.2. Effect of pH

The pH level present in wastewater can significantly influence the nitrification process during the treatment. Antoniou et al. (1990)

conducted an experimental study to investigate the effect of pH on the maximum specific growth rate of nitrifying bacteria. The study revealed that an optimal pH of roughly 7.8 was determined, and the effective maximum specific growth rate showed a monotonic increase as the temperature varied from 15-25°C.

Ciudad et al. (2007) conducted an experimental study on partial nitrification using an SBR. The study explored the effects of different operational modes and pH control techniques. Intermittent aeration and pH control resulted in significant improvements in partial nitrification, enhancing NH_4-N removal rates and nitrite accumulation.

2.3.2.3. Effect of Dissolved Oxygen (DO)

The nitrification process relies on oxygen to take place and thus is considered aerobic. As a result, the presence of DO is crucial in determining the nitrification rate.

Garrido et al. (1997) investigated the impact of dissolved oxygen (DO) on NO2-N in a Biofilm Airlift Suspension Reactor (BASR) for wastewater treatment. The study found that an ideal DO concentration of 2.5 mg/L resulted in complete conversion. DO levels between 1 and 2 mg/L achieved 50% conversion, while levels below 1 mg/L decreased NO2-N and NO3-N due to lower NH4-N conversion rates. In the range of 2.5 to 3.3 mg/L, changes were negligible, and the highest ammonia conversion was observed.

Guo et al. (2008) investigated the effects of DO concentration on NO_2-N accumulation in partial nitrification using a batch reactor. Results indicate that complete conversion occurred after 90 minutes at an optimal DO concentration of 3.5 mg/L, whereas the lowest conversion rate of around 66% was observed at a DO concentration of 0.5 mg/L after two

hours. On the other hand, both 2 mg/L and 6 mg/L DO concentrations resulted in 85% conversion after 90 minutes.

Liu (2012) study the impact of Solids Retention Time (SRT) and DO concentration on nitrification, both in the short and long term. The study found that nitrification was almost fully achieved in the 10, 20, and 40 days SRT reactors when the DO concentration was at least 0.37, 0.25, and 0.16 mg/L, respectively.

2.4. Local Studies about WWTP

Bahrain is a Small Island Developing State (SIDS) confronts numerous environmental challenges. The industrial and sewage waste has significant negative impact on Bahrain's marine ecosystem (Naser, 2020). Approximately 1.8 Million Cubic Meter (MCM) of industrial waste is released daily into Bahrain's marine environment. This waste contains various chemicals, such as trace metals and hydrocarbons, that have the potential to cause pollution (Bersuder et al., 2020).

Local research studies have examined various challenges related to wastewater, especially in Tubli bay, due to receiving high loads of sewage effluent. Tubli bay, located between Bahrain island and Sitra island in the eastern region of Bahrain, is renowned for its diverse marine and bird life and the lush mangrove forests that line its periphery.

Janahi et al. (2020) conducted a study to detect four types of enteric viruses in Tubli WWTP, analysing 59 influent and effluent samples using PCR analysis. It was found that the samples collected directly from Tubli bay did not reveal any evidence of the tested enteric viruses. However, it was noted that all samples of effluents contained a substantial concentration of faecal bacteriological markers.

Akhter & Madany (1991) evaluated wastewater quality. Influent, secondary, and tertiary effluent were analyzed between 1988 and 1989.

According to the analysis, the flow of wastewater received by the Tubli WWTP is classified as dilute and has low levels of Suspended Solid (SS) and Biochemical Oxygen Demand (BOD), with 142.8 mg/L and 124.6 mg/L, respectively. This can be attributed to excessive infiltration into the drains, resulting in elevated levels of Total Dissolved Oxygen (TDS) with 3488 mg/L in the wastewater.

The secondary effluent has BOD and SS concentrations of 6.9 and 2.4 mg/L, respectively, which is below the expected level due to the efficient treatment plant and low levels of organic matter in the raw wastewater. The tertiary effluent process ensures BOD and SS concentrations are below the standard value of 10 mg/L, making the resulting effluent suitable for most uses. The BOD and SS concentrations in the tertiary effluent are 1.5 mg/L and 2.0 mg/L, respectively. The study also suggests that there must be continuous monitoring of the coastal water and marine organisms near the discharge area to ensure that the effluent does not threaten the environment or public health.

A further investigation was carried out at Tubli bay to examine how the structure and composition of grey mangroves are influenced by the discharge of municipal wastewater and human-made sedimentation (Abou Seedo et al., 2017). It was observed that the structure and composition of the mangrove communities were similar across all study sites.

However, there was a positive effect on the height and diameter of the mangrove trees in the Tubli site due to the dilution effect of secondary treated wastewater. In other words, the wastewater discharge played a role in increasing the size of the mangrove trees in the area, potentially due to the nutrients and other elements present in the treated wastewater.

Al-Zubari et al. (2018) created a mathematical model to assess how climate change will impact municipal water management systems,

examining historical data from 2000 to 2012 and projecting forward to 2030. The vulnerability assessment showed that Bahrain's current municipal water management system is inefficient and is associated with high costs, which are predicted to increase over time if the current policies and management approach focused solely on supply-side management continue.

Also, the study found that implementing interventions such as decreasing the amount of generated wastewater and their associated collection and treatment costs could significantly reduce costs and make the system more efficient.

2.5. Predicting Wastewater Treatment Parameters Using AI and ML

In the field of WWT, the prediction of the effluent quality parameter is an important task that can help ensure the proper functioning of treatment plants and the protection of the environment. Traditionally, this has been done through laboratory testing, which can be time-consuming and expensive. However, with the advent of AI and ML techniques, it has become possible to predict effluent quality parameters more accurately and efficiently (Bekkari & Zeddouri, 2019).

Ye et al. (2020) discovered that although many different AI technologies are available, Artificial Neural Networks (ANNs) are among the most frequently used in environmental fields. This is likely due to the fact that ANNs are relatively easier to implement compared to other AI techniques. However, the effectiveness of ANNs is limited due to their low reproducibility and lack of comprehensive global search capabilities. Therefore, as AI technologies evolved rapidly, more researchers focus on hybrid methods such as Feed-forward Neural Network (FNN), Adaptive

Neuro-Fuzzy Inference System (ANFIS), and Genetic Algorithm-Artificial Neural Network (GA-ANN) instead of relying on single approaches, as shown in Fig. 7.

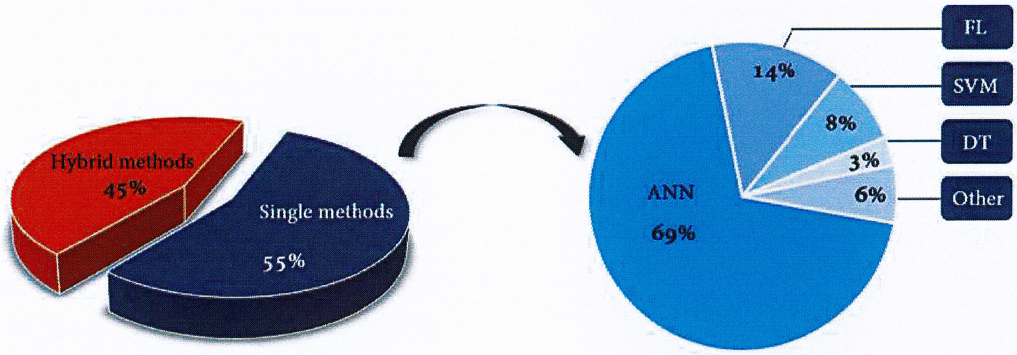


Fig. 7: The Statistics of Various AI Methods Utilized in the Reviewed Publications (Ye et al., 2020)

2.5.1. Total Nitrogen (TN)

Nitrogen is a significant cause of pollution in wastewater, and it needs to be reduced to an acceptable level before being discharged into the environment. Wastewater typically contains nitrogen in different forms, such as NH_3-N, NO_2-N, NO_3-N, and organically bound nitrogen. Therefore, it is essential to monitor the TN levels in the WWTP influent to ensure effective nutrient removal, minimize sludge production, and optimize the various stages of the treatment process (Salgot & Folch, 2018).

Bagherzadeh et al. (2021) utilized three supervised ML techniques: ANN, RF, and GBM, to predict the influent TN. The study used eight daily time-series influents as input variables, including pH, DO, COD, etc. In addition, four different scenarios were implemented based on feature selection to determine the most influential input variables that would yield the most accurate predictions based on three evaluation parameters: MAE, RMSE, and R^2. With 830 data points, only 30 were allocated for testing, while the remaining 800 were used for training. This training/testing split, with a significantly larger proportion of data used

for training, may lead to a model that is overfitted to the training data and lacks generalizability when applied to unseen data.

Among the scenarios tested, the one that included NH_4-N, COD, BOD, and DO as input variables yielded the best results, with an R^2 value of 0.58. However, the difference between this and the worst-performing scenario was relatively small, with only a marginal increase of 0.07 in R^2.

Conversely, Guo et al. (2015) predicted effluent TN using ANN and SVR. The SVR kernel was RBF with 50 as optimal regularization and 0.001 for gamma. The study used eight influent input variables, including temperature, IF, and influent TN. The dataset consisted of around 290 observations, with 200 used for training and the rest for testing. The R^2 values for ANN and SVR were almost the same, with 0.47 and 0.46, respectively. However, the SVR model suffered from overfitting, as evidenced by the training R^2 of 1. The study found that temperature had the most significant impact on TN, followed closely by influent TN and pH parameters.

Furthermore, Hvala & Kocijan (2020) utilized parallel Gaussian processes in the hybrid model structure to predict the effluent TN. The model was trained using 400 observations, while 250 observations were set aside as a test dataset. Ten different influent parameters were used as input parameters. The study demonstrated superior performance compared to the previous two studies, achieving an R^2 value of 0.814. This finding indicated that using parallel Gaussian processes in the hybrid model structure can be a practical approach for predicting effluent TN.

2.5.2. Total Kjeldahl Nitrogen (TKN)

The TKN test is used to quantify the presence of organic nitrogen and NH_3-N in wastewater. In municipal WWTP, TKN and NH_3-N levels are typically monitored at different stages of the treatment process. However, industrial WWTP tends to only measure the concentration of NH_3-N, disregarding the fact that organic nitrogen can also be converted into NH_3-N and thus contribute significantly to the overall NH_3-N load on the biological reactor (Liu et al., 2013). Therefore, it is a crucial factor in assessing water quality, as its levels falling outside the accepted range can pose a significant threat to aquatic ecosystems and public health.

Aguilar et al. (2021) employed multiple linear regression, SVR, and decision tree regression to estimate the TKN levels in the Colombia stream. The prediction was based on ten input variables such as: alkalinity, conductivity, pH, Total Suspended Solid (TSS), temperature, etc. The dataset used in this prediction was relatively small, consisting of only 30 instances collected from ten different sample points.

Five cross-validations were performed using three different evaluation parameters to ensure the reproducibility of the results. The decision tree regression provided the most reliable outcomes among the three models with RMSE of 0.29, MAE of 0.13, and R^2 of 0.84. On the other hand, the multiple linear regression produced the least accurate results with RMSE of 0.55, MAE of 0.43, and R^2 of 0.41.

An additional investigation was conducted on the Kavoor WWTP to forecast the effluent TKN based on the influent parameters using SVR and ANFIS (Manu & Thalla, 2017). The data collection period spanned from 2nd June 2014 to 30th August 2014, yielding a total of 88 instances. Out of these, 65 were used for training and 23 for testing. The study used seven influent parameters as inputs, including pH, TS, COD, temperature,

NH_3-N, TKN, and free ammonia. Based on the results, it appears that using SVR with an RBF kernel and regularization parameters of 15 and 0.15 for gamma led to superior performance compared to ANFIS. The RMSE obtained with SVR was 0.232, significantly lower than the RMSE of approximately 0.795 obtained with ANFIS.

2.5.3. Ammonium Nitrogen (NH_4-N)

Another critical parameter to monitor in wastewater is NH_4-N, a form of organic nitrogen generated during the decomposition of organic matter by bacteria. Monitoring NH_4-N levels is essential since high concentrations in water bodies can have detrimental effects on aquatic organisms (Caicedo et al., 2000).

Al-Ghazawi & Alawneh (2021) developed an ANN program that utilizes influent parameters such as IF, temperature, pH, BOD, COD, SS, and NH_4-N to predict the effluent NH_4-N and other effluent parameters. The authors utilized a trial-and-error approach to determine the optimal architecture for the neural network models. The best hyperparameters were identified as 0.01 Learning Rate (LR) and 0.9 momentum, with two hidden layers having varying numbers of neurons (not specified by the authors) and a 200 epochs training process with the sigmoid activation function.

The Wadi Arab dataset, comprising 487 days of data records for both influent and effluent parameters, was partitioned into a training set of 365 days and a testing set of 122 days. Various evaluation metrics were employed to assess the model performance, including MAE, RMSE, and R^2. The testing results for these metrics were 0.45, 0.63, and 0.26, respectively. The model for effluent NH_4-N was observed to suffer from overfitting due to significant disparities between the training and testing outcomes. Also, the neural model for NH_4-N exhibited strong sensitivity

to influent temperature, pH, BOD, and COD, while displaying moderate sensitivity to influent NH_4-N and only slight sensitivity to the IF.

On the other hand, El-Rawy et al. (2021) created a model to predict the removal efficiency of ammonia and other effluent parameters like TSS, BOD, and sulphide using different AI methods. The input data for the models were obtained from the samples of El-Berka WWTP data between 2011 and 2017, whereas the data collected in 2018 was reserved for testing the model's performance. The results primarily focused on COD, and it was noted that among the various predictions, the RMSE of ammonia exhibited the poorest performance, with a value of 6.45.

Finally, It can be concluded that no sole model can be universal due to the intricate nature of different WWTP processes. Consequently, based on the available data, creating more resilient and efficient models is necessary. Table (2) summarizes various models used in prior studies to predict different parameters of WWTP.

Table (2): Various AI Models and ML Algorithms Used to Predict Influent and Effluent Parameters of WWTP in Similar Studies

Reference	Inputs	Outputs	Data Size (Training, Testing)	Implemented Methods	Key finding
(Al-Ghazawi & Alawneh, 2021)	Influent: • IF • Temperature • pH • BOD • COD • SS • NH_4-N	Effluent: • COD • BOD • SS • NH_4-N	(365, 122)	• ANN	• The test results for NH_4-N for MAE, RMSE, and R^2 were 0.45, 0.63, and 0.26, respectively. • The effluent NH_4-N showed overfitting due to significant differences between training and testing outcomes. • The NH_4-N model was highly sensitive to influent temperature, pH, BOD, and COD, moderately sensitive to influent NH_4-N, and only slightly sensitive to the IF.
(El-Rawy et al., 2021)	Influent: • TSS • COD • BOD • NH3-N • Sulphide	Removal efficiency: • TSS • COD • BOD • NH3-N • Sulphide	(2920, 365)	• ANN • Backpropagation • Cascade-forward	• The results primarily focused on COD. • RMSE of ammonia exhibited the poorest performance, with a value of 6.45.

(Manu & Thalla, 2017)	Influent: • pH • TSS • COD • Temperature • NH_3-N • Free ammonia • TKN	Effluent: • TKN	(65, 23)	• ANFIS • SVR	• SVR with an RBF kernel led to superior performance compared to ANFIS. • The RMSE obtained with SVR was 0.232, significantly lower than the RMSE of approximately 0.795 obtained with ANFIS.
(Mehrani et al., 2022)	• Temperature • DO • pH	• N_2O	(15, 15)	• ANN • GBM • SVR	• Among the three models tested (ANN, SVM, and GBM), the ANN model had the highest R^2 value with 0.67, indicating the best fit to the test data. • The SVM model had the highest MSE and RMSE values (0.057 and 0.239, respectively), indicating the poorest performance in predicting the test data compared to the other two models.
(Bagherzadeh et al., 2021)	Influent: • DO • COD • BOD • NH_4-N	Influent: • TN	(800, 30)	• ANN • GBM • RF	• Among the scenarios tested, the one that included NH_4-N, COD, BOD, and DO as input variables yielded the best results, with an R^2 value of 0.58. • The difference between this and the worst-

					performing scenario was relatively small, with a marginal increase of 0.07 in R^2.
(Hamoda et al., 1999)	Influent: • Temperature • IF • BOD • TSS • First Stage: • BOD • TSS	Second Stage: • BOD • TSS • Third Stage: • BOD • TSS	(205, 23)	• Backpropagation	• Correlation coefficients for TSS using the model were 0.71 and 0.72 for secondary and tertiary treatments, respectively. • Corresponding BOD data had lower correlation coefficients of 0.65 and 0.74 for the same treatments.
(Guo et al., 2015)	Influent: • Temperature • TN • IF • SS • TN of food waste leachate • Months • COD	Effluent: • TN	(219, 60)	• ANN • SVR	• ANN and SVR yielded similar R^2 values of 0.47 and 0.46, respectively. • Due to its training R^2 value of 1, the SVR model appears to suffer from overfitting. • The temperature had the most significant impact on TN, with influent TN and pH parameters following closely behind in terms of their influence.

(Ly et al., 2022)	Influent: • IF • COD • NH_3 • TP • TN • Effluent: • COD • NH_3 • TN	Effluent: • TP	(7884, 876)	• SARIMAX • RF • SVR • GTB • ANFIS • LSTM	• Seasonal Autoregressive Integrated Moving Average (SARIMAX) consistently outperformed other methods in regression estimation with the lowest MSE, MAE, and MAPE values and the highest R^2 coefficient. • SARIMAX has reasonable computational time, which confirms its effectiveness in predicting effluent TP. • RF, SVM, Gradient Tree Boosting (GTB), and ANFIS were ineffective in handling large datasets with nonlinear and nonstationary characteristics.
(Hvala & Kocijan, 2020)	Influent: • IF • Temperature • COD • BOD • TN • NH4-N • TP • DO	Effluent: • TN • TP	(400, 250)	• Parallel Gaussian process in the hybrid structure.	• The R^2 value achieved in this model is high, with a score of 0.814, indicating a strong correlation between the predictor variables and the target variable. • Using parallel Gaussian processes in the hybrid model can effectively predict effluent TN.

2.6. Research Gaps

Scholarly studies have identified numerous gaps in knowledge, suggesting that there is still much to be learned and discovered.

Regarding global research on nitrification, it's worth noting that much of the research in this area has been conducted in laboratory settings. However, many of the effluent parameters predicted using AI, or ML do not directly relate to nitrification, whereas some studies have only focused on predicting NH_4-N and TN or TKN, which is insufficient for a comprehensive understanding of nitrification. Moreover, no study was found that predicts NO_2-N, which is a significant component of nitrification. It should be noted that most of these studies have utilized small data sets comprising less than 100 observations. This presents an issue as small sample sizes may not accurately represent the sampling population, resulting in unreliable predictions. Additionally, such data sets may not capture the full range of variation and complexity in the data, leading to biased or incomplete results.

To ensure the validity and reliability of predictions, it is essential to utilize larger and more representative data sets. The lack of ML techniques in the local studies pertaining to this field represents a significant research gap. Moreover, there is a limited number of research conducted on domestic WWTPs locally, with most studies focusing on Tubli bay and none utilizing ML or AI techniques. However, the current study stands out as it is exclusively focused on Muharraq WWTP and is considered the first to incorporate ML techniques in domestic WWT.

Chapter Three

Methods and Procedures

3.1. Research Methodology

3.2. Exploratory Data Analysis

3.3. Data Pre-Processing

Chapter Three
Methods and Procedures

3.1. Research Methodology

This chapter outlines and explains the methodologies applied when predicting nitrification performance. Fig. 8 displays the research framework, starting with data collection, including the data gathering process. Moreover, data pre-processing includes data cleaning, data partitioning, variable selection and outlier detection. Next, five algorithm methods have been implemented. These methods were selected for implementation based on their demonstrated effectiveness in similar datasets. Their track record of delivering reliable and accurate results in comparable data scenarios influenced the decision to utilize them. As a result, the application of these methods in the current study is anticipated to provide valuable insights and produce dependable outcomes. Finally, The implemented method's results were evaluated using four commonly employed statistical measurements: MAE, MAPE, RMSE, and R^2. These metrics play a crucial role in regression tasks as they provide valuable insights into model error and fit, enabling a comprehensive evaluation. By considering these metrics, one can effectively assess and compare different models, ensuring the selection of the most suitable one. Furthermore, the use of these evaluation metrics allows for meaningful comparisons with other studies.

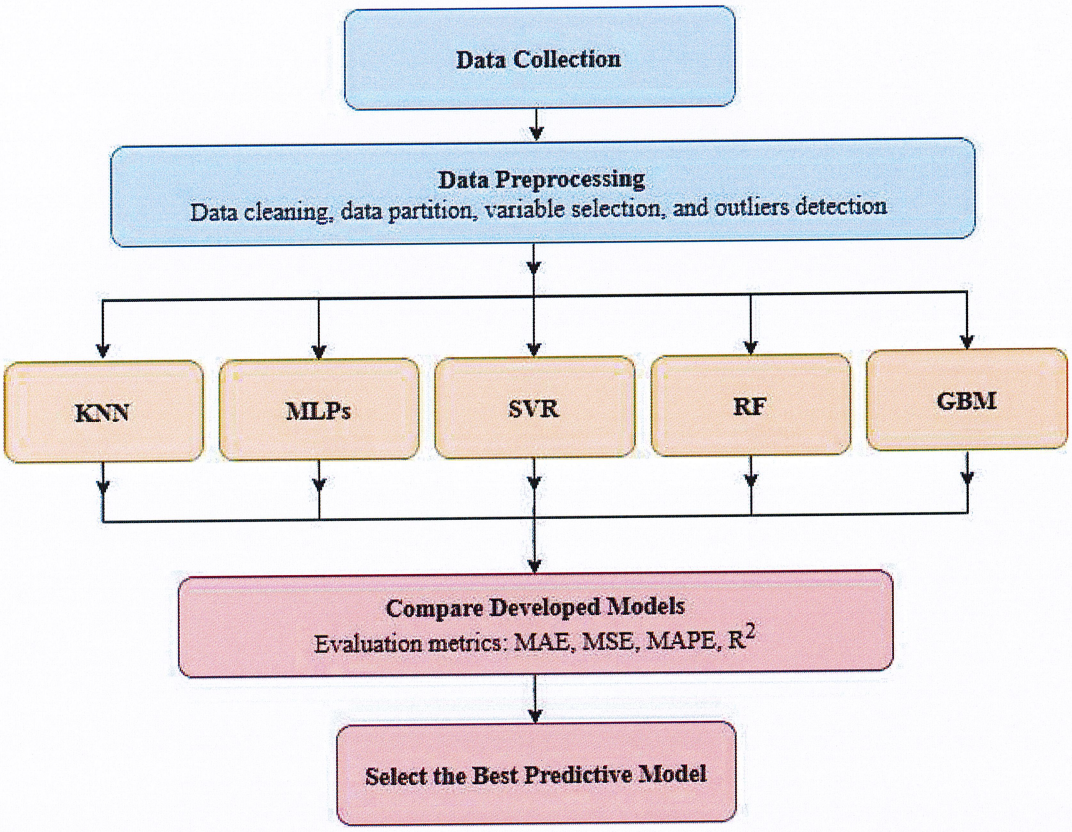


Fig. 8: Research Framework

3.2. Exploratory Data Analysis

The data was collected from Muharraq WWTP with 1729 instances from 1st Jan 2018 to 25th Sep 2022 and consisted of an influent and effluent parameter. The influent parameters consist of 20 attributes used as input variables to predict some of the effluent parameters, which consist of 18 attributes. In addition, the influent and effluent parameters have numerical and categorical attributes, as stated in Appendix (A).

3.3. Data Pre-Processing

The data were screened to ensure no missing values, and all measurements were formatted and logged correctly. During this process, five measurements were found to be missing from the BOD attribute and were replaced by the mean value. All numerical and categorical

measurements were found to be correctly formatted. Then, the data was split into ten cross-fold validation subsets, with one being used as the testing set and the rest as training sets. The model was run ten times using each subset as the testing data. Finally, the model's performance was judged by calculating the averages of the ten tests, as seen in Fig. 9 (Varma & Simon, 2006). This technique is powerful as it provides several benefits. Firstly, it maximizes the utilization of available data by partitioning it into multiple subsets or folds, enabling more comprehensive model training and testing. Secondly, it helps assess the model's generalization ability by averaging performance across multiple iterations, reducing the risk of overfitting or underfitting. Lastly, it offers a robust estimate of model performance by reducing dependency on a single data split.

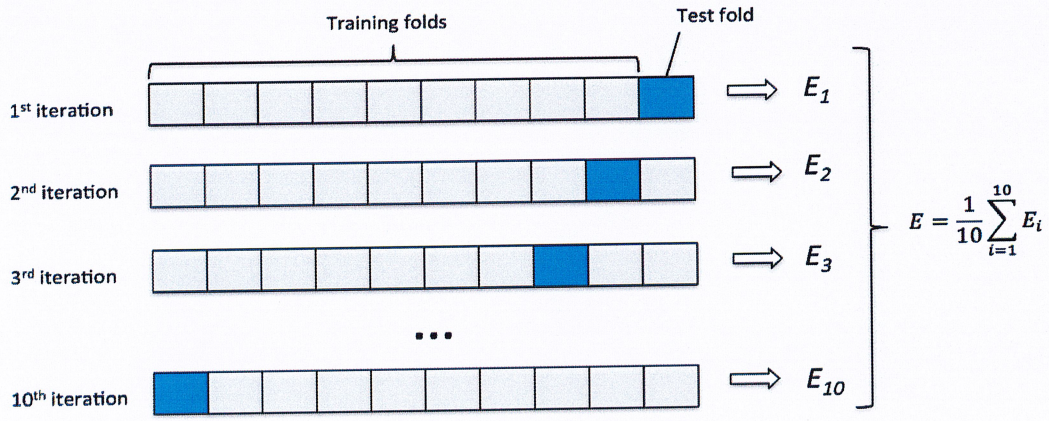


Fig. 9: Schematic of Ten Cross-Fold Validation

The variable selection is essential to make the model accurate and decrease complexity. This has been done by applying collinearity and Generalised Variance Inflation Factor (GVIF). Fox and Monette (1992) encourage the use of GVIFs, in order to easily compare them across dimensions. In general, if the value of GVIF is higher than 10, high multicollinearity is present between the measurements.

The purpose of outlier detection is to ensure that the model can run effectively when using testing data (Atkinson & Hawkins, 1981). However, a thorough understanding of data is a must, as predictors have to be fully understood to deal with the outliers. Therefore, according to the laboratory manager of Muharraq WWTP, most outliers are natural outliers, not errors.

For example, Fig. 10 shows that some values of the IF and SS are out of the standard range. The IF is usually out of the range due to rain or shut down of the upstream pumping station, while SS is due to the return of water from the plant. Accordingly, when beginning, the outliers will not be removed or substituted with mean or median values but will be after finding the optimal results to compare the optimal results with and without the outliers to study their effect.

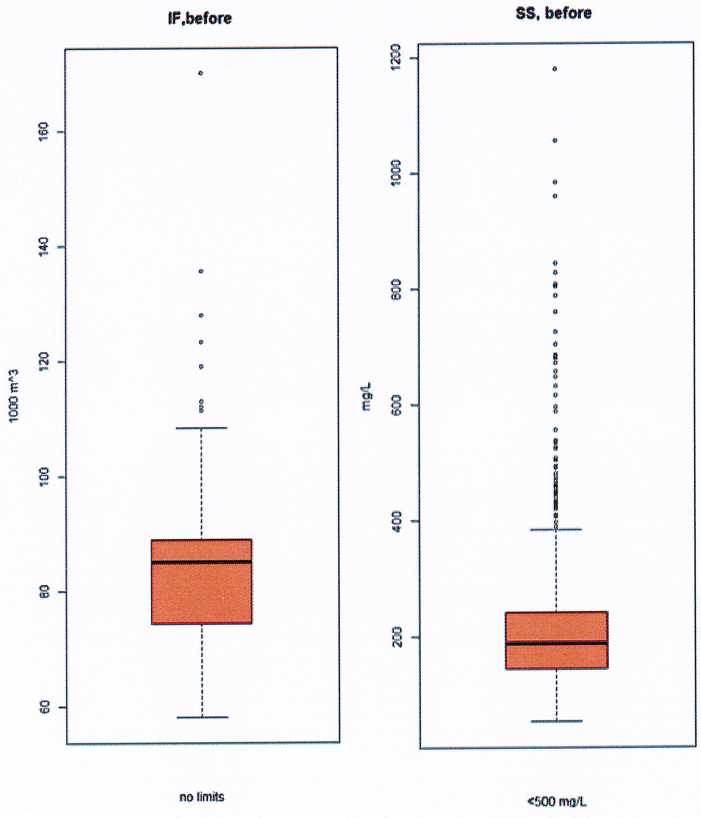


Fig. 10: Boxplots of IF and SS before Dealing with the Outliers.

Dealing with outliers has been processed separately based on the TSE range. Some predictors have TSE, and others do not, as represented in Table (3). Any outliers out of the TSE range have been substituted with the mean value. The predictors without the TSE range is dealt with based on Inter Quartile Range (IQR) (Hodge & Austin, 2004). Any value outside the range of ± (1.5 x IQR) was replaced with the mean value.

Table (3): The TSE Range of the Predictors with the Cause of Outliers

	TSE Range	**Cause of Outliers**
IF	---	Rain or shutdown of the upstream pumping station
SSDT	---	No outliers found
pH	6-10.3	No outliers found
SS	<500 mg/L	The return water from the plant
Conductivity	<12,000 µS/cm	Seasonal seawater or groundwater infiltration to the network
Sulphide	<80 mg/L	Increasing septicity of the sewage
OG	<340 mg/L	Contamination from sewage tanker
NH_3-N	---	The return water from the plant or contamination from the sewage tanker
Alkalinity	---	Seasonal seawater or the groundwater infiltration to the network for the effluent flow
COB	---	The return water from the plant or contamination from the sewage tanker

3.4. Data Description

Table (4) presents the data description, providing a comprehensive overview of the dataset's properties. This description aims to offer a clear understanding of the dataset, enabling effective interpretation and analysis.

Table (4): Wastewater Parameters and Definitions

Parameters	Type	Unit	Definition
IF	Influent	m^3	The rate at which wastewater enters a treatment system
SSDT	Influent	°C	One time grab sample measurement of temperature of influent at the influent station
pH	Influent	Unitless	The acid–base balance of water
SS	Influent	mg/L	Solid Particles present in untreated or partially treated wastewater and do not dissolve
Conductivity	Influent	µS/cm	The electrical conductivity of water
Sulfide	Influent	mg/L	Presence of sulfide compounds in untreated wastewater
OG	Influent	mg/L	Oil and grease level in wastewater
NH_3-N	Influent	mg/L	Ammonia Nitrogen level in wastewater
Alkalinity	Influent	mg/L	The water capacity to resist acidic changes
COB	Influent	Unitless	The Ratio between Chemical Oxygen Demand over Biochemical Oxygen Demand
NH_4-N	Effluent	mg/L	Ammonium Nitrogen level in treated water
NO_2-N	Effluent	mg/L	Nitrite Nitrogen level in treated water
TKN	Effluent	mg/L	Total Kjeldahl Nitrogen refers to the total amount of nitrogen present in untreated or partially treated wastewater in both organic and inorganic forms
TN	Effluent	mg/L	Total nitrogen encompasses all nitrogen compounds present in the system which are in organic and inorganic forms.

Chapter Four

Results and Discussion

4.1. Results

4.2. Discussion

Chapter Four

Results and Discussion

4.1. Results

The section illustrates the results of the variable selection process, provides descriptive statistics to help understand the evaluation metrics and outlines the behaviour of influent and effluent parameters. Further it includes the percentage of time that effluent variables were outside of the TSE range. Next, the results of the five implemented methods are given, with their optimal hyperparameters. Furthermore, the variable influence on each response will be shown. Finally, all five methods are compared based on four evaluation metrices.

4.1.1. Variable Selection

Starting with influent variables, the two categorical attributes, sample location and type, have been eliminated because their classes do not change. Moreover, more than 85% of the TDS values are missing, so they are removed.

In Fig. 11, collinearity is used to identify predictors that are highly correlated with each other. It was found that the SSDT and Dynamic Design Temperature (DDT) have 0.984 ACC. Also, the conductivity is highly correlated to the Chloride with 0.967 ACC because Chloride ions are major contributors to the conductivity of a solution. Surprisingly, all SS, COD, BOD, TN, and TP are highly correlated, with ACC is not below 0.825 when viewing all relations. Therefore, all DDT, Chloride, COD, BOD, TN, and TP have been eliminated, and alternatively SSDT, Conductivity, and SS were used.

It is better to predict the effluent quality from the influent parameters because it allows for proactive measures to be taken to improve the treatment process before the effluent is discharged into the environment. By predicting the effluent quality from the effluent itself, it would only be possible to identify any issues after the discharge has occurred, which could result in environmental harm or non-compliance with regulations. Since this study focuses on nitrification, the variables selected from the effluent parameters are NH_4-N, NO_3-N, NO_2-N, TKN, and TN. Unfortunately, all the values of the NO_3-N were repeated as 5 mg/L. For that reason, this attribute has been removed.

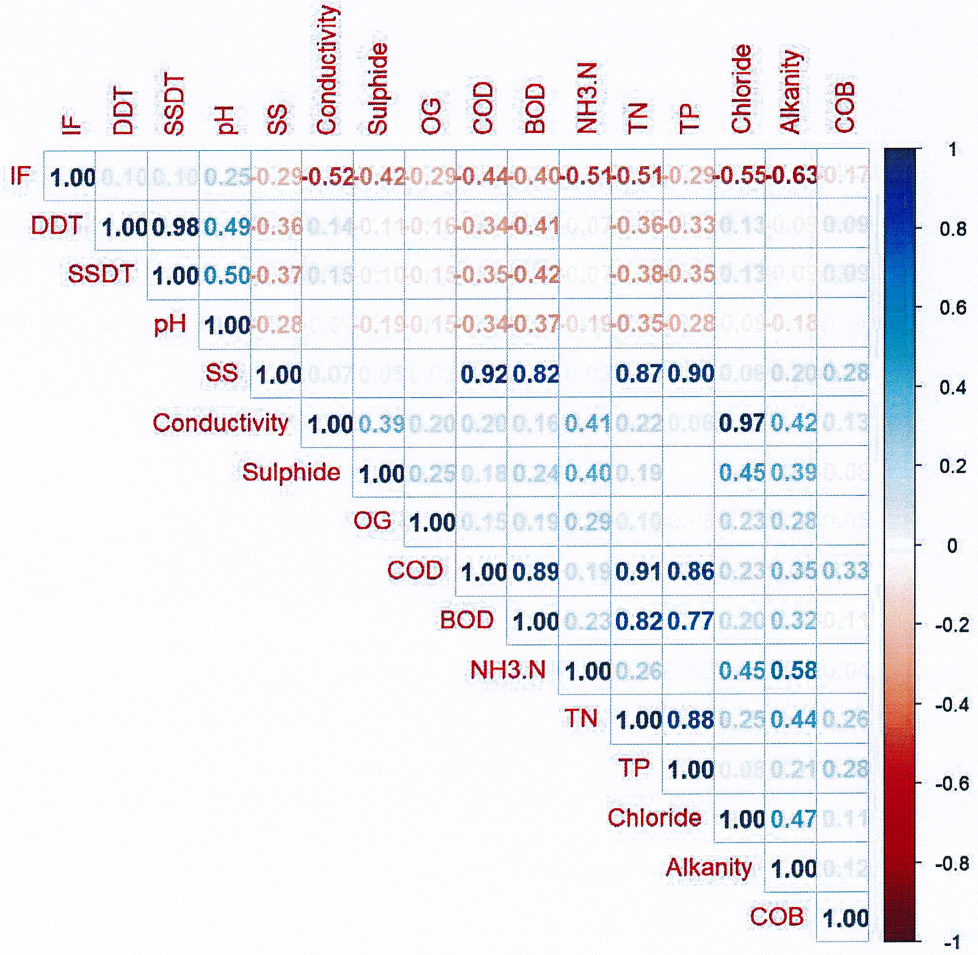


Fig. 11: The Collinearity between the Predictors

Table (5) outlines the Generalized Variance Inflation Factor (GVIF) before and after the variable selection.

Table (5): GVIF before and after Variable Selection

	GVIF before Variable Selection	GVIF after Variable Selection
IF	2.380	2.231
SSDT	32.178	1.597
pH	1.472	1.429
SS	10.090	1.434
Conductivity	17.217	1.601
Sulfide	1.537	1.414
OG	1.295	1.172
NH_3-N	1.889	1.770
Alkalinity	2.179	2.007
COB	15.280	1.189

4.1.2. Descriptive Statistics

Table (6) gives the statistical summary of laboratory measurements of influent and effluent parameters of the Muharraq WWTP after the data has been pre-processed. The statistical summary of the effluent variable will assist to understand evaluation metrics like MAE and the RMSE.

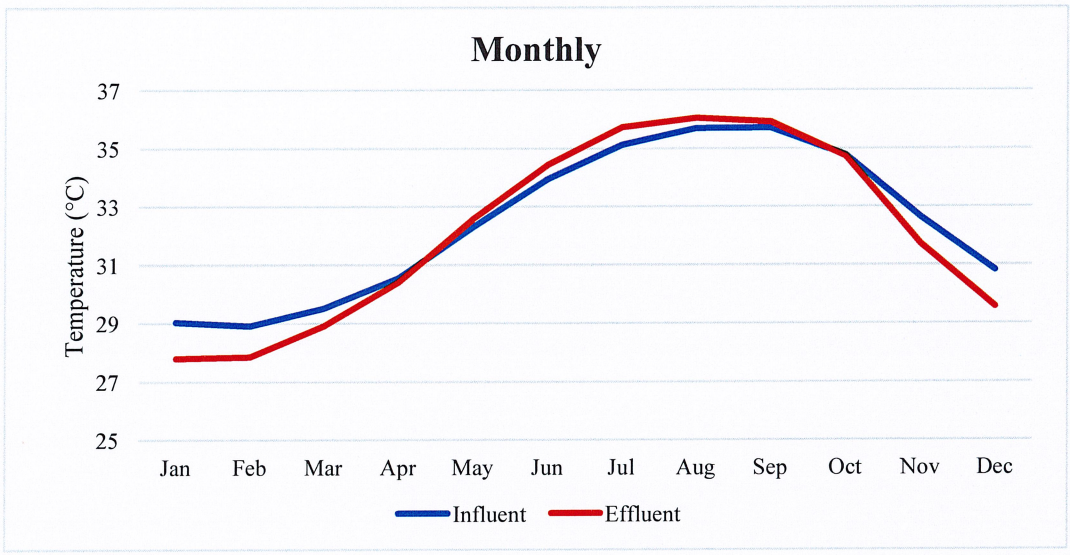
Table (6): Statistical Summary of Laboratory Measurements of Influent and Effluent Parameters

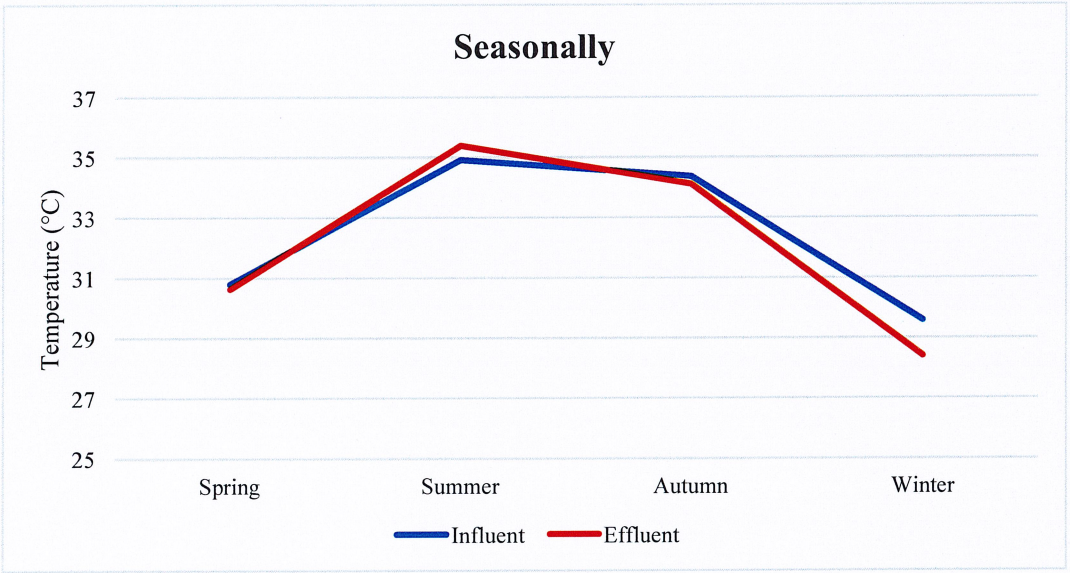
Parameters	Type	Unit	X_{min}	X_{mean}	X_{median}	X_{max}
IF	Influent	m^3	58,073	82,152	85,023	170,021
SSDT	Influent	°C	25.7	32.4	32.4	37.0
pH	Influent	Unitless	7.3	7.7	7.7	8.0
SS	Influent	mg/L	54.0	206.2	188.0	1180.0
Conductivity	Influent	µS/cm	1531	3100	3030	5990
Sulfide	Influent	mg/L	0.200	3.278	2.600	12.900
OG	Influent	mg/L	11.50	25.66	24.90	59.00
NH_3-N	Influent	mg/L	10.0	21.0	20.9	28.8
Alkalinity	Influent	mg/L	155.0	253.8	252.0	308.0
COB	Influent	Unitless	1.650	2.398	2.380	4.980
NH_4-N	Effluent	mg/L	0.100	0.620	0.500	7.500
NO_2-N	Effluent	mg/L	0.080	0.253	0.150	1.400
TKN	Effluent	mg/L	0.100	1.479	1.400	8.600
TN	Effluent	mg/L	2.000	3.799	3.600	11.400

4.1.3. Statistical Analysis

Fig. 12 shows the variation in influent and effluent temperature, measured by month, season, and year. It is observed that the monthly influent starts at a higher temperature in January, with influent and effluent temperatures crossing at around 30°C in April. Temperatures follow the same trajectory for the rest of the year, rising until August and afterwards gradually falling through December.

When measured seasonally, temperatures do not vary as much. Starting at a similar initial temperature of approximately 31°C, they increase during the spring and summer months until reaching a peak of roughly 35°C, then decrease gradually throughout the autumn and winter seasons. If measured yearly, average temperatures show the most variation. Beginning in 2018, influent was shown to have a lower standing temperature than effluent, with the two temperatures becoming equal in late 2019. At this point, the average temperature of the influent begins to rise, whereas the average effluent temperature begins to fall after this point. Both averages started a decline in 2021 from an influent temperature of 32.9°C and an effluent temperature of 32.1°C.





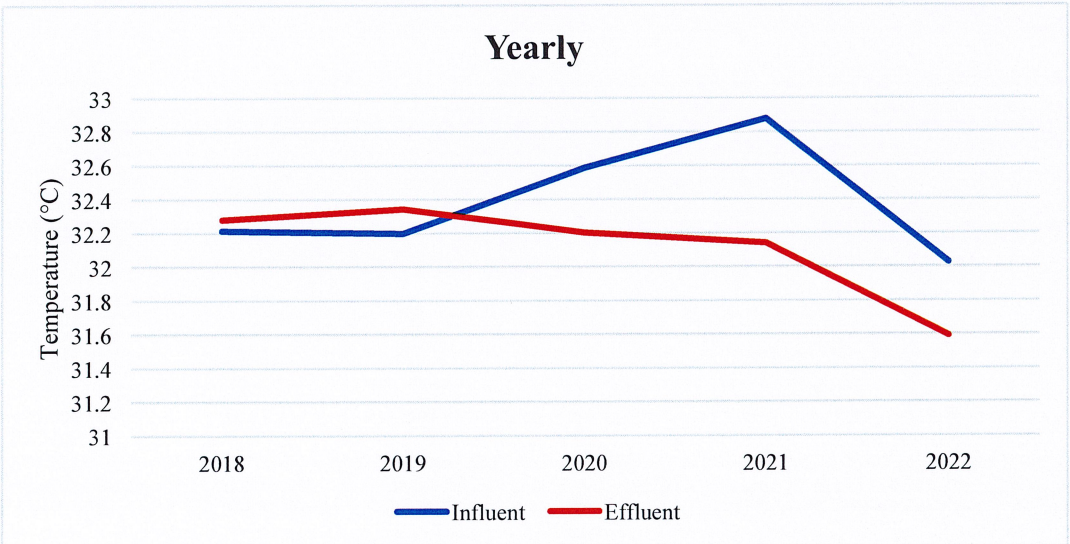


Fig. 12: Monthly, Seasonal, and Yearly Behavior of Influent and Effluent Temperature

Fig. 13 shows the concentration of Oil and Grease (OG), NH_3-N and TN of influent, again measured by month, season, and year. When measured monthly, the results show the highest fluctuation in results. TN rises from January to its peak in February of around 39 mg/L, decreasing until July and remaining relatively stable for every month measured. The concentration of OG fluctuates slightly during the first few months of the year before again becoming stable around June. NH_3-N does not fluctuate much, and levels remain relatively stable throughout all months.

The seasonal measures results show little change from those of the monthly figures. TN reaches a trough in summer before rising again to almost match its spring temperature during winter. OG falls by around 2 mg/L from spring to winter when measured by season. NH_3-N shows almost no change at all when measured by season.

The yearly statistics show TN to fall gradually from 2018 to 2021 before rising to around 35 mg/L. The OG and NH_3-N show minor variation over the years that measurements were recorded, both falling overall by approximately 4 mg/L. The behaviour of additional influent parameters was documented and recorded, and the outcomes are displayed in Appendix (C).

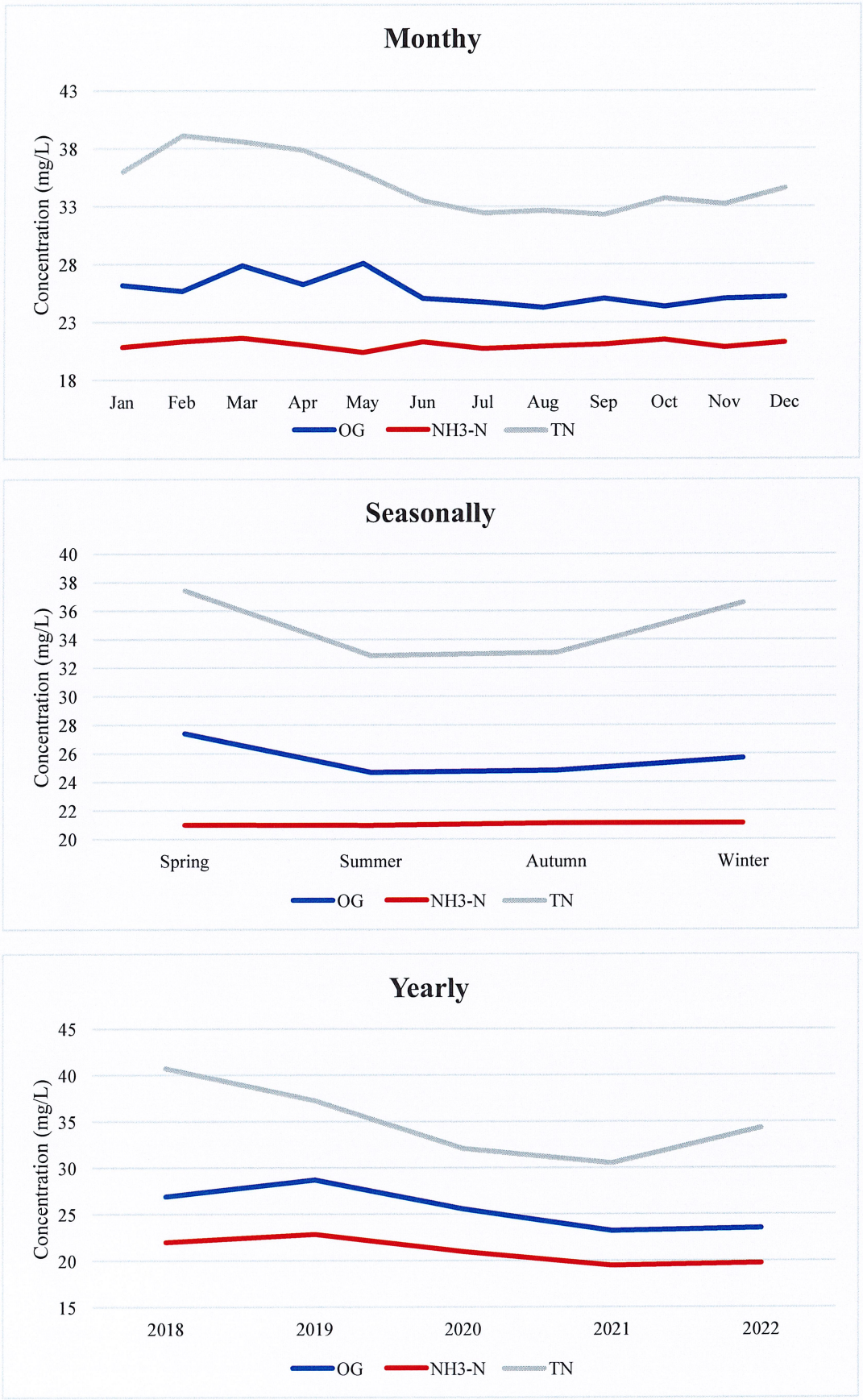


Fig. 13: Monthly, Seasonal, and Yearly Behavior of Influent Parameter

Fig. 14 shows the monthly, seasonally, and yearly behaviour of effluent parameters. It shows a slight variation in the concentration of NH_4-N, NO_2-N and TKN, regardless of time scale, with TN showing the most volatility in rsults.

Taking measurements by the month, TN does not show a tremendous overall increase; despite gradually rising until August, it falls again between August and December. On the other hand, NH_4-N and TKN spiked in August before settling back to previous levels, and NO_2-N showed slight fluctuation.

If measured seasonally, TN levels peak during summer before returning to a similar point in spring and winter of around 3.1 mg/L, with NH_4-N, NO_2-N and TKN remaining stable.

The yearly behaviours again do not show a significant overall increase or decrease. TN rose in 2021 before settling back at a rate of 3.5 mg/L in 2022. More effluent parameters were measured and shown in Appendix (C).

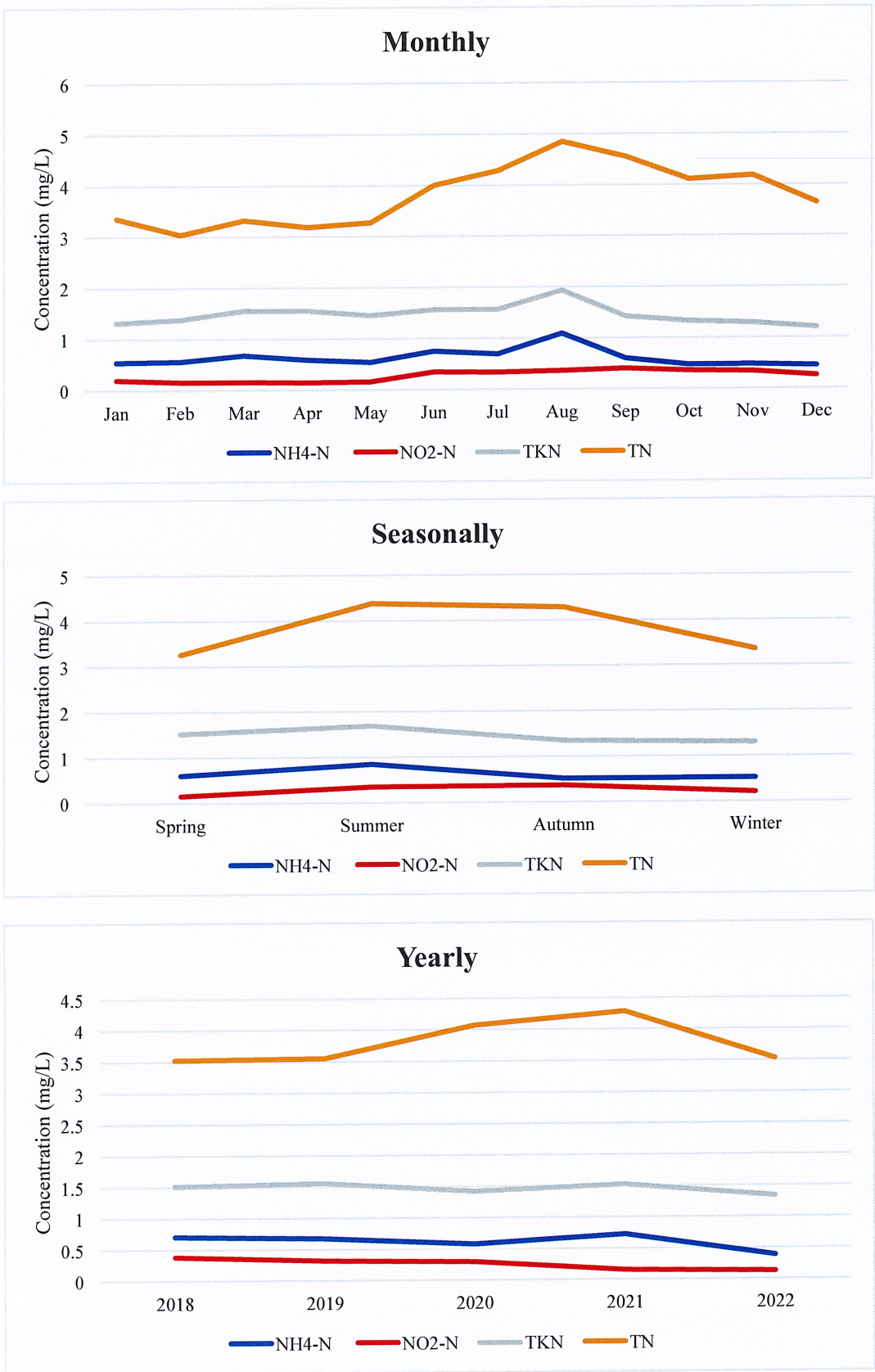


Fig. 14: Monthly, Seasonal, and Yearly Behavior of Effluent Parameter

Over the past five years, NH_4-N exceeded the acceptable range for 518 days out of 1,729, accounting for 29.96%, as shown in Table (7). In comparison, NO_2-N only surpassed acceptable limits 1.33% of the time or 23 days of the recorded period. TKN and TN were recorded outside acceptable ranges for nine and four days, respectively, equating to 0.52% and 0.23%. Notably, turbidity was outside acceptable levels for approximately half of the total period, while the remaining effluent parameters did not exceed acceptable levels throughout the entire period.

Table (7): The Effluent Parameters that are out of the TSE Range

NH_4-N	NO_2-N	TKN	TN	DO	TSS	pH	Turbidity	COD
29.96%	1.33%	0.52%	0.23%	0%	0.06%	0%	49.6%	0%

4.1.4. KNN Model Performance

Fig. 15 shows the MAE results obtained from KNN model. The model was run from K value 1 to 20 with a step size of 1, using NH_4-N, NO_2-N, TKN, and TN as responses. The MAE values for NH_4-N, TKN, and TN decreased the most between K values 1 and 5, with little to no reduction after K value 5. The initial MAE value for NH_4-N was 0.365 at K value 1, which decreased to its optimal value of 0.3003 at K value 20. For TKN, the initial value of MAE was 0.5211 at K value 1, fell to its optimal value of 0.3993 at K value 15, while the optimal value of TN was 0.5048 at K value 10, down from its initial value of 0.6169 at K value 1. These results indicate that the model's performance improved by raising the nearest neighbour values for NH_4-N, TKN, and TN. However, the MAE for NO_2-N showed little change throughout the K value increase, except for a marginal reduction of 0.0126 between K values 1 and 3.

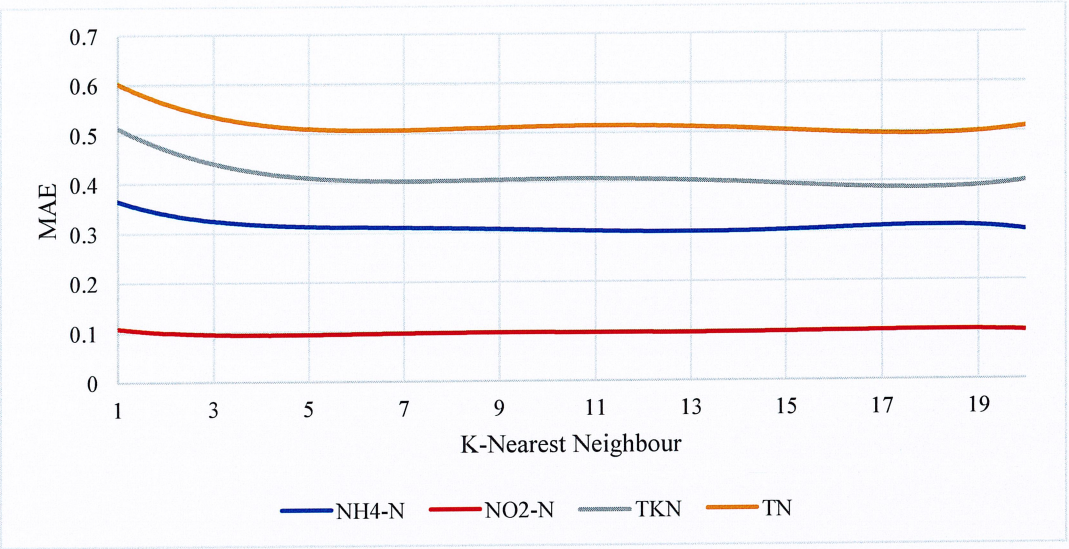


Fig. 15: The Effect of Different K Values on the Model performance

Table (8) presents the training and testing results of the KNN method, indicating no notable disparity between the training and testing outcomes for all the models. This suggests that there is no overfitting in the model, as demonstrated by the MAE values for the training data, which were 0.2858 for NH_4-N, 0.0867 for NO_2-N, 0.3736 for TKN, and 0.4546 for TN. The MAE values for the testing data were slightly higher than those for the training data, with values of 0.3003 for NH_4-N, 0.0958 for NO_2-N, 0.3993 for TKN, and 0.5092 for TN, but not significantly. In addition, the MAPE, RMSE, and R^2 values also showed consistency between the training and testing results, suggesting that the model performed well on the new data.

Table (8): Training and Testing Results of KNN using Optimal Hyperparameters

	NH_4-N		**NO_2-N**		**TKN**		**TN**	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
MAE	0.2858	0.3003	0.0867	0.0958	0.3736	0.3993	0.4546	0.5092
MAPE	80.81%	84.80%	52.33%	57.78%	84.89%	90.63%	61.88%	68.68%
RMSE	0.5196	0.5447	0.1369	0.1505	0.5955	0.6372	0.6533	0.7169
R^2	0.3187	0.2426	0.6386	0.5694	0.3149	0.2072	0.6274	0.5528

4.1.5. MLPs Model Performance

Regarding the impact of hyperparameters on the MAE of MLP models used to predict effluent parameters, a trial-and-error approach was employed to improve the model performance. The initial hyperparameters of the MLP model included a LR of 0.3, a momentum of 0.2, a single hidden layer with 6 neurons, and 500 epochs. The LR and momentum were varied during training. The LR ranged from 0.001 to 0.1, with a step size of 0.005, and then from 0.1 to 1 with a step size of 0.05. Additionally, the number of hidden layers was altered, ranging from one to three layers, with the number of neurons changing from 5 to 25 in increments of one. The number of epochs was also adjusted, varying from 100 to 5000.

Starting with the LR, It was observed that as the value of the LR increased, the MAE also increased. Fig. 16 illustrates that a lower LR allows the model to learn more quickly and converge to a better solution. Specifically, the MAE value for TN was the highest among the parameters, starting at 0.6 with an LR of 0.001 and increasing to 0.86 with an LR of 0.99. Additionally, the MAE for TKN ranged between 0.43 and 0.64 when the LR was near zero and followed an upward trend as the LR grew. The line representing NH_4-N showed a gradual upward inflexion from around 0.4 as the LR increased. Lastly, the lowest MAE value was observed for NO_2-N, which remained around 0.1 when the LR was near zero and increased gradually as the LR increased.

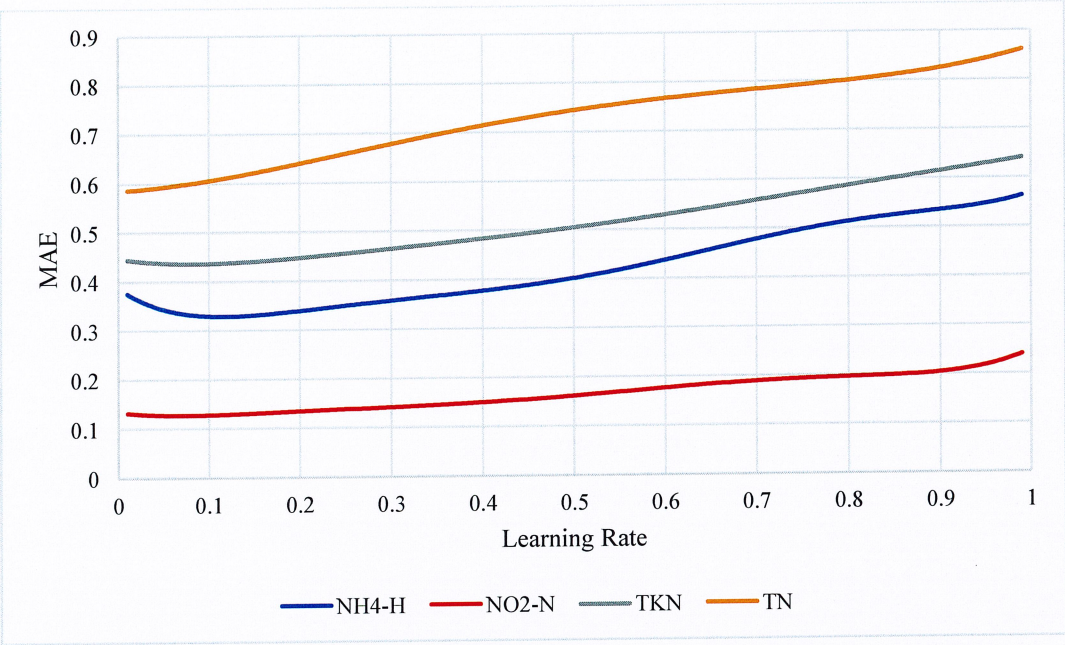


Fig. 16: The Effect of LR on MLPs Performance

Regarding the momentum, it was observed that as the value of the momentum increased, MAE also increased, as outlined in Fig. 17. This suggests that a lower momentum value helps the model to converge faster and avoid getting stuck in local minima. NO_2-N was the only measure not to show a significant change as momentum increased, with all other measures showing a severe upward trend once momentum reached around 0.8.

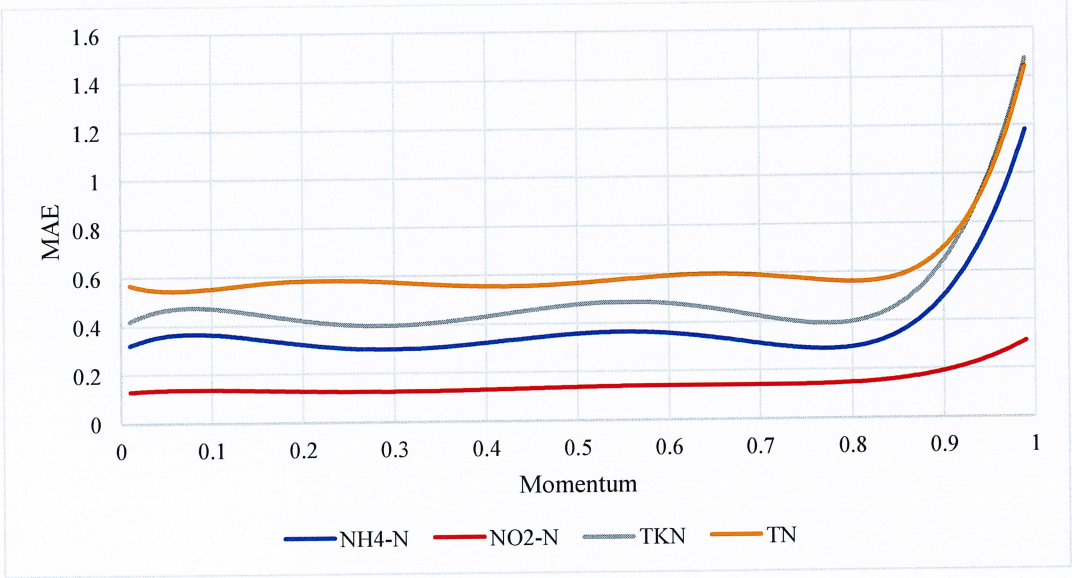
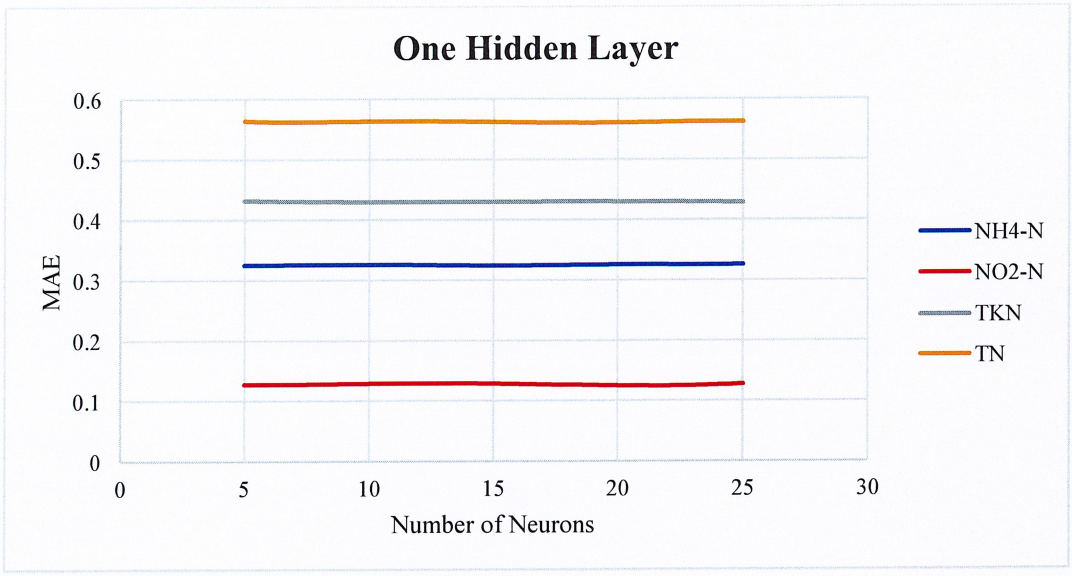
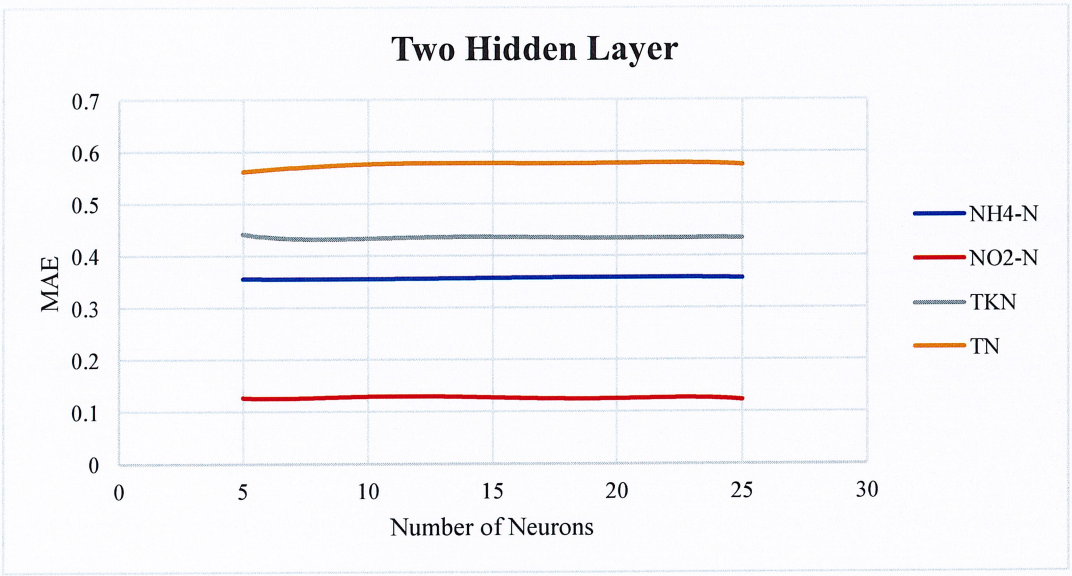


Fig. 17: The Effect of Momentums on MLPs Performance

Although the number of hidden layers did not significantly impact the MAE, there was a noticeable change in NO_2-N when three hidden layers were added. As a result, the MAE for NO_2-N reduced by 0.02, as depicted in Fig. 18. This result suggests that a deeper neural network architecture can improve the accuracy of predictions for NO_2-N. However, it is worth noting that the effect of adding more hidden layers was relatively minor and did not have a significant impact on the other measures. Therefore, careful consideration should be given to the trade-off between computational complexity and potential improvements in the model performance when selecting the appropriate neural network architecture.





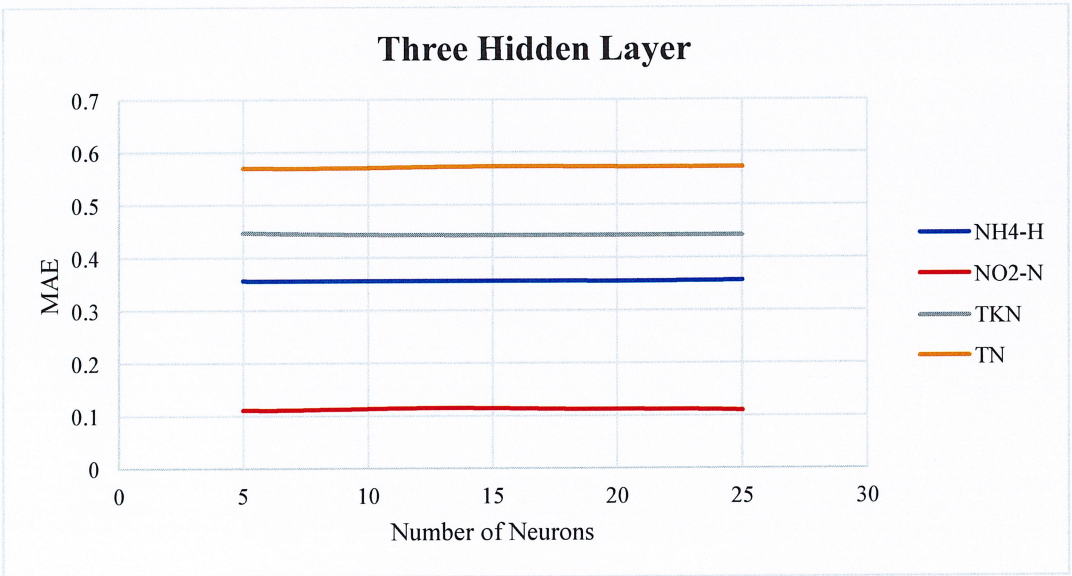


Fig. 18: The Impact of Adding Hidden Layers on MLP Performance

Finally, the number of epochs also impacted the MLPs performance. It was observed that the only real changes occurred before the number of epochs increased to 1000, though overall, more changes can be observed throughout Fig. 19. For TN, the MAE decreased from 0.63 to less than 0.54, the MAE of the TKN varied from 0.46 to 0.42, for NO_2-N, the change was from 0.14 to 0.097, and for NH_4-N, the MAE was reduced by around 0.04. This suggests that training the model for more iterations can thoroughly learn the data and achieve better accuracy.

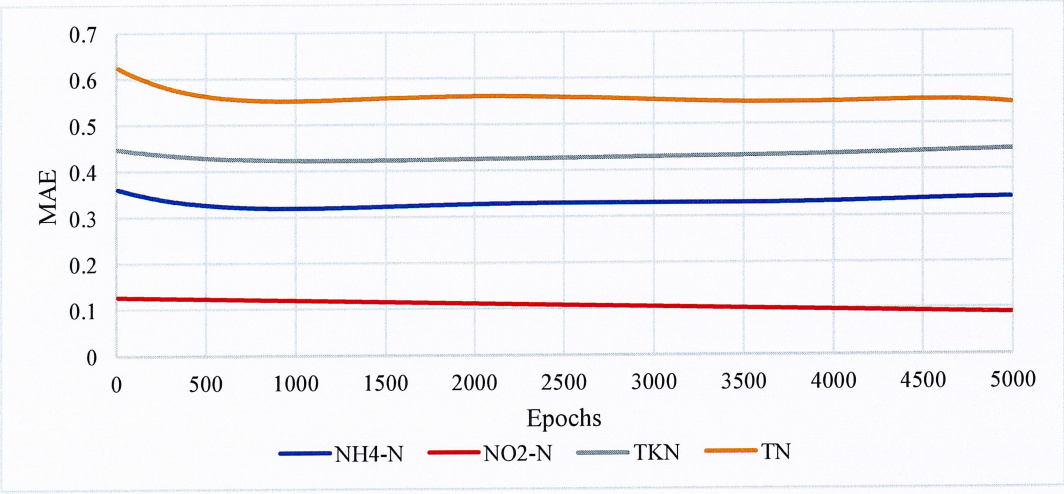


Fig. 19: Effect of Number of Epochs on MLP Convergence

The results show the optimal parameters for each output, NH_4-N, NO_2-N, TKN, and TN, that were found through training and testing an MLPs model. The optimal parameters include the LR, momentum, number of neurons, and the epochs for each output. The optimal parameters were determined by evaluating the model's performance using four evaluation metrics: MAE, MAPE, RMSE, and R^2. For NH_4-N, the optimal parameters were an LR of 0.1, a momentum of 0.4, 1 hidden layer with 13 neurons, and 1500 epochs, resulting in an MAE of 0.3201. For NO_2-N, the optimal parameters were an LR of 0.1, a momentum of 0.2, 23 neurons in the first layer, 8 neurons in the second layer, 25 neurons in the third layer, and 2000 epochs, resulting in an MAE of 0.0971. For TKN, the ideal parameters were an LR of 0.08, a momentum

of 0.2, 24 neurons, and one hidden layer, resulting in an MAE of 0.4233. For TN, the optimal parameters were an LR of 0.05, a momentum of 0.01, 21 neurons, and one hidden layer, resulting in an MAE of 0.5455.

Table (9) displays the outcomes of the MLPs method for both the training and testing data, indicating that there are little differences in the evaluation metrics for all the models. This suggests that there is no overfitting present, which is supported by the MAE values for the training data. Additionally, the consistency between the MAPE, RMSE, and R^2 values for both the training and testing data demonstrates that the model was able to perform effectively when presented with new data.

Table (9): Training and Testing Results of MLPs using Optimal Hyperparameters

	NH_4-N		NO_2-N		TKN		TN	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
MAE	0.3011	0.3201	0.0799	0.0971	0.4077	0.4233	0.4878	0.5455
MAPE	85.13%	90.40%	48.24%	58.53%	92.63%	96.08%	66.41%	74.23%
RMSE	0.5083	0.5555	0.1077	0.1538	0.6236	0.6635	0.6731	0.7767
R^2	0.3544	0.2367	0.7669	0.5721	0.3298	0.1694	0.6103	0.4832

4.1.6. SVR Model Performance

SVR was used for prediction with three kernels: polynomial, PUK and RBF. Tuning hyperparameters are used to determine the best fitting model for each response. A polynomial function is employed with three different exponent values, 1, 2, and 3, with regularisation starting at 0.0001 and increasing by a factor of 10, as shown in Fig. 20. Higher degrees of exponent values refer to a more complex model. Different regularization parameters are also used to check and compare the models. The model with exponent 3 showed slightly lower MAE scores for the NH_4-N with regularization 0.1 (0.2871), NO_2-N with regularization 1 (0.1065), TKN with regularization 0.1 (0.4039), and TN class with regularization 10 (0.5244).

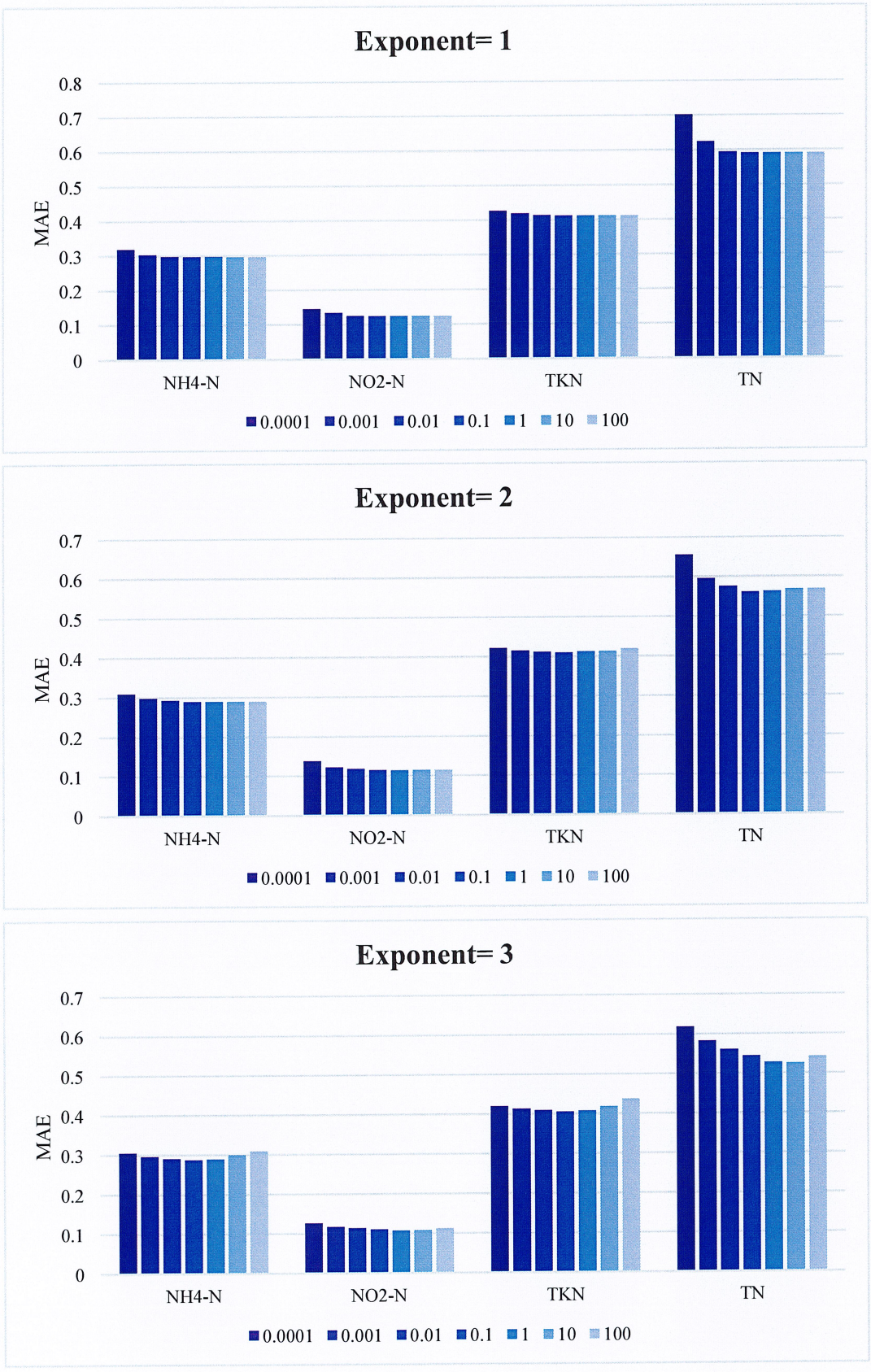


Fig. 20: How the Performance of SVR is Impacted by Varying Exponent Values

RBF kernel is employed with four different Gamma values, starting at 0.001 and growing by a factor of 10. As a result, the regularizations were also raised by a factor of 10, as shown Fig. 21. The Gamma parameter determines the width of the Gaussian curve used in the RBF kernel, with a smaller Gamma value resulting in a wider curve and a higher Gamma value resulting in a narrower curve. The model with Gamma 1 also showed lower MAE scores for the NH_4-N with regularization 1 (0.2803), NO_2-N with regularization 10 (0.1026), TKN with regularization 1 (0.3984), and TN with regularization 10 (0.5043) slightly.

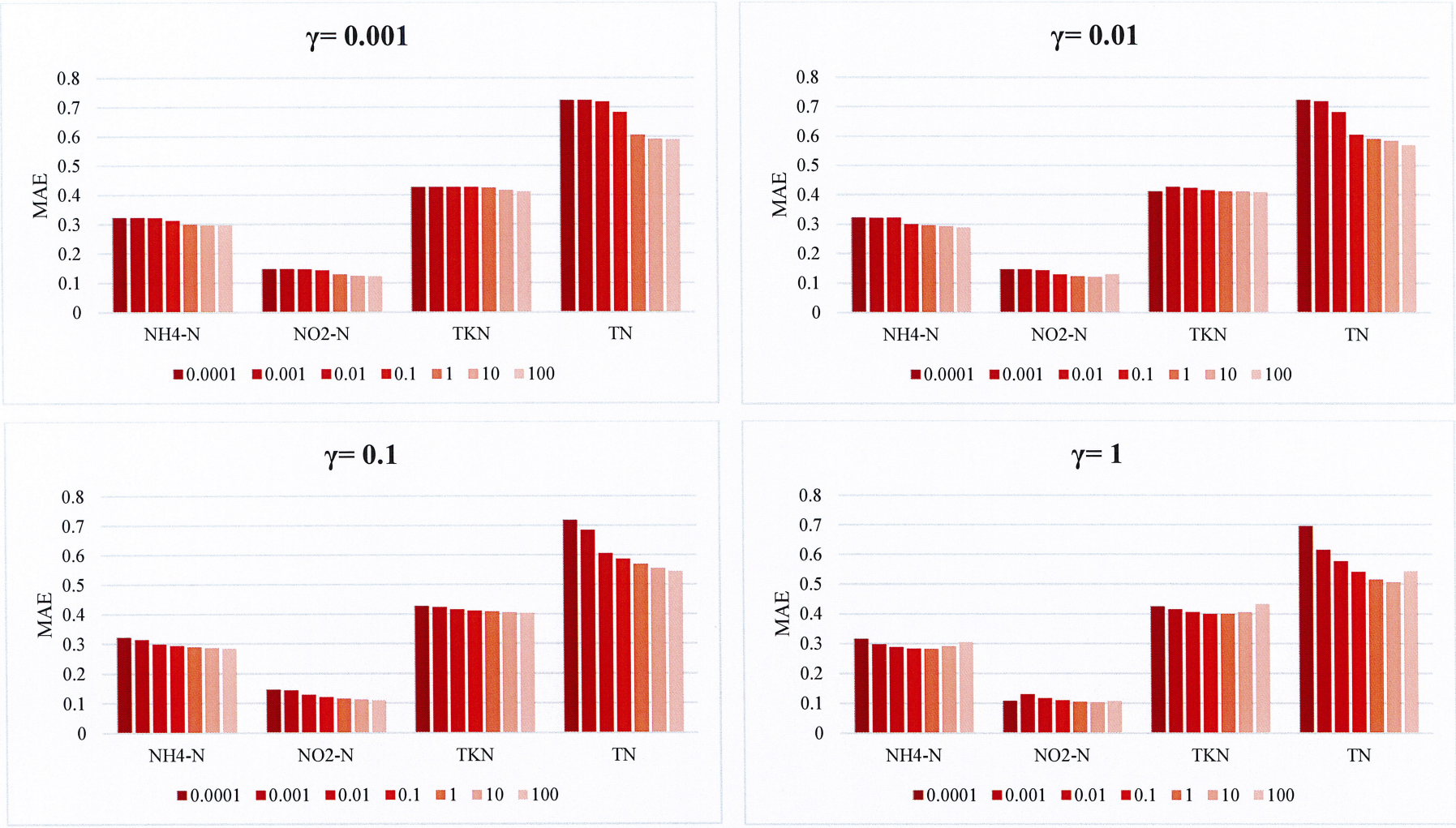


Fig. 21: The Effect of the Different Gamma Values on SVR Performance

PUK kernel were set the Omega and Sigma parameter values to 1. Omega scales the magnitude of the coefficients in the polynomial expansion where a higher degree refers to a complex model. On the other hand, Sigma smooths the polynomial expansion, where lowering the sigma value produces a more complex model. The results varied based on the choice of the regularization parameter, as observed in Fig. 22. The model achieved the lowest MAE score for NH_4-N with a regularization value of 0.1 (0.2759), NO_2-N with regularization 1 (0.0943), TKN with regularization 0.1 (0.3859), and TN with regularization 0.1 (0.4873).

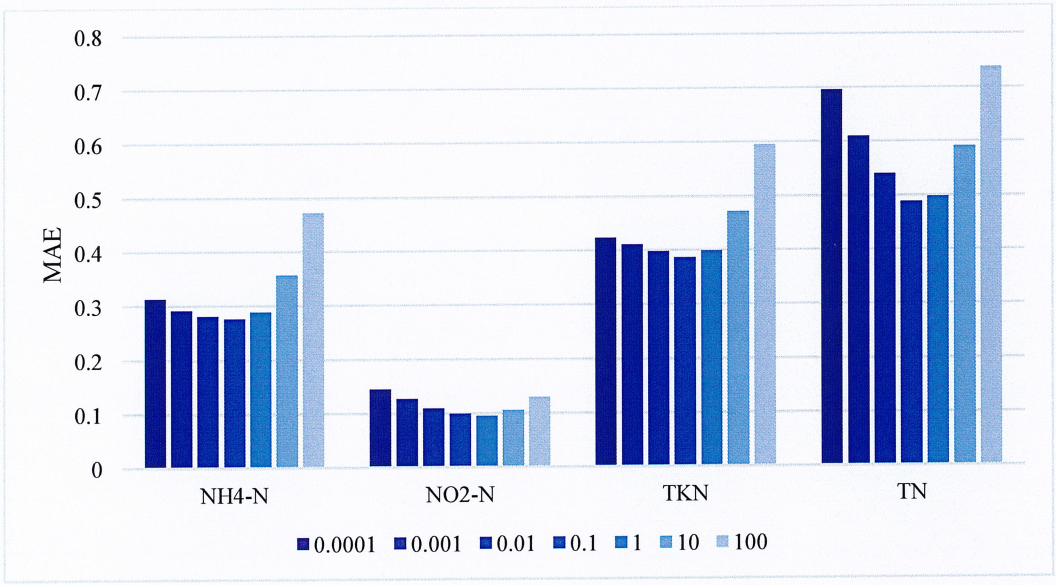


Fig. 22: The MAE Result of PUK Kernal

Looking at each of the kernel's performances, it is noticeable that the PUK kernel has the lowest MAE scores of all four classes and outperforms the RBF and Polynomial kernels, as shown in Table (10). Moreover, the PUK kernel also has the lowest MAPE values of all four responses, indicating that the predictions are better regarding percentage deviation from the actual values. The RBF kernel has slightly higher MAE scores compared to PUK. The Polynomial kernel has the highest MAE scores and MAPE values, implying that it is not the best choice for

these datasets. The R^2 values are higher for the PUK kernel but lower for the other kernels, indicating that the PUK kernel best fits all four classes.

Table (10): Comparing SVR Kernels using their Optimal Hyperparameters

		NH_4-N	NO_2-N	TKN	TN
PUK	MAE	0.2759	0.0943	0.3859	0.4873
	MAPE	77.92%	56.85%	87.58%	66.31%
	RMSE	0.5967	0.1575	0.6741	0.7413
	R^2	0.1426	0.5317	0.1662	0.5480
RBF	MAE	0.2803	0.1026	0.3984	0.5043
	MAPE	79.17%	61.86%	90.42%	68.61%
	RMSE	0.6121	0.1794	0.6890	0.7447
	R^2	0.0999	0.4016	0.1173	0.5203
Polynomial	MAE	0.2871	0.1065	0.4039	0.5214
	MAPE	81.08%	64.21%	91.66%	71.36%
	RMSE	0.6265	0.1881	0.7096	0.7657
	R^2	0.0748	0.3441	0.0581	0.4958

Table (11) displays little differences between the training and testing results for all the responses. The closeness of the errors suggests that there is no significant gap between them, indicating that the model did not overfit. Similarly, the MAPE, RMSE, and R^2 values for both the training and testing data were comparable, indicating that the model was able to perform effectively when given new data. Even though MAPE shows a high percentage and that's because when dealing with small values, the MAPE can potentially yield high percentages even for relatively small absolute differences. This is because the percentage difference can become large when the actual value is close to zero.

Table (11): Training and Testing Results of SVR using Optimal Hyperparameters

	NH_4-N		NO_2-N		TKN		TN	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
MAE	0.2295	0.2759	0.059	0.0943	0.3222	0.3859	0.4102	0.4873
MAPE	64.88%	77.92%	35.63%	56.85%	73.20%	87.58%	55.84%	66.31%
RMSE	0.5732	0.5967	0.1291	0.1575	0.636	0.6741	0.6882	0.7413
R^2	0.4345	0.1426	0.7080	0.5317	0.4635	0.1662	0.7563	0.5480

4.1.7. RF Model Performance

The results presented in Fig. 23 were obtained by running the RF model and recording its output, as the depth of the tree was increased from 1 to 15, with a step size of one. Each response exhibited a similar pattern, with the highest reduction in MAE occurring between maximum depths of 1 to 6 and gradually levelling out as the maximum depth increased. The optimal MAE values for each response were observed at maximum depths of 14 or 15. For NH_4-N, the optimal value of 0.2875 was obtained at a maximum depth of 14, while for NO_2-N, the optimal MAE of 0.0878 was also achieved at a maximum depth of 14. The optimal conditions for TKN and TN were both observed at a maximum depth of 15, with MAE values of 0.3812 and 0.4583, respectively.

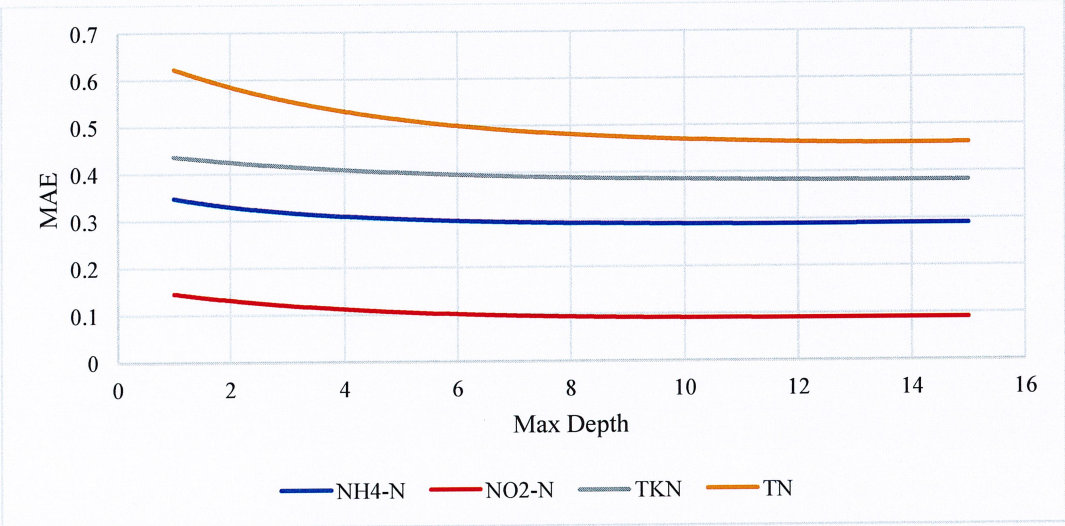


Fig. 23: The Effect of the Tree Depth on RF Performance

The results of the RF method for both training and testing are shown in Table (12), and these results demonstrate a considerable gap between the training and testing errors across all models, indicating the potential for overfitting in the model, which is supported by the MAE values of 0.1493 for NH_4-N, 0.0345 for NO_2-N, 0.1727 for TKN, and 0.1832 for TN in the training result. On the other hand, the MAE values for the

testing data were much higher than those for the training data, with values of 0.2873 for NH_4-N, 0.0878 for NO_2-N, 0.3812 for TKN, and 0.4598 for TN. Furthermore, the MAPE, RMSE, and R^2 values also suggest that overfitting is occurring, which aligns with the MAE results.

Table (12): Training and Testing Results of RF using Optimal Hyperparameters

	NH_4-N		NO_2-N		TKN		TN	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
MAE	0.1493	0.2873	0.0345	0.0878	0.1727	0.3812	0.1832	0.4598
MAPE	42.22%	81.14%	20.81%	52.96%	39.25%	86.51%	24.94%	62.56%
RMSE	0.2323	0.526	0.0509	0.1331	0.2550	0.6192	0.2554	0.6595
R^2	0.9266	0.2948	0.9651	0.6652	0.9665	0.2510	0.9588	0.6217

4.1.8. GBM Model Performance

For the GBM algorithm, the initial values of hyperparameters were set, shrinkage was set as 0.01 and n.minobsinnode was set as 10 and n.trees at 100. A trial-and-error approach was employed to improve the model performance. Three different distribution parameters (Gaussian, Laplace, t-distribution) were tried, with shrinkage values ranging from 0.001 to 0.1 with a step size of 0.005, and then from 0.1 to 1 with a step size of 0.05. Regarding the observations, a step size of one was used from 1 to 25, while the number of trees ranged from 50 to 100 with a step size of 50.

The MAE for Gaussian distribution parameter impact on NH_4-N was recorded at 0.3648, this being the lowest recorded MAE and the most successful result, with Laplace distribution showing a value of 0.3865 and t-distribution a value of 0.3985 as shown in Fig. 24. For NO_2-N readings, Laplace distribution was the best performer at 0.0997, with Gaussian and t-distribution providing only slightly higher levels of MAE at 0.115 and 0.1095 respectively. For TKN, the Laplace was also the best performer, with an MAE of 0.4052, only marginally less than the t-distribution and slightly better than the Gaussian distribution by around 0.02. All results

taken of TN are relatively similar, with Gaussian performing better, by a marginal amount. Overall, the widest variation in results are from NH_4-N, and the narrowest variation are from TN.

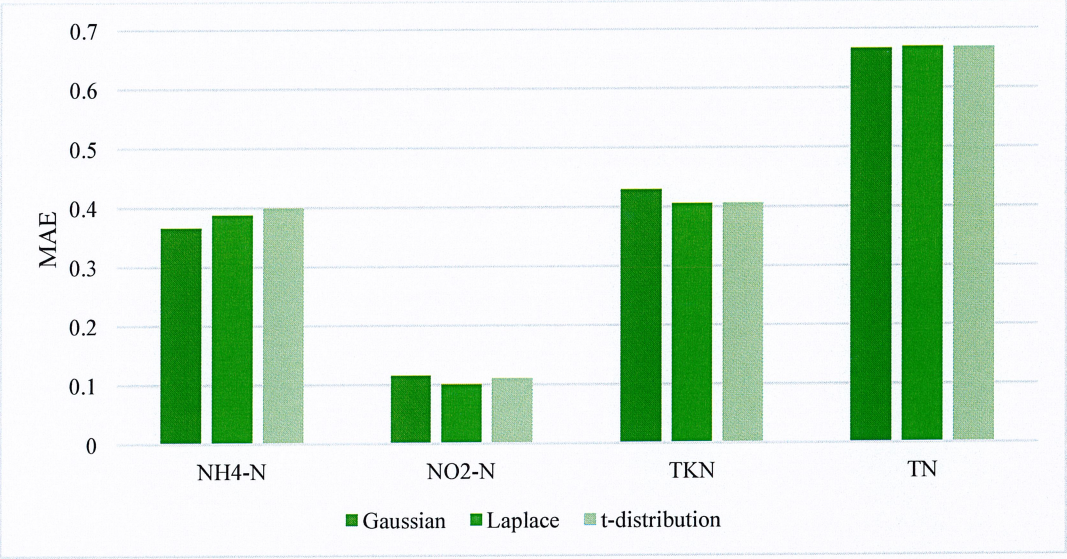
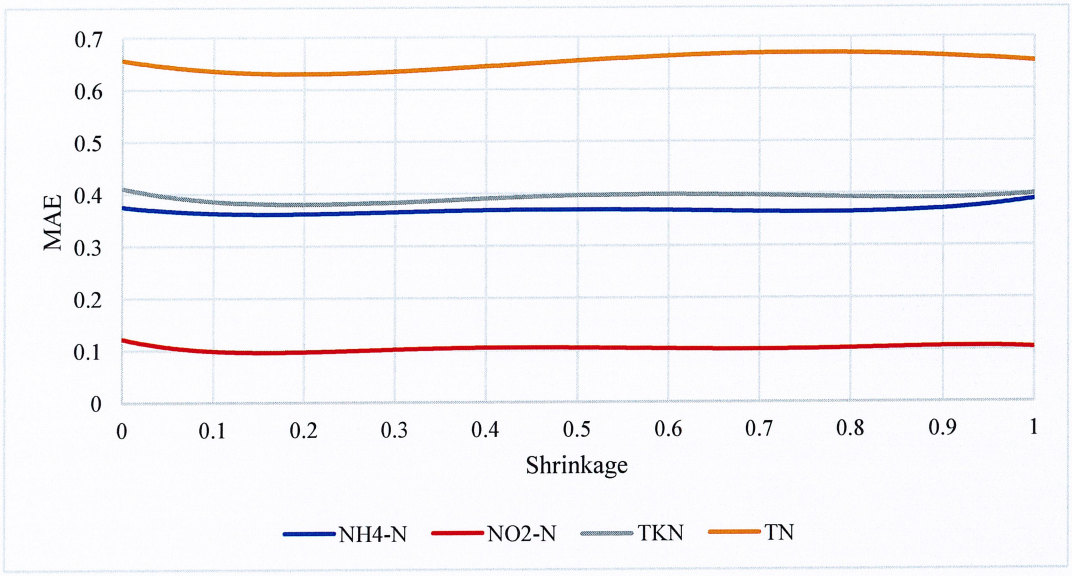
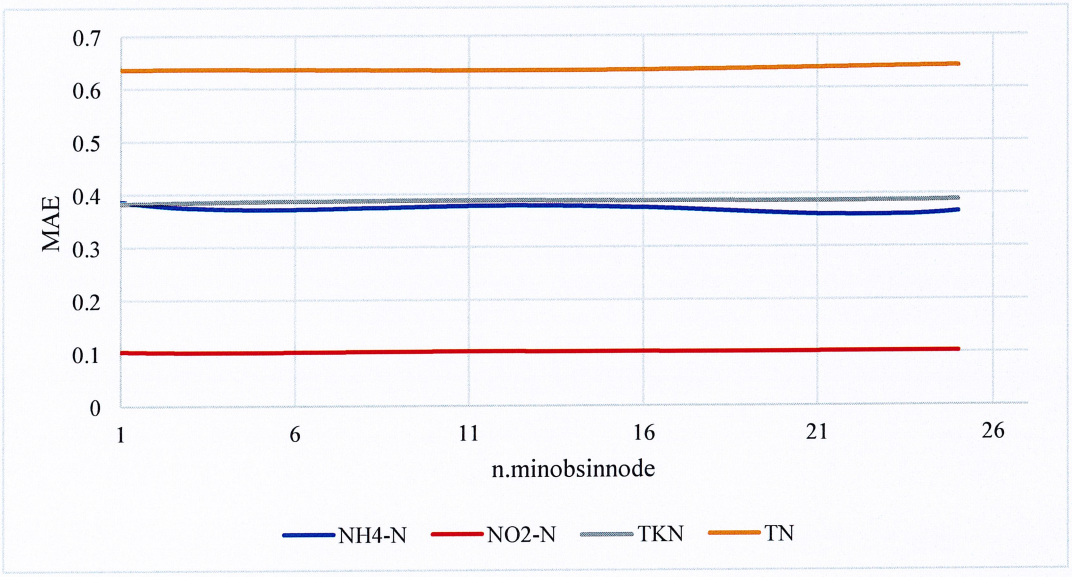


Fig. 24: The Effect of Distribution Function on MAE

Shrinkage did not show a wide variation in its results, as displayed in Fig. 25. For NH_4-N, MAE was lowest at a shrinkage rate of 0.055 (0.3611). For NO_2-N, the best possible MAE was found with a shrinkage rate of 0.45 (0.0992). TKN results show an optimum MAE at a shrinkage rate of 0.3 (0.382). The results of TN show the lowest MAE when the shrinkage rate is at 0.2 (0.6355). The n.minobsinnode has a range of 1 - 25, with a step size of 1, and n.trees used a range from 50-1000 with a step size of 50. The MAE of n.minobsinnode and n.trees was also measured. NH_4-N was the only parameter to show noticeable variation across all methods, with all others remaining reasonably stable.





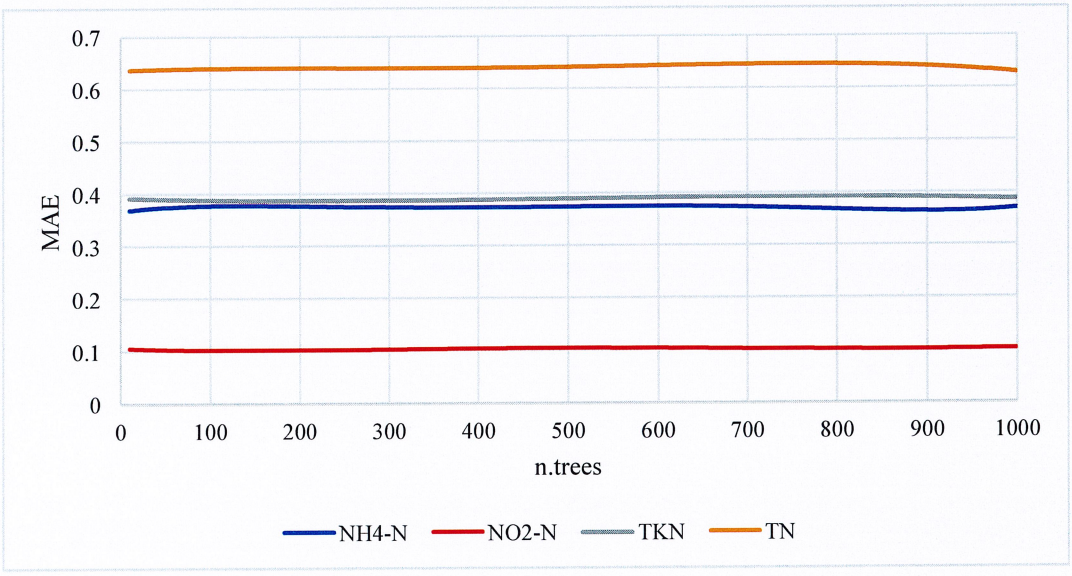


Fig. 25: The Effect of Shrinkage, n.minobsinnode, and n.trees on GBM Performance

The optimal hyperparameters for NH_4-N occur when shrinkage is at 0.85, the n.minobsinnode is at 24, and there are 500 trees. Regarding NO_2-N, the parameters that give the lowest MAE are with shrinkage at 0.45, n.minobsinnode at 10 and with 900 trees. The optimal hyperparameters for TKN are a shrinkage rate of 0.3, n.minobsinnode at 7, with 40 trees. TN however, shows optimal hyperparameters when shrinkage is at 0.2, n.minobsinnode is at 3, and 60 trees are used.

The results presented in Table (13) indicate a minimal discrepancy between the training and testing outcomes for all responses. The similarity between the errors suggests that there is no noteworthy difference between them, which implies that the model did not overfit. Additionally, the MAPE, RMSE, and R^2 values for both the training and testing data were alike, indicating that the model performed well when given new data. For more detailed information regarding the GBM predictions, please refer to Appendix (D), which presents the predicted and actual values of the testing data.

Table (13): Training and Testing Results of GBM using Optimal Hyperparameters

	NH_4-N		NO_2-N		TKN		TN	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
MAE	0.3098	0.3567	0.0682	0.0959	0.2854	0.3832	0.4592	0.6269
MAPE	71.69%	80.73%	42.85%	57.62%	84.85%	86.89%	58.96%	79.65%
RMSE	0.5248	0.6104	0.1289	0.1574	0.4879	0.5600	0.6928	0.8836
R^2	0.4185	0.2124	0.5987	0.4261	0.3879	0.1911	0.6582	0.4059

4.1.8.1. Variable Influence

According to the relative variable influence method of GBM illustrated in Fig. 26, the analysis indicated that the SSDT parameter had the most significant influence on the NH_4-N response, contributing to 56.90% of the total. OG followed in second place, representing 13.01%. This highlights a significant disparity in the predictive power of the variables. Conversely, the NH_3-N and SS parameters had a relatively minor influence, each contributing approximately 6%. In contrast, the

COB, conductivity, and alkalinity variables had minimal impact on NH_4-N.

Regarding the NO_2-N response, there was a relatively even influence distribution among the various parameters, except sulphide. Sulphide demonstrated a notably lower effect compared to the other predictors, accounting for just 1.72% of the total. This figure starkly contrasts the most influential parameter, IF, which had a significant impact of 15.77%. The remaining predictors had influence levels ranging from 6.47% to 13.87%.

The parameters for the TKN response show three holding a similar amount of influence, making up 62.76% in total. From the most influential parameters of IF, SS, and conductivity, there was a drop in the significance of the influencing parameters, with OG and pH representing around 8% of the total. Upon measuring the influential parameters for the TN response, it becomes evident that SSDT and SS are the most dominant factors, accounting for a substantial 87.20% of the total influence. The subsequent lowest level of impact was the IF attribute, with a comparatively modest share of 5.20%. Meanwhile, the remaining parameters had a negligible influence on the TN response.

The analysis of the NH_4-N and TN responses highlighted SSDT as the most influential attribute, accounting for over half of the total influence in each case. Conversely, the NO_2-N and TKN responses showed a more balanced influence distribution among the parameters.

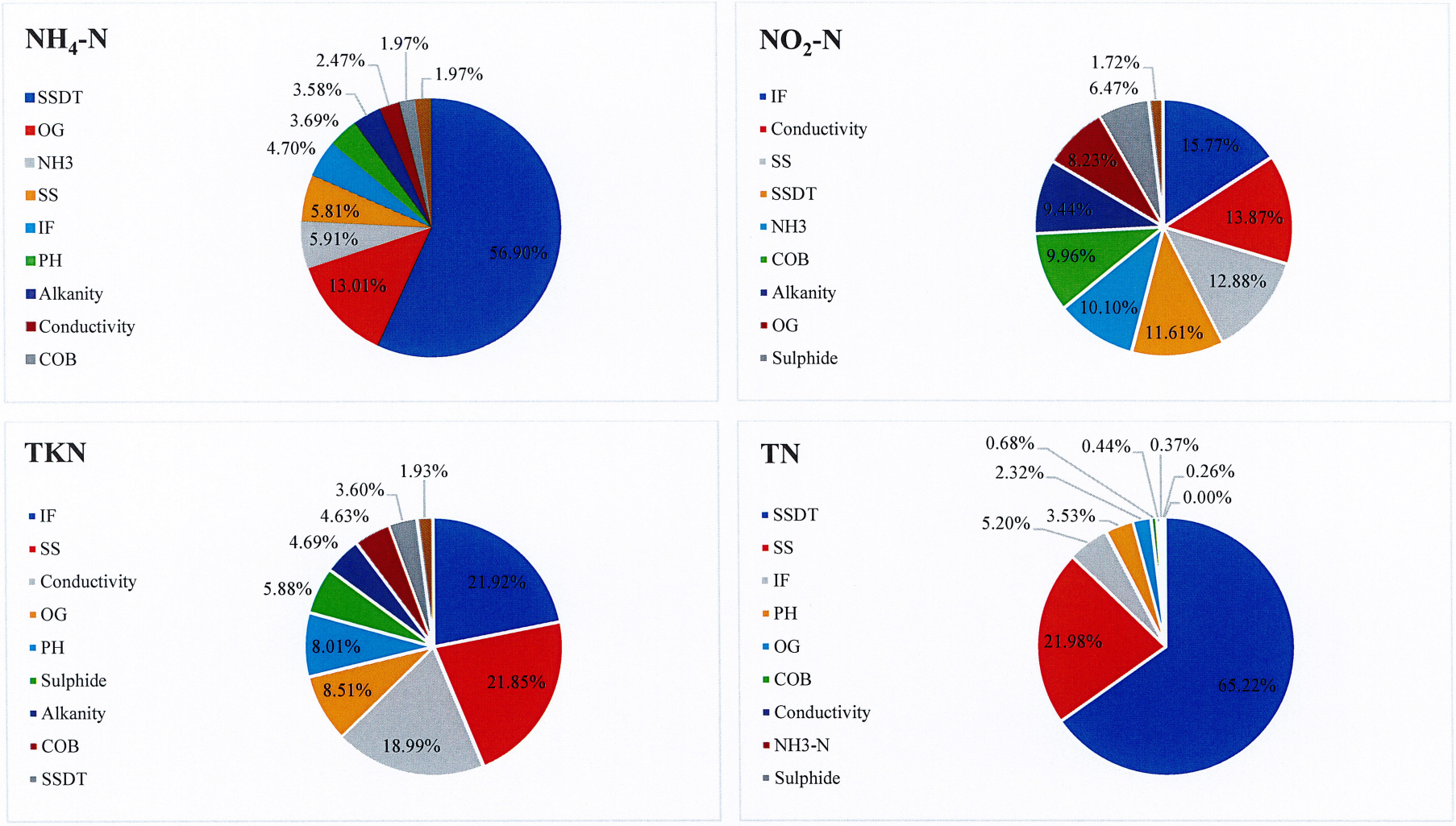


Fig. 26: The Influence of influent Variables on Each Response

4.1.9. Model Comparison

Fig. 27, displays the MAE, MAPE, RMSE, and R^2 results for various models using NH_4-N, NO_2-N, TKN, and TN as response variables. These models were compared to evaluate their performance. The SVR model exhibited the lowest MAE values for NH_4-N, with a value of 0.2759. The RF and KNN models followed closely, with values of 0.2873 and 0.3003, respectively. The MLP and GBM models were the least effective for NH_4-N while the SVR model was the most effective for NH_4-N, and the RF model exhibited best performance for all other response variables. For NO_2-N, the RF model outperformed the other models by a significant margin and was the only one that demonstrated a substantial improvement in MAE for NO_2-N compared to the other models. The RF model also performed well for TKN, recording a value of 0.3812, with the GBM and SVR models trailing closely. However, the MLP model performed the worst for TKN. Moreover, the RF model was the most effective for TN, recording a value of 0.4583, only slightly better than the SVR model. Remarkably, it significantly outperformed the MLP and KNN models, exhibiting the largest margin of improvement over the GBM model.

When evaluating the MAPE parameter, the results indicate no significant difference among the top performers compared to MAE. Similar to MAE, the analysis suggests that the RF model is the most effective for the NO_2-N, TKN, and TN responses, while the SVR model performs best for NH4-N. Additionally, based on R^2 values, it is evident that the RF model outperforms the others for all responses, with optimal values of 0.2948 for NH_4-N, 0.6652 for NO_2-N, 0.251 for TKN, and 0.6217 for TN. Furthermore, the RMSE testing results follow similar pattern as R^2, confirming that the RF model is the most effective.

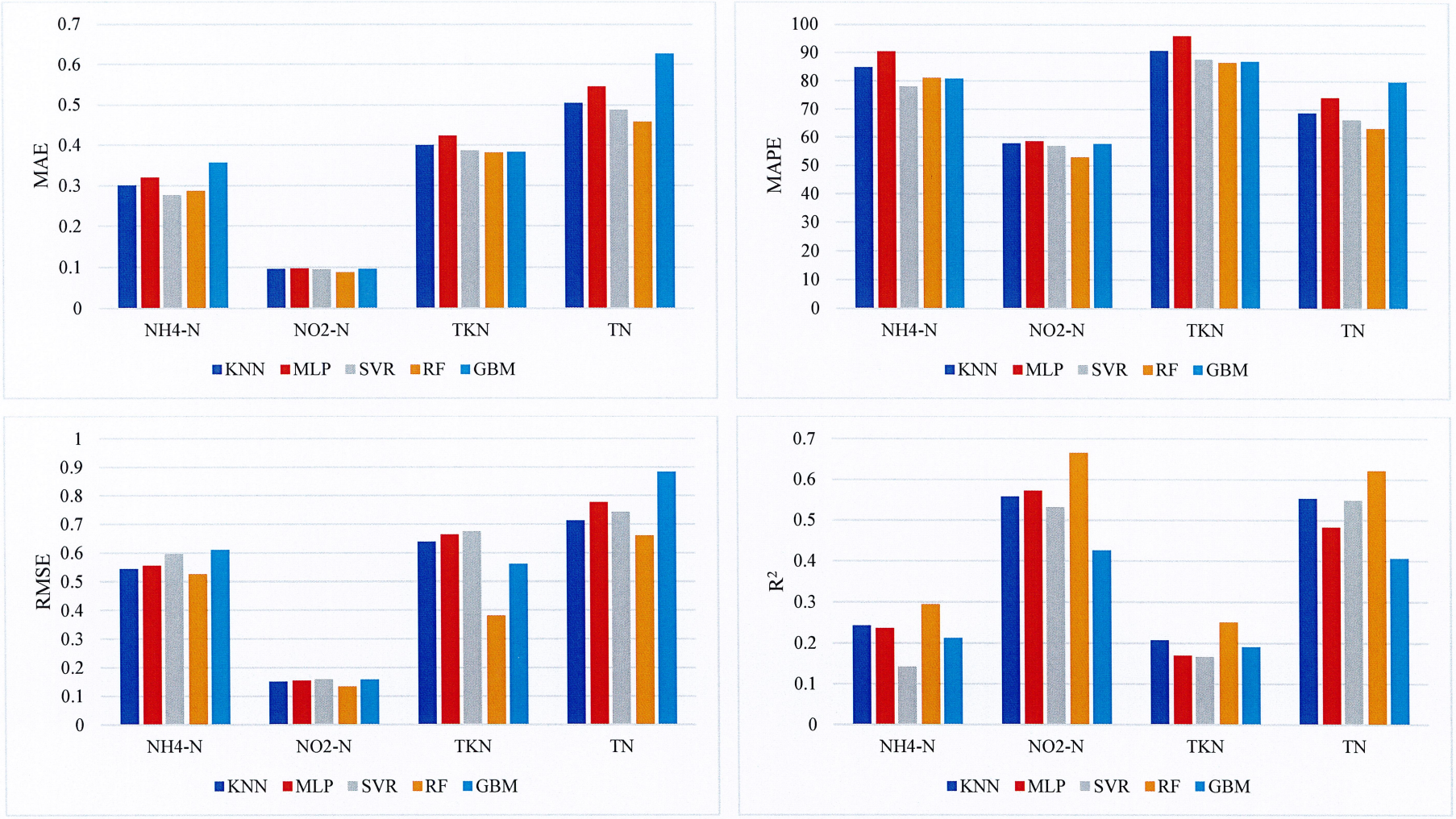


Fig. 27: The Evaluation Metrics of the Best Result out of Each Method

4.1.10. Impact of Outlier Removal

Fig. 28 displays the boxplots for IF and SS both before and after removing outliers using IQR and TSE range. Similar procedures were applied for the other variables, and they are presented in Appendix (B).

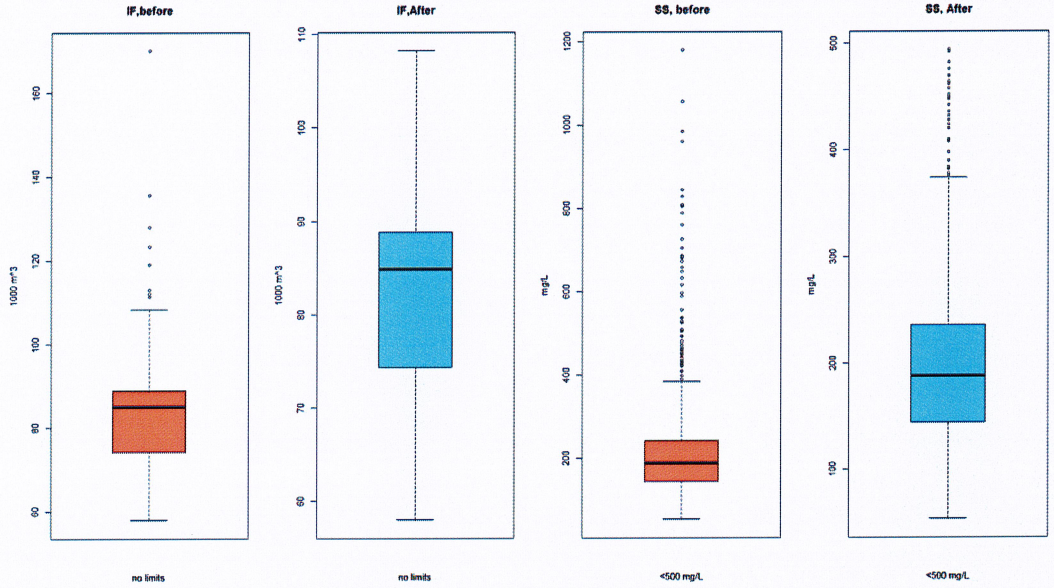


Fig. 28: Boxplots of IF and SS before and after Dealing with the Outliers

Table (14) displays the outcomes of the best model and their optimal hyperparameters, both before and after removing the outliers. The outcomes did not differ significantly, indicating that the outliers did not significantly impact the model's results. This is supported by the MAE values for the results with outliers, which were 0.2759 for NH_4-N, 0.0878 for NO_2-N, 0.3812 for TKN, and 0.4598 for TN. The MAE values for the results without outliers were very similar, with values of 0.2729 for NH_4-N, 0.0884 for NO_2-N, 0.3816 for TKN, and 0.4599 for TN. Moreover, the consistency of MAPE, RMSE, and R^2 values between both outcomes aligns with the MAE results. The reason for the minimal impact is that the outliers exist in only 5 out of 10 variables and in small quantities.

Table (14): Best Model Results with Optimal Hyperparameters before and after Outliers Removal

	NH_4-N		NO_2-N		TKN		TN	
	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers	With Outliers	Without Outliers
MAE	0.2759	0.2729	0.0878	0.0884	0.3812	0.3816	0.4598	0.4599
MAPE	77.92%	77.07%	52.96%	53.30%	86.51%	86.60%	62.56%	62.58%
RMSE	0.5967	0.5883	0.1331	0.1334	0.6192	0.6185	0.6595	0.6593
R^2	0.1426	0.1661	0.6652	0.6632	0.2510	0.2526	0.6217	0.6222

4.2. Discussion

This study aimed to predict the nitrification performance of WWTP using ML approaches to predict the effluent concentrations of NH_4-N, NO_2-N, TKN, and TN. Five different ML algorithms utilized including KNN, MLPs, SVR, RF, and GBM. The performance of these algorithms was evaluated using four metrics: MAE, MAPE, RMSE, and R^2. However, the MAE was preferred over other measurements for the following reasons: Firstly, it is not sensitive to outliers like RMSE. Therefore, MAE is simple relatively to understand and interpret. Secondly, MAE is better than MAPE when actual values are smaller because MAPE is a percentage-based metric that can be affected by small absolute errors. In contrast, MAE is unaffected by the actual values' scale, making it more appropriate for evaluating accuracy.

Starting with the results of the KNN model, the decrease in MAE values with increasing K is due to the algorithm's robustness to noise in the datasets. When K is small, the model is more prone to overfitting, which can result in higher MAE values. However, as K increases, the model becomes less sensitive to noise, resulting in better predictive performance and lower MAE values. This trade-off highlights the need to balance performance and model complexity. Despite its simplicity, the KNN model demonstrates excellent performance compared to the best

model, with relative differences in MAE of 8.84%, 9.34%, 4.75%, and 10.15% for NH_4-N, NO_2-N, TKN, and TN, respectively.

For MLP, even after changing multiple hyperparameters, it did not yield good results. This could be simply because the dataset might not be well-suited for MLP. For instance, if the relationships between the input features and the output variable are not complex, using a complex model like MLP might not be necessary, and simpler models like KNN or SVR might perform better.

For SVR, it gives the best MAE result in NH_4-N with 0.2759, and it shows competitive performance for NO_2-N, TKN, and TN, which was close to the best result. This is due to the fact that SVR is a kernel-based method that maps the input data into a high-dimensional feature space, where it becomes easier to find a linear boundary to separate the data.

The PUK kernel is also flexible in that it allows for tuning the distance parameter and exponent parameter, which can lead to better performance. Additionally, the PUK kernel has fewer hyperparameters than other kernel functions, making it less prone to overfitting and easier to tune.

Regarding the RF results, it performs well across all four parameters, ranking at the top for NO_2-N, TKN, and TN and second for NH_4-N. Despite overfitting, its ability to reduce variance through averaging predictions of multiple trees improves its performance on unseen data. Increasing the max depth does not significantly affect the MAE values, with stable results between max depths 11-15. Thus, a max depth of 11-15 may balance model complexity and predictive performance.

Lastly, GBM provides non-consistent results, where it performs the worst in NH_4-N and TN, but it is close to the best in NO_2-N and TKN. This may be because it is a more complex model than the others and

could face the same issue as MLPs. Overall, the models exhibit comparable performance as the difference between the best and worst results for all four parameters is relatively small.

According to the variable influence method, SSDT has the most significant impact on nitrification, and these results are consistent with Guo et al. (2015). The effluent and influent TN behaviour also support this during the seasons shown in Fig. 13 and Fig. 14, where TN shows the largest variation. The SSDT directly affects the growth and activity of microorganisms responsible for the nitrification process. The optimal temperature range for growth and activity is around 30°C (Adel et al., 2021). While higher temperatures can accelerate the nitrification process, microorganisms can become stressed and lose their ability to perform efficiently above a certain temperature threshold. Therefore, maintaining an optimal temperature range is crucial to ensure efficient and effective nitrification in WWT processes.

Furthermore, while the literature indicates that pH significantly impacts nitrification, its influence in this particular case appears to be minor, as its value remained relatively stable (between 7.3 and 8.0), as shown in Table (6).

The IF significantly impacts nitrification in WWT systems by affecting hydraulic and solids retention times. High IF rates can reduce the time for microorganisms to perform nitrification, while low IF may lead to higher removal efficiencies.

Managing flow rates is crucial to ensure optimal treatment efficiency and nitrification performance (Zhu et al., 2007). The OG variable also exhibits a moderate influence on NH_4-N. This can be attributed to the fact that OG can interfere with the biological processes in WWTPs, reducing their efficiency in removing ammonia. In addition, OG can cause

equipment blockages and operational issues, further affecting the treatment process.

Finally, the SS and high conductivity in wastewater can negatively affect the nitrification process. High levels of SS can physically interfere with the growth and activity of the microorganisms responsible for nitrification by reducing the availability of oxygen and acting as a substrate for heterotrophic microorganisms. Similarly, high conductivity levels can increase the osmotic stress on the nitrifying microorganisms, reducing their growth and activity. High conductivity can also lead to increased alkalinity, as illustrated in Fig. 11 with 0.42 ACC, further reducing the efficiency of the nitrification process. Therefore, controlling the SS and conductivity levels in wastewater is crucial for maintaining optimal conditions for nitrification in WWT (Metcalf & Eddy, 2003).

To compare the results of this study with previous studies, it is important to consider few rules. Firstly, all inputs should be influent variables because using effluent as an input is not logical. Secondly, it is also important to consider the size of the data because large data sets improve prediction accuracy by increasing input variety, better representing the population, improving generalization, and being more robust to noise.

Starting with NH_4-N prediction, the current study exhibited superior performance versus Al-Ghazawi and Alawneh (2021) study in terms of all three metrics: MAE (0.2759 versus 0.45), RMSE (0.526 versus 0.63), and R^2 (0.2948 versus 0.26). The current study's model produced lower errors and a higher R^2, indicating a better fit between the predicted and observed NH_4-N values. It is worth mentioning that the data size of this study is four times larger than that of Al-Ghazawi and Alawneh (2021), which makes the results more reliable.

Regarding TKN, the study by Manu & Thalla (2017) achieved better results than the present study in terms of both R^2 (0.83 versus 0.251) and slightly lower RMSE (0.232 versus 0.3812). However, it's worth noting that the present study had a significantly larger dataset (1729 versus 88), making it a more practical and robust model due to the larger data used for training and testing. Additionally, the influent TKN in the dataset was used as an input in the study conducted by Manu & Thalla (2017), and it had a cross-correlation of 0.92 with the effluent TKN. This creates an unfair comparison because the present study cannot access the influent TKN. Even if it were available, there would be no reason to predict the effluent TKN when the influent TKN, which has a cross-correlation of 0.92, is already available. Lastly, based on the R^2 values for predicting TN, Hvala & Kocijan (2020) demonstrated the highest predictive accuracy with an R^2 of 0.814, while Guo et al. (2015) exhibited the lowest predictive accuracy with an R^2 of 0.47. The current study falls between these two studies with an R^2 of 0.62, suggesting that it yielded acceptable results. However, further improvement can be achieved by employing hybrid methods to enhance the accuracy of TN prediction.

Chapter Five

Conclusion and Future Work

5.1. Conclusion

5.2. Future Work

Chapter Five
Conclusion and Future Work

5.1. Conclusion

The objective of this research is to utilize supervised ML models to predict nitrification performance and statistically analyze the behavior of influent wastewater characteristics and effluent data at Muharraq WWTP for the period (2018-2022). The study accurately predicted the concentrations of NH_4-N, NO_2-N, TKN, and TN in the effluent using ten influent variables.

According to statistical analysis, the influent and effluent TN were the variables that exhibited the most seasonal variation that could be attributed to the ambient temperature. Furthermore, the effluent NH_4-N and turbidity exceeded the acceptable range over the past four years, accounting for 29.96% and 49.6%, respectively.

The study found that SSDT and DDT have high correlation of 0.984 ACC. Conductivity is highly correlated with Chloride with 0.967 ACC as Chloride ions are a major contributor to conductivity. Surprisingly, all SS, COD, BOD, TN, and TP are also highly correlated, with ACC not below 0.825 between each relation. Therefore, SSDT, Conductivity, and SS were used instead of DDT, Chloride, COD, BOD, TN, and TP.

The variation of hyperparameters of the models contributed to improving the results, especially the K value in KNN and the kernels in SVR. The MAE values for the effluent exhibited the highest decrease between K values of 1 and 5, but with little reduction after K value 5. Additionally, the PUK kernel in SVR outperformed other kernels in all

outputs due to its flexibility in allowing tuning of both the distance parameter and exponent parameter, ultimately resulting in improved performance.

The results showed that the SVR model was the most effective in predicting NH_4-N, while the RF model outperformed other models in predicting NO_2-N, TKN, and TN with MAE values of 0.2759, 0.0878, 0.3812, and 0.4583, respectively. Overall, these results are more reliable than the previous studies based on the data size. Furthermore, it showed better performance for NH_4-N and was comparable to the best results in TKN and TN.

Additionally, the effluent NH_4-N and TN indicated that SSDT was the most significant influent variable, accounting for over half of the overall impact in both cases. Conversely, the responses of NO_2-N and TKN showed a more evenly distributed influence among the parameters. Finally, the outliers did not affect the results, probably because they exist in only 5 out of 10 variables and in small quantities.

This study could potentially lead to improvements in the understanding of efficiency and effectiveness of the nitrification process in Muharraq WWTP, which would have positive environmental impacts. Furthermore, the successful application of ML models demonstrates the potential of applying AI technology in the field of wastewater management to optimize performance.

5.1.1. Research Limitations and Weaknesses

Despite the promising results obtained from the supervised ML models used in this study, there were some limitations that should be considered. First, the attribute NO_3-N had to be removed from the analysis because all plant's recorded values were 5 mg/L. This had a negative impact on the study because NO_3-N is a significant factor in

5.2. Future Work

This research represents the first step towards more in-depth investigations of wastewater quality in Bahrain using ML and possibly lays the foundation for future studies. I recommend continued research to gain more comprehensive understanding of the various parameters that influence wastewater quality in the region. Additionally, the unique characteristics of Bahrain's wastewater quality suggest that other parameters could be tested beyond the scope of this study. To achieve this, more supervised ML modellings methods, such as ridge and lasso, can be utilized to reduce overfitting and improve feature selection. Hybrid approaches that combine multiple models may also be beneficial in achieving higher accuracy by leveraging the strengths of each component model while compensating for their weaknesses.

Al-Noaimi, M. A. (2005). *Water Use and Management in Bahrain: An Overview.* Damascus: Ministry of Municipalities and Agriculture Water Resources Directorate in Kingdom of Bahrain.

Alpaydın, E. (2010). *Introduction to Machine Learning.* Cambridge: MIT Press.

Al-Zubari, W. K., El-Sadek, A. A., Al-Aradi, M. J., & Al-Mahal, H. A. (2018). Impacts of climate change on the municipal water management system in the Kingdom of Bahrain: Vulnerability assessment and adaptation options. *Climate Risk Management, 20*, 1-16. doi:10.1016/j.crm.2018.02.002

Antoniou, P., Hamilton, J., Koopman, B., Jain, R., Holloway, B., Lyberatos, G., & Svoronos, S. A. (1990). Effect of temperature and ph on the effective maximum specific growth rate of nitrifying bacteria. *Water Research, 24*(1), 97-101. doi:10.1016/0043-1354(90)90070-M

Atkinson, A. C., & Hawkins, D. M. (1981). Identification of Outliers. *Biometrics, 37*(4), 1-3. doi:10.2307/2530182

B. Uestuen, W. M. (2006). Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel. . *Chemometrics and Intelligent Laboratory Systems, 81*, 29-40.

Baek, J., & Choi, Y. (2020). Deep neural network for predicting ore production by truck-haulage systems in open-pit mines. *Applied Sciences (Switzerland), 10*(5), 1-25. doi:10.3390/app10051657

Bagherzadeh, F., Mehrani, M. J., Basirifard, M., & Roostaei, J. (2021). Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms. *Journal of Water Process Engineering, 41*, 1-7. doi:10.1016/j.jwpe.2021.102033

Bekkari, N., & Zeddouri, A. (2019). Using artificial neural network for predicting and controlling the effluent chemical oxygen demand in wastewater treatment plant. *Management of Environmental Quality: An International Journal, 30*(3), 1-17. doi:10.1108/MEQ-04-2018-0084

Bersuder, P., Smith, A. J., Hynes, C., Warford, L., Barber, J. L., Losada, S., . . . Lyons, B. P. (2020). Baseline survey of marine sediments collected from the Kingdom of Bahrain: PAHs, PCBs, organochlorine pesticides, perfluoroalkyl substances, dioxins, brominated flame retardants and metal contamination. *Marine Pollution Bulletin, 161*, 1-10. doi:10.1016/j.marpolbul.2020.111734

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). *WEKA Manual.* Hamilton: University of Waikato.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Caicedo, J. R., Van Der Steen, N. P., Arce, O., & Gijzen, H. J. (2000). Effect of total ammonia nitrogen concentration and pH on growth rates of duckweed (Spirodela polyrrhiza). *Water Research, 34*(15), 3829-3835. doi:10.1016/S0043-1354(00)00128-7

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247–1250. doi:10.5194/gmd-7-1247-2014

Chang, C.-C. &.-J. (2011). LIBSVM: A library for support vector machines. . *ACM Transactions on Intelligent Systems and Technology (TIST), 2(3)*, 27.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society, 158*(3), 419-466. doi:10.2307/2983440

Chen, C., Liaw, A., & Brieman, L. (2004). *Using random forest to learn imbalanced data.* Berkeley: Statistics Department at the University of California.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). San Francisco: Association for Computing Machinery (ACM).

Cheng, M., Fang, F., Kinouchi, T., Navon, I. M., & Pain, C. C. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. *Journal of Hydrology, 590*, 1-13. doi:10.1016/j.jhydrol.2020.125376

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science, 110*(1), 195-239. doi:10.1007/s10994-020-05962-y

Cho, K. H. (2014). Achieving enhanced nitrification in communities of nitrifying bacteria in full-scale wastewater treatment plants via optimal temperature and pH. *Separation and Purification Technology, 132*, 697-703. doi:10.1016/j.seppur.2014.06.027

Ciudad, G., González, R., Bornhardt, C., & Antileo, C. (2007). Modes of operation and pH control as enhancement factors for partial nitrification with oxygen transport limitation. *Water Research, 41*(20), 4621– 4629. doi:10.1016/j.watres.2007.06.036

Coskuner, G., & Jassim, M. S. (2008). Development of a correlation to study parameters affecting nitrification in a domestic wastewater treatment plant. *Journal of Chemical Technology and Biotechnology, 83*(3), 299–308. doi:10.1002/jctb.1808

Cutler, A., Cutler, D. R., & Stevens, J. R. (2018). Random Forests. In C. Zhang, & Y. Ma, *Ensemble Machine Learning* (pp. 157-175). Springer.

Daims, H., Lebedeva, E. V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., . . . Bendinger. (2015). Complete nitrification by Nitrospira bacteria. *Nature, 528*(7583), 504-509. doi:10.1038/nature16461

El-Rawy, M., Abd-Ellah, M. K., Fathi, H., & Ahmed, A. K. (2021). Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. *Journal of Water Process Engineering, 44*. doi:10.1016/j.jwpe.2021.102380

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association, 87*(417), 178-183. doi:10.1080/01621459.1992.10475190

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232. doi:10.1214/aos/1013203451

Garrido, J. M., Van Benthum, W. A., Van Loosdrecht, M. C., & Heijnen, J. J. (1997). Influence of dissolved oxygen concentration on nitrite accumulation in a biofilm airlift suspension reactor. *Biotechnology and Bioengineering, 53*(2), 168–178. doi:10.1002/(SICI)1097-0290(19970120)53:2<168::AID-BIT6>3.0.CO;2-M

Gnida, A., Wiszniowski, J., Felis, E., Sikora, J., Surmacz-Górska, J., & Miksch, K. (2016). The effect of temperature on the efficiency of industrial wastewater nitrification and its (geno)toxicity. *Archives*

of Environmental Protection, 42(1), 27–34. doi:10.1515/aep-2016-0003

Gubry-Rangin, C., Hai, B., Quince, C., Engel, M., Thomson, B. C., James, P., . . . Nicol, G. W. (2011). Niche specialization of terrestrial archaeal ammonia oxidizers. *Proceedings of the National Academy of Sciences of the United States of America, 108*(52), 21206-21211. doi:10.1073/pnas.1109000108

Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y. M., Park, J. p., . . . Cho, K. H. (2015). Prediction of effluent concentration in a wastewater treatment plant using machine learning models. *Journal of Environmental Sciences (China), 32*, 90-101. doi:10.1016/j.jes.2015.01.007

Guo, J., Peng, Y., Huang, H., Wang, S., Ge, S., Zhang, J., & Wang, Z. (2010). Short- and long-term effects of temperature on partial nitrification in a sequencing batch reactor treating domestic wastewater. *Journal of Hazardous Materials, 179*(1-3), 471–479. doi:10.1016/j.jhazmat.2010.03.027

Guo, X., Kim, J. H., Behera, S. K., & Park, H. S. (2008). Influence of dissolved oxygen concentration and aeration time on nitrite accumulation in partial nitrification process. *International Journal of Environmental Science and Technology, 5*(4), 527-534. doi:10.1007/BF03326050

Hamoda, M. F., Al-Ghusain, I. A., & Hassan, A. H. (1999). Integrated wastewater treatment plant performance evaluation using artificial neural networks. *Water Science and Technology, 40*(7), 55-65. doi:10.1016/S0273-1223(99)00584-3

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and Additive Trees. In *The Elements of Statistical Learning* (pp. 337-387). Springer.

Hodge, V. J., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review, 22*(2), 85-126. doi:10.1007/s10462-004-4304-y

Hvala, N., & Kocijan, J. (2020). Design of a hybrid mechanistic/Gaussian process model to predict full-scale wastewater treatment plant effluent. *Computers and Chemical Engineering, 140*, 1-12. doi:10.1016/j.compchemeng.2020.106934

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679 – 688. doi:10.1016/j.ijforecast.2006.03.001

Janahi, E. M., Mustafa, S., Parkar, S. F., Naser, H. A., & Eisa, Z. M. (2020). Detection of enteric viruses and bacterial indicators in a sewage treatment center and shallow water bay. *International Journal of Environmental Research and Public Health, 161*, 1-13. doi:10.1016/j.marpolbul.2020.111734

Liaw, A., & Wiener, M. (2002). Classification and Regression with Random Forest. *R News, 2*(3), 18-22.

Liu, G. (2012). *NITRIFICATION PERFORMANCE OF ACTIVATED SLUDGE UNDER LOW DISSOLVED OXYGEN CONDITIONS.* MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY.

Liu, G., Xu, X., Zhu, L., Xing, S., & Chen, J. (2013). Biological nutrient removal in a continuous anaerobic-aerobic-anoxic process treating synthetic domestic wastewater. *Chemical Engineering Journal, 225*, 223-229. doi:10.1016/j.cej.2013.01.098

Ly, Q. V., Truong, V. H., Ji, B., Nguyen, X. C., Cho, K. H., Ngo, H. H., & Zhang, Z. (2022). Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants. *Science of*

The Total Environment, 832, 1-12. doi:10.1016/j.scitotenv.2022.154930

Maier, H. R., Morgan, N., & Chow, C. W. (2004). Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. *Environmental Modelling and Software, 19*(5), 485-494. doi:10.1016/S1364-8152(03)00163-4

Manu, D. S., & Thalla, A. K. (2017). Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the removal of Kjeldahl Nitrogen from wastewater. *Applied Water Science, 7*(7), 3783-3791. doi:10.1007/s13201-017-0526-4

Mehrani, M.-J., Bagherzadeh, F., Zheng, M., Kowal, P., Sobotka, D., & Mąkinia, J. (2022). Application of a hybrid mechanistic/machine learning model for prediction of nitrous oxide (N2O) production in a nitrifying sequencing batch reactor. *Process Safety and Environmental Protection, 162*, 1015-1024. doi:10.1016/j.psep.2022.04.058

Metcalf, E., & Eddy, H. (2003). *Wastewater engineering: treatment and reuse.* New York: McGraw-Hill.

Naser, H. A. (2020). Technology for environmental management in Small Island Developing States: the case of Bahrain. *Current Opinion in Environmental Sustainability, 59*, 1-8. doi:10.1016/j.cosust.2022.101219

Natekin, A., & Knoll, A. (2013). Gradient Boosting Machines, a Tutorial. *Frontiers in Neurorobotics, 7*(21), 21-36. doi:10.3389/fnbot.2013.00021

Okabe, S., Aoi, Y., Satoh, H., & Suwa, Y. (2014). Nitrification in Wastewater Treatment. In *Nitrification* (pp. 405-433).

Ontañón, S. (2020). An overview of distance and similarity functions for structured data. *Artificial Intelligence Review, 53*(7), 5309-5351. doi:10.1007/s10462-020-09821-w

Park, Y. S., & Lek, S. (2016). Developments in Environmental Modelling. In *Developments in Environmental Modelling* (Vol. 28, pp. 123-140). doi:10.1016/B978-0-444-63623-2.00007-4

Rachmatullah, M. I., Santoso, J., & Surendro, K. (2021). Determining the number of hidden layer and hidden neuron of neural network for wind speed prediction. *PeerJ Computer Science, 7*, 1-19. doi:10.7717/PEERJ-CS.724

Reuters, T. (1992). Retrieved from ISI Web of Science: https://mjl.clarivate.com/search-results

Rodriguez-Sanchez, A., Gonzalez-Martinez, A., Martinez-Toledo, M. V., Garcia-Ruiz, M. J., Osorio, F., & Gonzalez-Lopez, J. (2014). The Effect of Influent Characteristics and Operational Conditions over the Performance and Microbial Community Structure of Partial Nitritation Reactors. *Water, 6*(1), 1905-1924. doi:10.3390/w6071905

S.K. Shevade, S. K. (1999). Improvements to the SMO Algorithm for SVM Regression. *IEEE Transactions on Neural Networks* .

Salgot, M., & Folch, M. (2018). Wastewater treatment and water reuse. *Current Opinion in Environmental Science and Health, 2*, 64-74. doi:10.1016/j.coesh.2018.03.005

Segal, M. R. (2004). Machine Learning Benchmarks and Random Forest Regression. *Biostatistics, 5*(3), 363-376. doi:10.1093/biostatistics/5.3.363

SINGH, A., KOTIYAL, V., SHARMA, S., NAGAR, J., & LEE, C.-C. (2020). A Machine Learning Approach to Predict the Average Localization Error With Applications to Wireless Sensor

Networks. *IEEE Access, 8*, 208253-208263. doi:10.1109/ACCESS.2020.3038645

Smola, A. J. (2004). A tutorial on support vector regression. *Statistics and Computing, 14(3)*, 199-222.

Van Kessel, M. A., Speth, D. R., Albertsen, M., Nielsen, P. H., Op Den Camp, H. J., Kartal, B., . . . Lücker, S. (2015). Complete nitrification by a single microorganism. *Nature, 528*(7583), 1-5. doi:10.1038/nature16459

Vapnik, V. N. (1998). Statistical learning theory. *New York: Wiley.*

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model . *BMC Bioinformatics, 7*(1), 1-8. doi:10.1186/1471-2105-7-91

Wang, B., Sharma, J., Chen, J., & Persaud, P. (2021). Ensemble Machine Learning Assisted Reservoir Characterization Using Field Production Data–An Offshore Field Case Study. *Energies, 14*(1052), 1-20. doi:10.3390/en14041052

Ward, B. B. (2008). Nitrification in Marine Systems. In D. D. Capone, D. D. Bronk, M. R. Mulholland, & E. J. Carpenter, *Nitrogen in the Marine Environment* (pp. 199-261). Amsterdam: Elsevier. doi:10.1016/B978-0-12-372522-6.00005-0

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79-82. doi:10.3354/cr030079

Ye, Z., Yang, J., Zhong, N., Tu, X., Jia, J., & Wang, J. (2020). Tackling environmental challenges in pollution controls using artificial intelligence: A review. *Science of the Total Environment, 699*, 1-28. doi:10.1016/j.scitotenv.2019.134279

Appendix (A): Sample of Influent and Effluent Data of Muharraq WWTP

Table (15): Sample of Influent Laboratory Measurements Data of Muharraq WWTP

Sampling Date	IF	Sample Location	Sample Type	DDT	SSDT	pH	SS	Conductivity	S	OG	COD	BOD	NH_3-N	TN	TP	Chloride	Alkalinity	TDS	COB
Unit	m^3/day	---	---	^0C	^0C	---	mg/L	µS/cm	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	---
TSE Limit	---	---	---	15-37	20-35	6 - 10.3	<500	<12,000	<80.0	<340	220- 900	115 - 450	---	15- 40	3-15	---	---	---	---
1-Jan-18	69,500	AS-501	Composite	29.0	29.6	7.8	204	3,340	4.3	38.4	463	177	21.9	40.6	5.3	768	296	---	2.61
2-Jan-18	69,679	AS-501	Composite	29.5	29.3	7.6	194	3,780	3.2	21.3	408	180	23.5	41.4	4.9	886	290	---	2.27
3-Jan-18	69,868	AS-501	Composite	29.2	29.6	7.7	230	3,680	4.2	23.5	447	200	23.9	40.8	5.7	858	296	---	2.24
4-Jan-18	70,368	AS-501	Composite	29.3	29.4	7.6	200	3,570	3.8	39.3	428	162	19.8	38.7	5.0	824	279	---	2.64
5-Jan-18	71,849	AS-501	Composite	29.3	29.5	7.4	334	3,480	4.8	30.3	634	262	22.0	43.9	4.9	796	290	---	2.42
6-Jan-18	70,742	AS-501	Composite	29.1	29.3	7.5	342	3,600	5.6	22.5	624	244	24.3	47.6	7.2	775	285	1,904	2.56
7-Jan-18	71,556	AS-501	Composite	29.4	29.7	7.5	222	3,530	6.6	29.7	507	184	23.0	41.4	5.5	789	282	---	2.75
8-Jan-18	69,653	AS-501	Composite	29.2	29.3	7.6	234	3,250	4.2	25.6	509	201	25.8	44.0	5.4	789	274	---	2.53
9-Jan-18	69,510	AS-501	Composite	27.9	28.0	7.6	202	3,320	3.4	24.1	471	209	23.7	40.0	5.6	768	290	---	2.25
10-Jan-18	69,741	AS-501	Composite	29.0	29.0	7.6	274	3,240	3.2	35.1	553	202	20.0	40.9	6.9	768	282	---	2.74
11-Jan-18	67,822	AS-501	Composite	28.8	29.0	7.6	194	3,080	3.4	32.7	435	171	26.1	39.3	5.2	727	290	---	2.54
12-Jan-18	70,303	AS-501	Composite	29.0	29.1	7.5	232	3,220	5.6	24.3	499	199	23.0	40.7	5.7	699	298	---	2.51
13-Jan-18	69,535	AS-501	Composite	28.8	29.0	7.5	266	3,250	4.0	42.3	500	194	22.6	41.6	5.9	755	285	1,856	2.58
14-Jan-18	69,276	AS-501	Composite	28.9	28.6	7.7	316	3,210	4.0	34.2	597	217	26.5	47.3	7.5	755	308	---	2.75
15-Jan-18	68,366	AS-501	Composite	28.8	29.2	7.7	210	3,140	3.6	38.7	452	190	23.8	41.5	5.7	720	290	---	2.38
16-Jan-18	69,095	AS-501	Composite	29.0	28.9	7.7	254	3,190	4.0	29.4	512	196	21.7	39.3	5.8	720	280	---	2.61
17-Jan-18	68,539	AS-501	Composite	28.5	28.9	7.6	260	3,310	4.0	27.9	531	204	23.4	44.3	6.5	734	295	---	2.61
18-Jan-18	70,065	AS-501	Composite	28.9	28.9	7.8	220	3,220	4.4	23.8	453	193	25.1	46.4	6.2	716	298	---	2.35

Table (16): Sample of Effluent Laboratory Measurements Data of Muharraq WWTP

Sampling Date	Sample Location	Sample Type	pH	Turbidity	OG	TSS	COD	BOD	NH_4-N	NO_3-N	NO_2-N	TKN	TN	DO	Conductivity	Alkalinity	ET
Unit	---	---	---	NTU	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	µS/cm	mg/L	^{0}C
TSE Discharge Limit	---	---	6.5 – 8.5	<2	<5	<10	<40	<10	<1	<10	<1	<5	<10	>2	---	---	---
1-Jan-18	AS-504	Composite	7.5	2	1	3	15	2	1.2	5	0.3	2.1	3.5	8.8	3,260	178	29
2-Jan-18	AS-504	Composite	7.5	2	1	3	17	2	1.3	5	0.3	2	3.4	9.2	3,270	176	28.5
3-Jan-18	AS-504	Composite	7.6	2	1	3	17	2	1.2	5	0.4	2	3.6	9.2	3,320	180	28.4
4-Jan-18	AS-504	Composite	7.5	2	1	4	16	2	0.6	5	0.2	1.4	3.2	9	3,460	172	28.3
5-Jan-18	AS-504	Composite	7.6	2	1	3	19	2	0.5	5	0.2	1.6	3.1	9.1	3,420	172	28.4
6-Jan-18	AS-504	Composite	7.6	2	1	4	17	2	0.9	5	0.3	2.1	3.4	8.9	3,280	179	28.6
7-Jan-18	AS-504	Composite	7.5	2	1	3	19	2	1.2	5	0.3	1.8	3.5	9.1	3,400	179	28.6
8-Jan-18	AS-504	Composite	7.5	2	1	3	17	2	1.4	5	0.4	1.9	3.7	9.2	3,460	183	28.5
9-Jan-18	AS-504	Composite	7.5	2	1	2	17	2	0.7	5	0.3	1.6	3.4	8.9	3,400	173	28.1
10-Jan-18	AS-504	Composite	7.5	2	1	2	16	2	0.6	5	0.3	1	3.2	9.2	3,370	172	28.2
11-Jan-18	AS-504	Composite	7.5	2	1	4	17	2	0.6	5	0.3	1.4	3.1	8.9	3,210	168	28
12-Jan-18	AS-504	Composite	7.5	2	1	3	18	2	0.6	5	0.3	1.1	3.1	9.1	3,240	177	27.8
13-Jan-18	AS-504	Composite	7.4	2	1	3	18	2	0.9	5	0.3	1.6	3.3	8.9	3,150	177	27.9
14-Jan-18	AS-504	Composite	7.3	2	1	3	18	2	1	5	0.3	1.8	3.4	9	3,100	177	28
15-Jan-18	AS-504	Composite	7.5	2	1	2	18	2	1.2	5	0.4	2.1	3.8	9.1	3,110	170	28.3
16-Jan-18	AS-504	Composite	7.5	2	1	3	19	2	0.9	5	0.4	1.3	3.3	9.2	3,130	175	28.2
17-Jan-18	AS-504	Composite	7.5	2	1	3	17	2	0.6	5	0.3	1.3	3.2	9.1	3,120	171	27.6
18-Jan-18	AS-504	Composite	7.5	2	1	3	18	2	0.5	5	0.2	1.2	3.3	8.5	3,250	175	27.6

Appendix (B): Boxplots of Influent Parameters

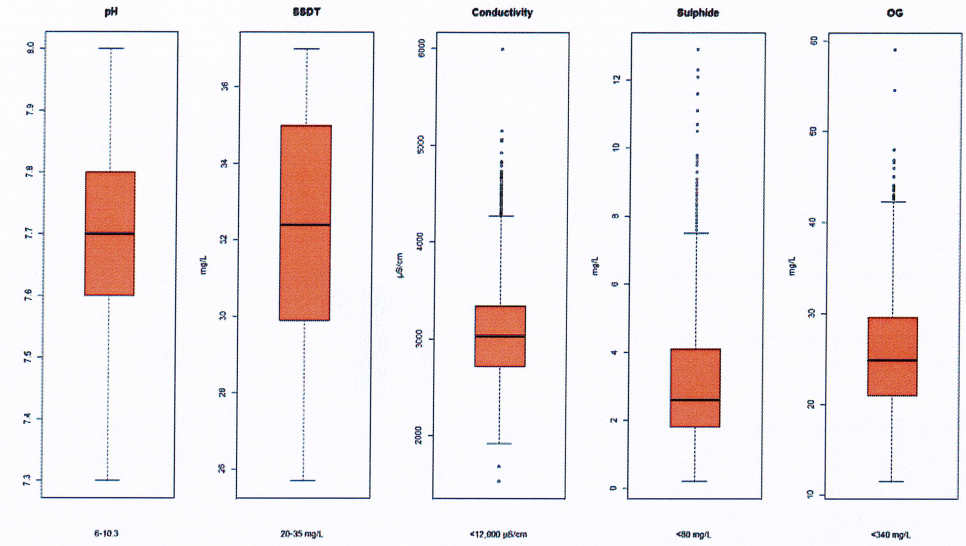


Fig. 29: The Boxplots of Different Variables that their Outliers below the TSE Range

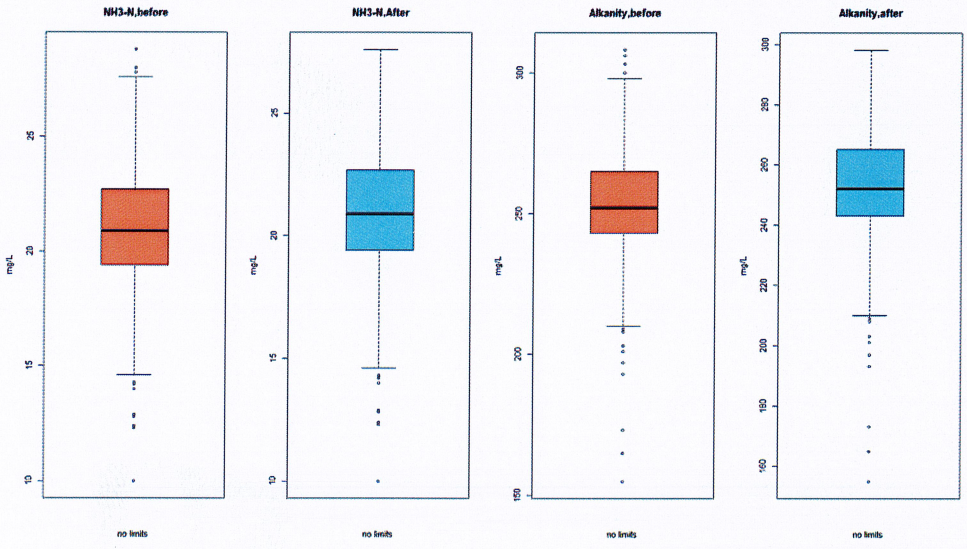


Fig. 30: The Boxplots of NH_3-N and Alkalinity before and after Removing the Outliers

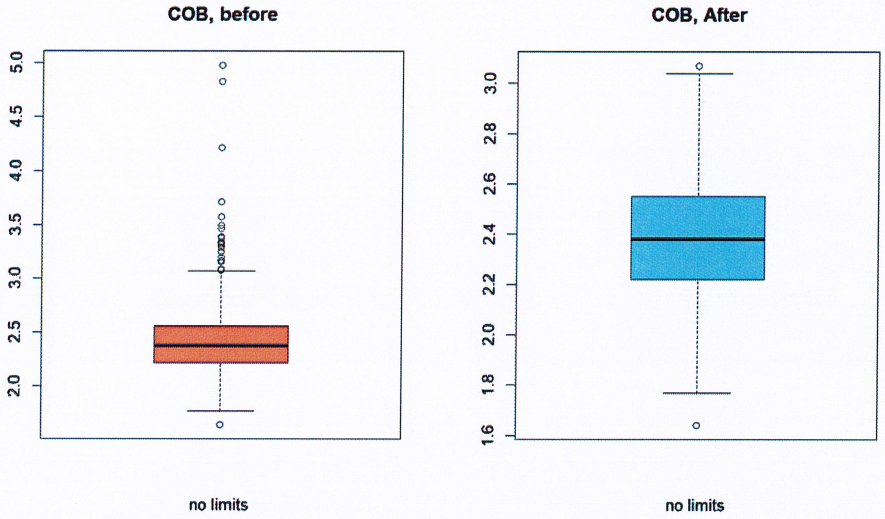


Fig. 31: The Boxplots of COB before and after Removing the Outliers

Appendix (C): Statistical Analysis

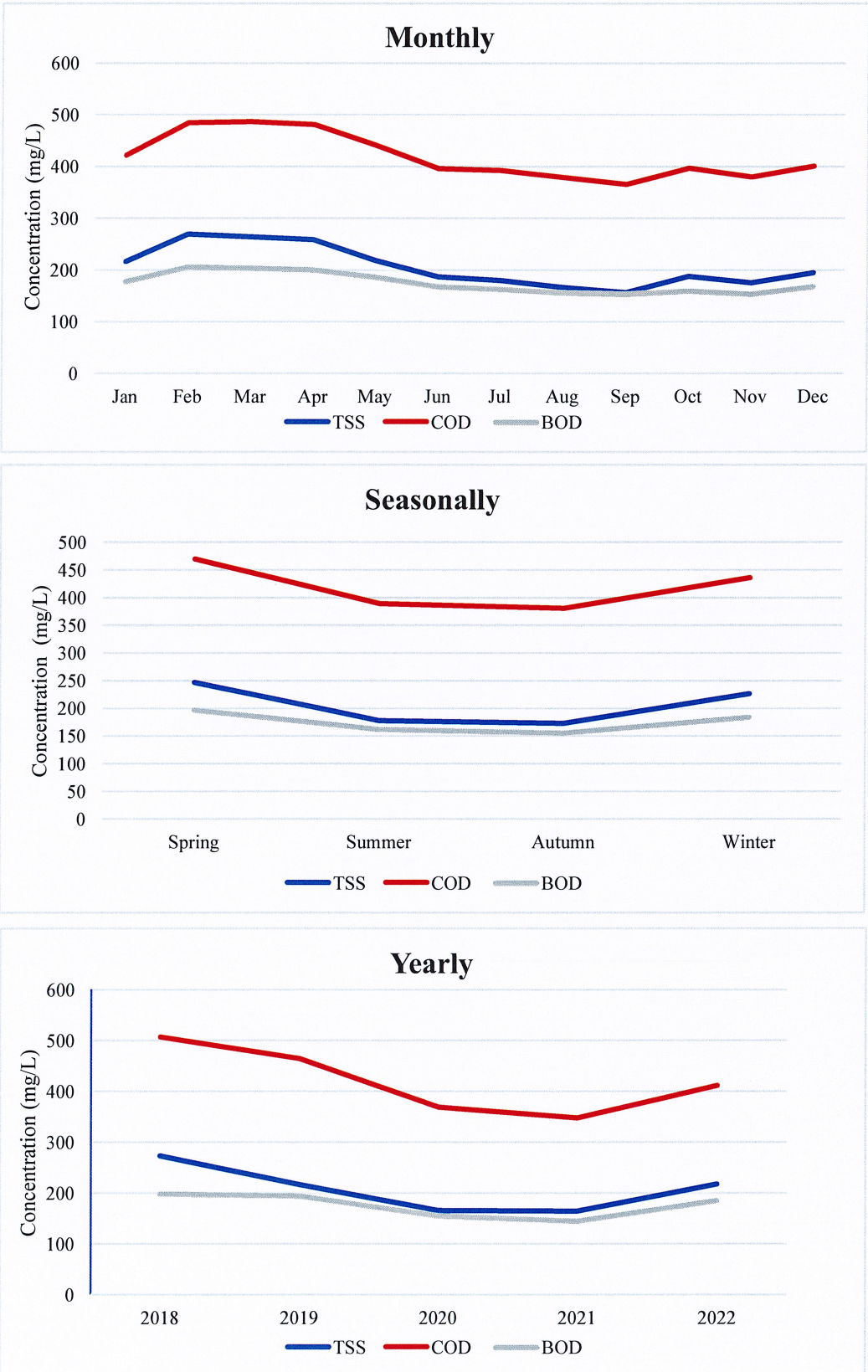
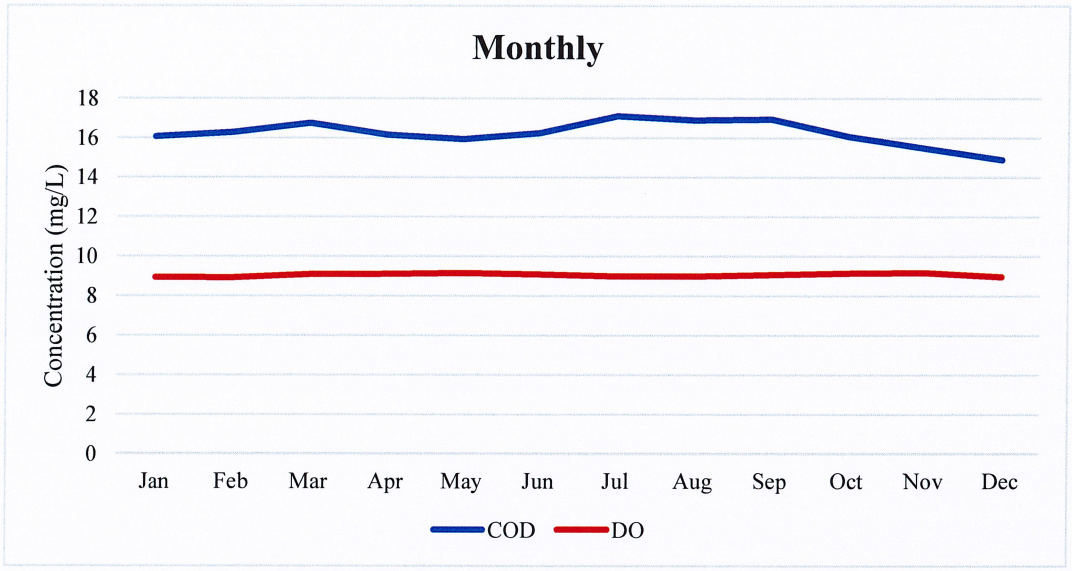
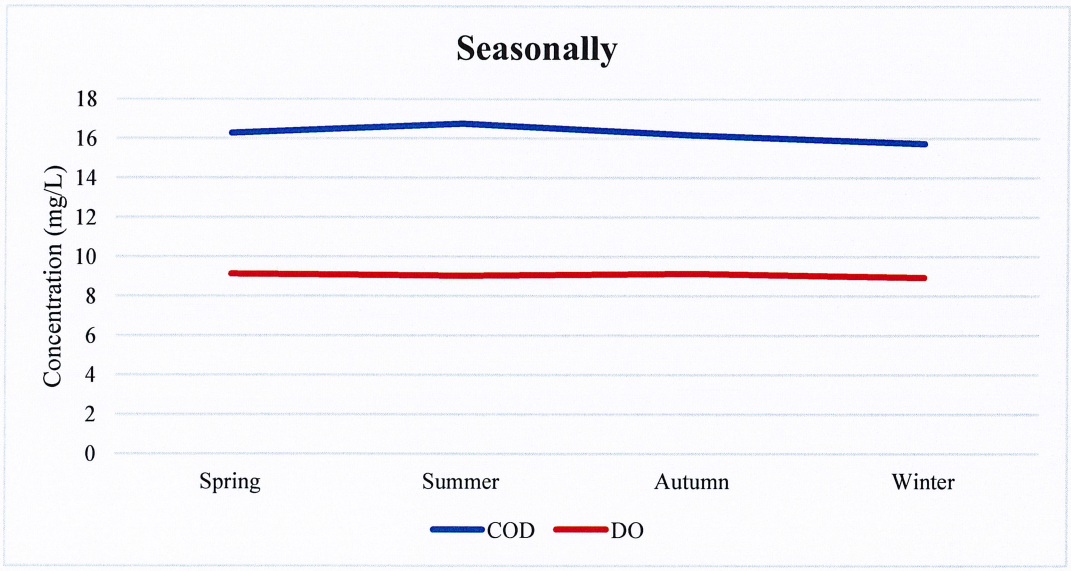


Fig. 32: Monthly, Seasonal, and Yearly Behavior of Influent Variables





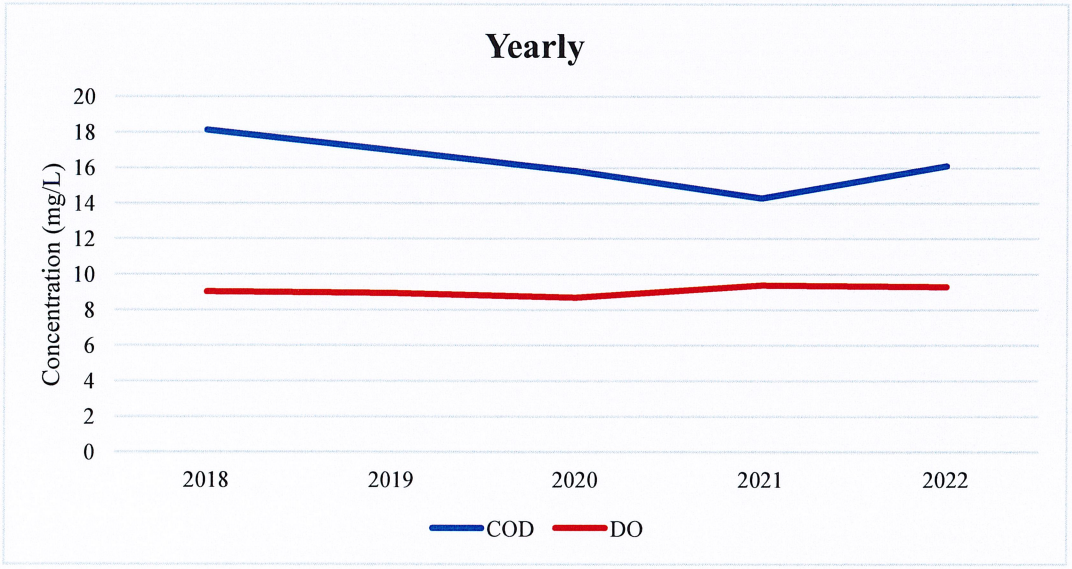


Fig. 33: Monthly, Seasonal, and Yearly Behavior of Effluent Variables

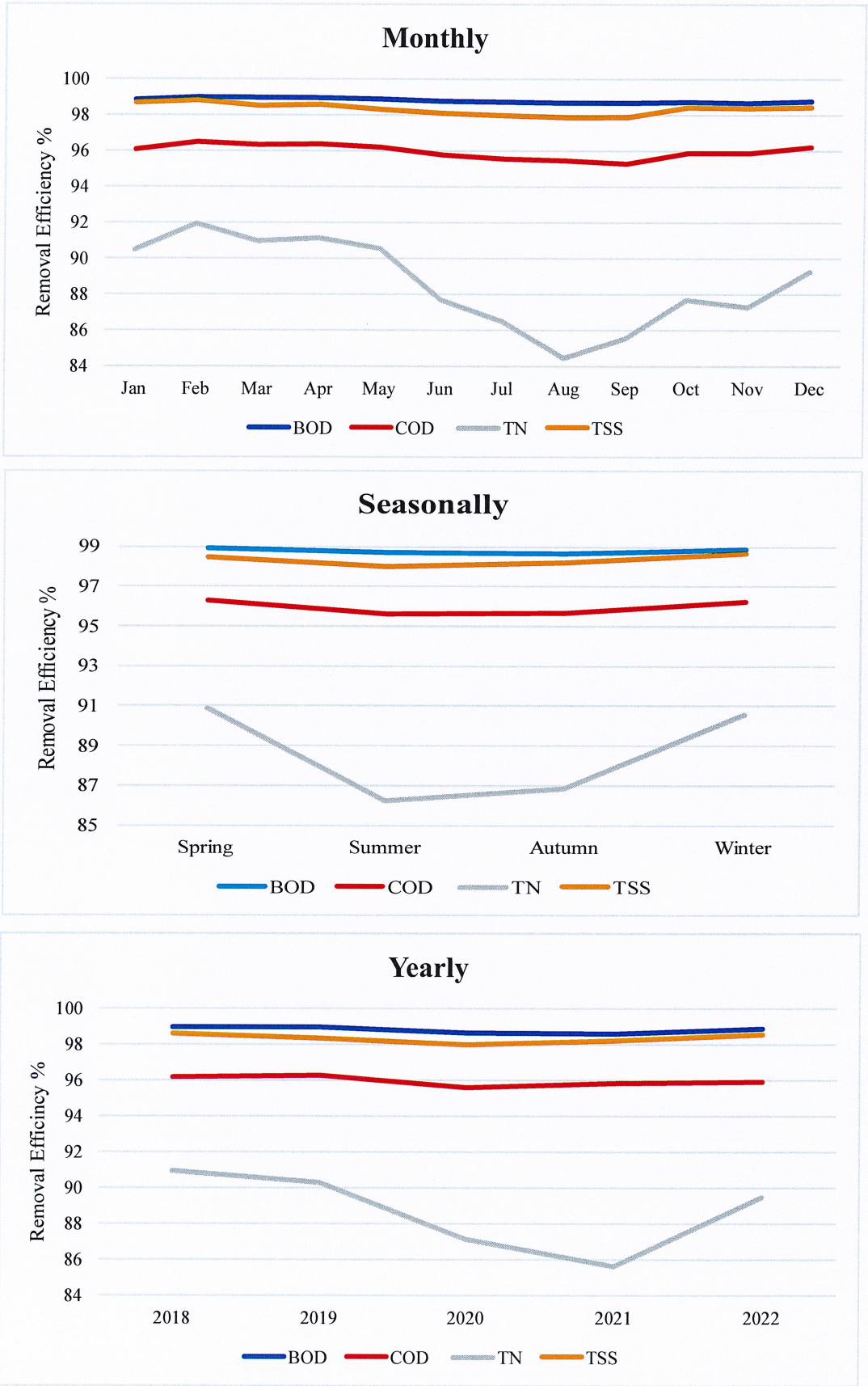


Fig. 34: Monthly, Seasonal, and Yearly Behavior of Removal Efficiency of Different Variables

Appendix (D): Predicted versus Actual values of the GBM Model

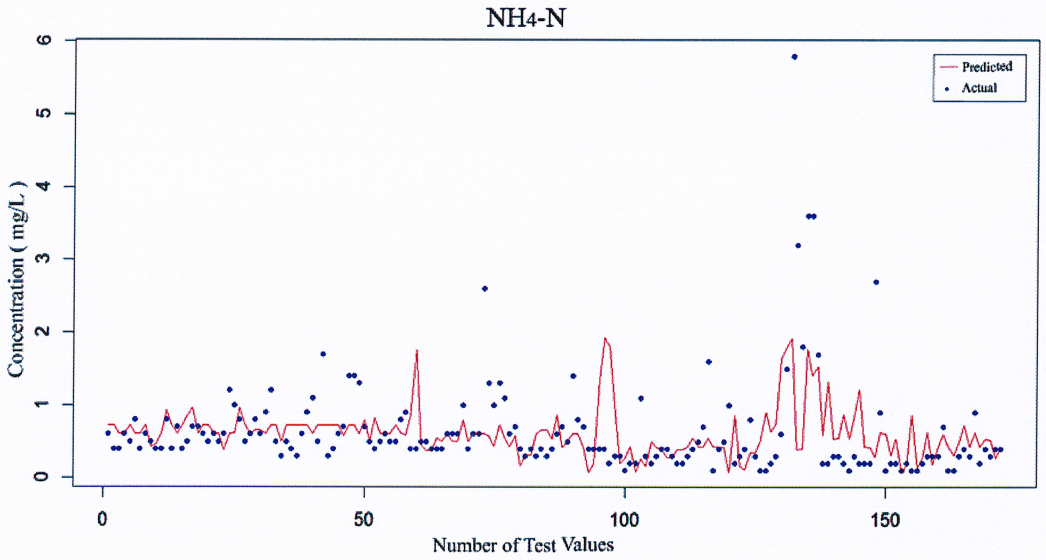


Fig. 35: Predicted vs True Values of the Test Data of NH_4-N

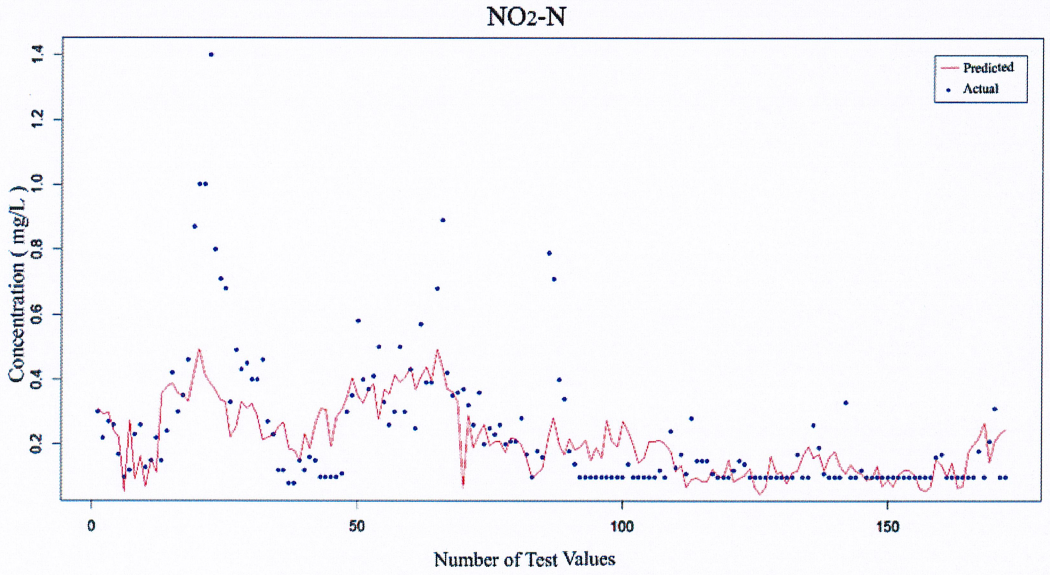


Fig. 36: Predicted vs True Values of the Test Data of NO_2-N

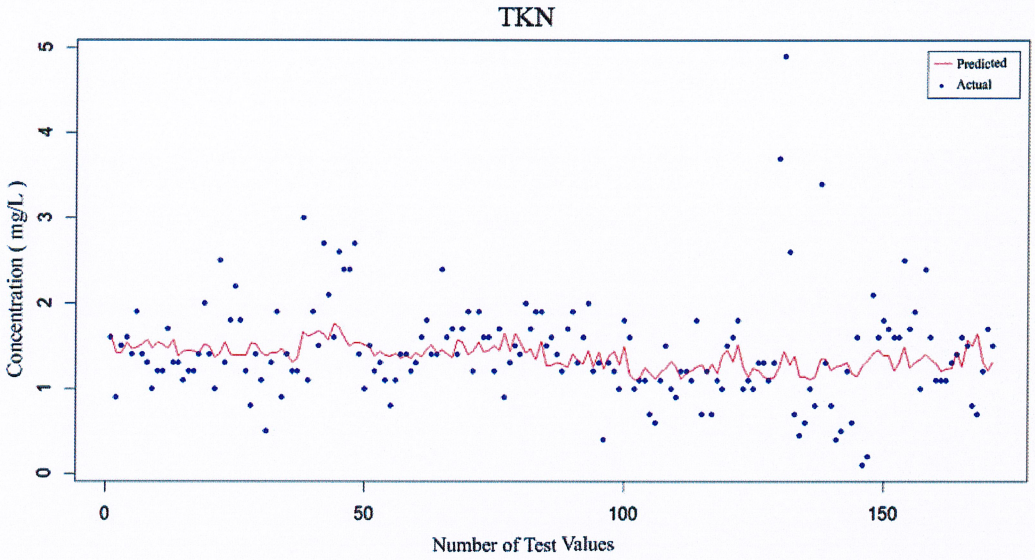


Fig. 37: Predicted vs True Values of the Test Data of TKN

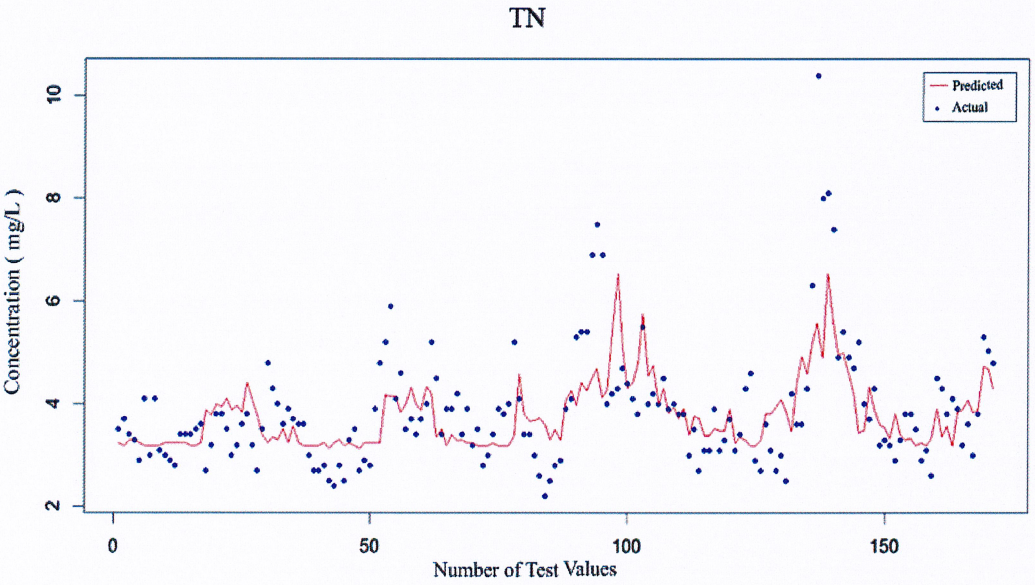


Fig. 38: Predicted vs True Values of the Test Data of TN

جامعة البحرين

كلية العلوم

قسم برامج الدراسات العليا



تنبُّؤ أداء النترجة لمحطة المحرق لمعالجة مياه الصرف الصحي بإستخدام تقنيات التعلم الآلي

إطروحة مقدمة كجزء من متطلبات الحصول على درجة الماجستير
في علوم وتحليل البيانات الضخمه

إعداد

سلمان يوسف حماد

٢٠١٦٧٥٧٢

إشراف

د . مجيد صفر جاسم
أستاذ مشارك

جامعة البحرين

مملكة البحرين
مارس ٢٠٢٣

الملخص

أصبح التعلم الآلي اليوم أحد أساليب النمذجة الحديثة الأكثر سرعة وفعالية في دراسة وتحليل العمليات الكيميائية والبيولوجية. تقوم محطة معالجة مياه الصرف الصحي بإزالة الملوثات من مياه المجاري، وتنتج مياه معالجة تتماشى مع المعايير البيئية قبل عملية التصريف. يتمثل الهدف من مرحلة النترجة في محطة معالجة مياه الصرف الصحي في إزالة النيتروجين بإستخدام البكتيريا الآزوتية ، حيث إن النيتروجين يمكن أن يكون ضارًا للكائنات الحية التي تعيش في المسطحات المائية حيث يتم إطلاقه فيها. تستخدم هذه الدراسة التعلم الآلي المراقب للتنبؤ بأداء عملية النترجة لمياه الصرف الصحي لمحطة مدينة المحرق بإستخدام عشر متغيرات من المياه الداخلة للتنبؤ بتركيز التفريغ لكلٍ من نيتروجين الأمونيوم، نيتروجين النتريت، نيتروجين كيلدال الكلي، وإجمالي النتروجين. تم استخدام خمسة أساليب للتعلم اللآلي المراقب، بما في ذلك خوارزمية أقرب جار، الشبكة العصبية ذات المستقبلات المتعددة الطبقات، انحدار المتجه الداعم، خوارزمية الغابة العشوائية، آلة التعزيز الأشتقاقي. تم استخدام أربعة مقاييس تقييم لتقييم أداء كل نموذج، متوسط الخطأ المطلق، متوسط الخطأ المطلق المئوي، جذر متوسط مربع الخطأ، ومعامل التحديد. أظهرت النتائج أن انحدار المتجه الداعم كان الأكثر فعالية في التنبؤ بنيتروجين الأمونيوم بينما تفوقت خوارزمية الغابات العشوائية على غيرها من النماذج في تنبؤ كلاً من نيتروجين النتريت، نيتروجين كيلدال الكلي، وإجمالي النتروجين بمتوسط خطأ مطلق بلغ ٠,٢٧٥٩، ٠,٠٨٧٨، ٠,٣٨١٢، ٠,٤٥٨٣ بالتوالي. بالإضافة إلى ذلك، فإن النتائج تشير إلى أن حالة الثبات لدرجة الحرارة هي المتغير الأكثر تأثيراً على تركيز كِلا من أمونيوم النتروجين و إجمالي النتروجين. تظهر النتائج إلى أن نموذجي الغابة العشوائية و انحدار المتجه الداعم يمكنهما إدارة عملية النترجة بفعالية، مما يؤدي إلى إنتاج مياه معالجة تفي بمعايير الري في البحرين وتحمي النظام البيئي.